Vocabulary Learning and Instruction

Volume 6, Number 1,

November 2017

doi: http://dx.doi.org/10.7820/vli.v06.1.2187-2759

VLI Editorial Team

Editorial Board: Jeffrey Stewart, Luke Fryer, Raymond Stubbe, Aaron Gibson, Peter Carter

Reviewers: Paul Meara, Norbert Schmitt, John A. Read, Stuart Webb, John P. Racine, Tomoko Ishii, Tim Stoeckel, Dale Brown, Jon Clenton, Stuart McLean, Peter Thwaites, Tatsuya Nakata, Kiwamu Kasahara, Masumi Kojima, James Rogers, Yuko Hoshino, Vivienne Rogers, Yu Kanazawa

Copy Editors: Alex Cameron, Andrew Gallacher, Peter Harold, Mark Howarth, Linda Joyce, Tim Pritchard, Andrew Thompson, Alonzo Williams, Ian Mumby, Pete Westbrook, Melanie González, Stephen Cutler

The Editorial Team expresses a sincere thank you to Mana Ikawa, who designed the cover for the print version of *VLI*.

Copyright © 2017 *Vocabulary Learning and Instruction*, ISSN: Online 2187-2759; Print 2187-2767. All articles are copyrighted by their respective authors.

Vocabulary Learning and Instruction

Volume 6, Number 1, November 2017
doi: http://dx.doi.org/10.7820/vli.v06.1.2187-2759

Table of Contents

Articles **Page**

Letter from the Editor iv
Raymond Stubbe

Regular Article

The Impact of Learner-Related Variables on Second Language Incidental Vocabulary Acquisition through Listening 1
Sarvenaz Hatami

Vocab@Tokyo Articles

The Impact of Semantic Clustering on the Learning of Abstract Words 21
Tomoko Ishii

Responding to Research Challenges Related to Studying L2 Collocational Use in Professional Academic Discourse 32
Birgit Henriksen and Pete Westbrook

The Use of Psycholinguistic Formulaic Language in the Speech of Higher Level Japanese Speakers of English 48
Stephen F. Cutler

Profiling Lexical Diversity in College-level Writing 61
Melanie C. González

i-lex v1 and v2: An Improved Method of Assessing L2 Learner Ability to See Connections between Words? 75
Ian Munby

Letter from the Editor

Dear Readers,

It is my great pleasure to offer you the November 2017 issue of *Vocabulary Learning and Instruction* (VLI). In the following pages, you will find a selection of papers presented at the Vocab@Tokyo Conference; a most enjoyable event that was held on September 12-14, 2016. This issue also features a regularly submitted article by Sarvenaz Hatami of California State University, Long Beach.

As a reminder, *VLI* is an open-access international journal that provides a peer-reviewed forum for original research related to vocabulary acquisition, instruction, and assessment. Submissions are encouraged from researchers and practitioners in both first language, and EFL and ESL contexts.

Please enjoy this issue,
Raymond Stubbe,
Editor, *VLI*

Vocabulary Learning and Instruction
Volume 6, Issue 1, November 2017
http://vli-journal.org

The Impact of Learner-Related Variables on Second Language Incidental Vocabulary Acquisition through Listening

Sarvenaz Hatami
California State University Long Beach
doi: http://dx.doi.org/10.7820/vli.v06.1.Hatami

Abstract

Little is known about the complex process of L2 incidental vocabulary acquisition from listening and the factors that contribute to its success. To expand our knowledge in this area, the present study investigated the impact of five learner-related variables on L2 incidental vocabulary acquisition from listening. These variables were gender, L2 vocabulary knowledge, amount of L2 listening (for academic purposes and pleasure), level of enjoyment, and (self-reported) level of comprehension. Ninety-nine Iranian English as a foreign language (EFL) learners at pre-intermediate levels of English proficiency were randomly assigned to a listening group and a control group. Sixteen target words were chosen in a graded reader and were then replaced by 16 English-like non-words. The participants listened to the graded reader containing the 16 non-words and completed a vocabulary post-test immediately after the listening session. The post-test measured participants' knowledge of five different dimensions of word knowledge at the level of recognition. The findings revealed that while gender and amount of L2 listening appear to have no impact on incidental vocabulary gains from listening, L2 vocabulary knowledge, level of enjoyment, and level of comprehension are important facilitating factors.

1 Introduction

Before the 1970s, listening was assumed to be a receptive language skill in which listeners passively assimilate messages from incoming speech (Morley, 1984, as cited in Murphy, 1991). Today, listening comprehension is described as a far more complex process, critical to second-language (L2) acquisition, and the most difficult of the four language skills to learn (Vandergrift, 2004). Not unexpectedly, incidental vocabulary acquisition from listening is also a complex process involving many different factors. In his review essay on factors affecting the incidental acquisition of L2 vocabulary from oral input, Ellis (1994) emphasized that very little attention had been paid to this area of research. Surprisingly, after more than 20 years, the need for further study still exists. While there is a considerable amount of research on L2 incidental vocabulary acquisition through reading, research on L2 incidental vocabulary acquisition through listening is scarce (Brown, Waring, & Donkaewbua, 2008; van Zeeland & Schmitt, 2013a; Vidal, 2003). As a result, little is known about the development of vocabulary knowledge from L2 listening and the word-, text-, task-, and learner-related variables that

Please cite this article as: Hatami, S. (2017). The Impact of Learner-Related Variables on Second Language Incidental Vocabulary Acquisition through Listening. *Vocabulary Learning and Instruction, 6* (1), 1–20. doi: 10.7820/vli.v06.1.Hatami

play a role in this process. Nevertheless, the importance of L2 incidental vocabulary acquisition through listening cannot be underestimated, and children's sizeable vocabulary development in their first language (L1), before learning to read, attests to this (Ellis, 1994).

The objective of this work, therefore, was to explore some of the learner-related variables that might contribute to L2 incidental vocabulary acquisition from listening. Studies have shown that listening is a less effective input mode than reading for L2 incidental vocabulary acquisition (Brown et al., 2008; Hatami, 2017; Vidal, 2011). L2 learners have also reported that listening is their least preferred input mode when compared to reading and reading-while-listening (Brown et al., 2008). In order to better understand and ultimately reduce the complications learners face in L2 incidental vocabulary acquisition from listening, more needs to be known about this complex process and the factors that contribute to its success.

The learner-related variables chosen for inclusion in the present study were gender, L2 vocabulary knowledge, amount of L2 listening (for academic purposes and pleasure), level of enjoyment, and (self-reported) level of comprehension. L2 research has shown that these variables play a role in incidental vocabulary acquisition from reading (e.g., Elgort & Warren, 2014); in this study, the aim was to determine whether these learner-related variables also play a role in L2 incidental vocabulary acquisition from listening. Reading and listening, despite their differences, share important comprehension processes; for instance, they both involve decoding and interpretation using two basic knowledge sources: linguistic knowledge and world knowledge (Vandergrift & Baker, 2015). Because of such important similarities, and also because L2 listening research is limited, "it is common practice for listening researchers to use reading-based findings as their starting point" (van Zeeland, 2014, p. 1007).

In addition to evidence from L2 reading research, a number of listening studies, although not directly focused on incidental vocabulary acquisition, indirectly suggest that some of the learner-related variables chosen for this study might play a role in incidental vocabulary acquisition from listening. Regarding the role of gender, for instance, differences between males and females have been reported in strategy use while listening in the L2 (Bacon, 1992). However, there are also studies which have failed to show any significant gender differences in L2 listening comprehension ability (e.g., Bacon, 1992; Feyten, 1991; Vandergrift, 2006) or strategy use (Vandergrift, 1997). Furthermore, L2 vocabulary knowledge has been shown to be an important factor for successful L2 listening comprehension (Mecartty, 2000; Stæhr, 2009). And finally, enjoyment and L2 listening comprehension have been shown to be closely related (Ducker & Saunders, 2014).

2 Literature Review

2.1 Factors Affecting L2 Incidental Vocabulary Acquisition through Listening

L2 research has shown that both reading and listening can be a source of incidental vocabulary acquisition. However, while L2 incidental vocabulary acquisition through reading and the factors involved have been widely examined,

research on factors contributing to L2 incidental vocabulary acquisition through listening, particularly learner-related factors, is very limited. Vidal (2003) investigated the impact of two learner-related variables, L2 proficiency and lecture comprehension, on incidental vocabulary acquisition from academic listening with 116 university-level Spanish EFL learners. The findings revealed that both L2 proficiency and lecture comprehension impact the degree to which vocabulary is gained from academic listening: the higher the level of L2 proficiency and lecture comprehension, the greater the vocabulary gains. Moreover, in a study with 172 Chinese university EFL learners at pre-intermediate to intermediate levels of language proficiency, Chang (2012) examined the relationship between metacognitive listening awareness, listening comprehension, and incidental vocabulary acquisition, and found that they are related; however, the correlations were generally not strong.

In addition, a few studies have examined the word- and task-related variables that could play a role in L2 incidental vocabulary acquisition through listening. These variables include: frequency of word occurrence in the text (Brown et al., 2008; Hatami, 2017; van Zeeland & Schmitt, 2013a; Vidal, 2003, 2011); predictability from word form and parts (i.e., unpredictable, deceptively transparent, morphologically predictable, similar to L1); word type (i.e., low-frequency, technical, academic); type of elaboration (i.e., explicit, implicit, no elaboration) (Vidal, 2003, 2011); part of speech; concreteness (van Zeeland & Schmitt, 2013a); and repetition of the listening text (Chang, 2012).

As the above review indicates, the L2 studies that have explored the impact of certain variables on L2 incidental vocabulary acquisition from listening are too few in number to allow any general conclusions. Therefore, attempts at establishing previous findings or exploring new variables would be worthwhile, and this is what this study set out to accomplish, by focusing on learner-related variables, which previous research has examined to a surprisingly limited extent.

2.2 Depth of Vocabulary Knowledge

Depth of vocabulary knowledge is a broad, vague, and imprecisely-defined construct, which L2 researchers have conceptualized and measured in different ways (Schmitt, 2014). A common approach to describing depth of vocabulary knowledge is to draw a distinction between receptive vocabulary knowledge (i.e., the ability to comprehend lexical items during reading or listening) and productive vocabulary knowledge (i.e., the ability to produce lexical items during speaking or writing). One way of conceptualizing depth of vocabulary knowledge is to think of vocabulary knowledge along a continuum of mastery; as more and more is acquired about a word, mastery of the word gradually shifts from the receptive end toward the productive end of the continuum (Melka, 1997). However, if there really is a continuum, the location of "the threshold at which the word passes from receptive to productive status" is unclear (Read, 2000, p. 154). Other scholars have viewed the acquisition of depth of vocabulary knowledge along a set of discrete stages (rather than along a continuum) and have used progression scales for its measurement (e.g., Wesche & Paribakht, 1996); however, such scales have long been the subject of criticism (Schmitt, 2010). Meara (1997) proposes another conceptualization of depth of vocabulary knowledge, which tends to focus

on the lexicon as a whole, rather than on isolated words. According to Meara, the mental lexicon consists of an interrelated network of lexical items; when a new word is acquired, it is integrated into this network of already-known items. Based on this view, the greater the degree of integration of a word into this network, in other words, the greater the number of links between a word and its related items (through associations, collocations, etc.), the greater its depth. However, as Schmitt (2014) points out, while this approach seems very promising, "unfortunately, our understanding of lexical organization is not yet advanced enough to pursue this direction in a tangible way" (p. 943). Another well-known approach to conceptualizing depth of vocabulary knowledge is termed the *dimensions* or *components* approach (Read, 1997; Schmitt, 2010). In this approach, which is known as one of the most effective ways of measuring depth of vocabulary knowledge (Nation & Webb, 2011), vocabulary knowledge is broken down into various isolated dimensions (see Table 1 for the nine different dimensions of word knowledge proposed by Nation, 1990, 2001). Receptive or productive knowledge of each dimension is then separately measured.

In the present study, the vocabulary knowledge gained through listening was conceptualized and measured using the dimensions approach. While the dimensions approach has been used quite extensively in L2 reading studies on incidental vocabulary acquisition, only two of the L2 listening studies reviewed above have used the dimensions approach to measure incidental vocabulary gains (i.e., Hatami, 2017; van Zeeland & Schmitt, 2013a). Other L2 listening studies have either used a scale (i.e., Chang, 2012; Vidal, 2003, 2011) or have only measured one or two dimensions of word knowledge, that is, written form and/or form-meaning connection (i.e., Brown et al., 2008; Chang, 2012). As van Zeeland and Schmitt point out, since "learning gains from listening have [been] found to be small, even significantly smaller than those from reading, the dimensions approach should serve particularly well in revealing the smallest increments in learning" (p. 611).

3 The Present Study

In this study, the impact of five learner-related variables on L2 incidental vocabulary acquisition through listening was examined. The learner-related variables were gender, L2 vocabulary knowledge, amount of L2 listening (for academic purposes and pleasure), level of enjoyment, and (self-reported) level of comprehension. Although word-, text-, and task-related variables can also play an important role in L2 incidental vocabulary acquisition from listening, the primary focus here was on variables related to the learner/listener, in the hope that additional research will extend to other variables.

Table 1. What Is Involved in Knowing a Word (Nation, 1990, 2001)

Form	Meaning	Use
Spoken form	Form-meaning connection	Grammatical functions
Written form	Concept and referents	Collocations
Word parts	Associations	Constraints on use (e.g., register, frequency)

4 Method

4.1 Participants

Ninety-nine undergraduate students (57 males, 42 females) majoring in engineering at a high-ranking university in Iran participated in this study. The participants ranged in age between 18 and 24 years (M = 19.58, SD = 1.36) and all spoke Farsi as their L1. They had all formally studied English for 7 years at school before entering university, and none had ever lived in an English-speaking country. The vocabulary knowledge of the participants was determined using the Vocabulary Levels Test (VLT) (Schmitt, Schmitt, & Clapham, 2001). The mean scores on the 2,000, 3,000, and 5,000 word levels of the VLT were 23.07, 15.13, and 7.43, respectively (maximum score at each level = 30). Only participants with at least 50% mastery of the 2,000 word level were included in the study. This cut-off point was determined to ensure that participants had knowledge of the running words in the listening text and could therefore understand the text with little or no difficulty. None of the participants reported hearing difficulties. All participants received cash incentives (equivalent to $10 CAD) for their participation. The participants were randomly assigned to a listening group (n = 51) and a control group (n = 48).

4.2 Materials

4.2.1. Target words

For the purposes of this study, non-words were used, that is, words created by a complete change in the form of already known, common concepts (Waring & Takaki, 2003). Learning such non-words is the simplest level of learning a new word, as it only involves learning a new label, and not a new concept (Nation, 2001). Nevertheless, using such non-words in L2 research for lower proficiency learners such as those in this study is acceptable and common. This is because, firstly, in the initial stages of L2 learning, L2 vocabulary acquisition does not involve learning many new concepts; rather, learners typically acquire L2 word forms for their already-existing L1 concepts (Nation & Webb, 2011). Secondly, providing 95%–98% lexical coverage—which is known to provide adequate comprehension of spoken texts (Stæhr, 2009; van Zeeland & Schmitt, 2013b)—with truly unknown target words (i.e., words whose form and meaning are both unknown) often necessitates heavy adaptation of the text, since truly unknown words are typically infrequent words that tend to occur with other low-frequency words (Webb, 2005); and the heavy adaptation of the text can compromise the ecological validity of the study.

Sixteen words in the listening text were chosen as target words. The target words were then substituted throughout the text with 16 non-words (see Appendix A). Several steps were taken to ensure that the non-words looked like plausible English words and were equivalent, as much as possible, in terms of learning difficulty. First, 46 non-words, all two-syllabic and five or six letters in length, were chosen from Meara's (2013) list of imaginary words. Next, three TESL experts judged the non-words with regard to their plausibility as real English words. Consequently, 16 of the 46 non-words were excluded due to one of the following

reasons: the non-word had irregular pronunciation and/or spelling, contained a real English word, was a popular English first/last name, or looked French. A questionnaire was then developed for the remaining 30 non-words and was administered to five native English speakers (mean age = 38 years) and five Iranian non-native English speakers (mean age = 29.8 years). The questionnaire asked the respondents, in a yes/no question, whether each non-word resembled a real English word. It also required the respondents to rate each non-word, on a scale of 1–5, in terms of its spelling and pronunciation difficulty (1 = *very easy*; 5 = *very difficult*). Based on the responses to the questionnaire, 16 of the non-words were selected to be used in the study. These 16 non-words were rated as plausible English words by at least 8 of the 10 respondents to the questionnaire, and their average spelling difficulty and pronunciation difficulty were rated lower than 3 on the 5-point scale.

4.2.2. Listening material

The listening text chosen for this study was *The Monkey's Paw*, an elementary-level graded reader selected from the Oxford Bookworms series. To ensure that participants had knowledge of all the running words in the text, the text was further simplified. First, the researcher (a native Farsi speaker) changed the proper nouns which were thought to be unfamiliar to the participants to more familiar ones (e.g., *Herbert* was changed to *Jack*). In addition, using the BNC-COCA-25 vocabulary profiler available at www.lextutor.ca/vp/, words in the text that were beyond the 1,000 word-level were either substituted with words from this level or eliminated. The final text contained 4,231 words, and after inserting the non-words, a lexical coverage of 95.84% was reached. A lexical coverage of 95%–98% has been shown to provide adequate comprehension of spoken texts (Stæhr, 2009; van Zeeland & Schmitt, 2013b).

The final version of the text with the inserted non-words was audio-recorded as it was read aloud by a TESL professor who was a native speaker of Canadian English. The duration of the narration was 36 min, with an average speech rate of 117.5 words per minute.

4.3 Instruments

4.3.1. Language background questionnaire

The language background questionnaire, translated into Farsi, was designed to collect a range of information about the participants. In addition to demographic information (i.e., gender, age, native country, native language, other languages spoken, and proficiency levels in those languages), participants reported whether or not they had lived in an English-speaking country and how long they had studied English outside of school and university. Moreover, the participants were asked to estimate the amount of time that they spent in a typical week listening to English materials for academic purposes (e.g., lectures, language learning CDs) and for pleasure (e.g., movies, radio, audio books). The two purposes for listening were separated, in order to help learners more accurately calculate their

amount of L2 listening in a typical week. Because of the EFL context of the learners and the very low possibility of learners engaging in English conversations, conversational listening was not included in the questionnaire.

4.3.2. Vocabulary Levels Test

The VLT, originally developed by Nation (1983) and updated and validated by Schmitt et al. (2001), was used in this study to measure L2 vocabulary knowledge. The VLT, which is a test of receptive vocabulary knowledge, consists of four sections that represent four distinct word frequency levels (the 2,000, 3,000, 5,000, and 10,000 frequency levels) as well as a section for academic vocabulary. In this study, because the 10,000 word level appeared to be beyond the vocabulary knowledge of the participants, only the sections related to the 2,000, 3,000, and 5,000 word levels were administered. In scoring, each word correctly chosen was awarded one point. Because each section had 30 test items, and three sections were used in this study, the maximum possible score was 90. Cronbach's alpha for the entire test (all three sections together) was 0.85.

4.3.3. Vocabulary post-test

To capture the vocabulary knowledge gained through listening, five dimensions of word knowledge were selected from the nine proposed by Nation (1990, 2001): spoken form, written form, part of speech, syntagmatic association, and form-meaning connection. All these five dimensions were measured at the level of recognition, and therefore, the vocabulary post-test consisted of five tests (see Appendix B). The post-test was adapted from the works of Webb (2005), Chen and Truscott (2010), and van Zeeland and Schmitt (2013a).

Each of the five tests appeared on two consecutive pages facing each other, with 8 (of the 16) target words on one page and another 8 on the next. Moreover, on the back of the last page of the post-test, two 5-point scales were provided to measure learners' level of enjoyment from listening to the story and level of understanding of the story (see Appendix C). Following Webb (2005), the tests were sequenced so that any possibility of learning effect was avoided. For example, recognition of the written form preceded recognition of form-meaning connection because the correct response to the former was provided in the latter. Instructions for all sections of the post-test appeared in both Farsi and English. Participants were asked to avoid making any changes to the answers they had provided in previous sections of the post-test and, as they were taking the post-test, they were carefully supervised to ensure this. (An easier and more reliable alternative could have been to collect the tests from the participants immediately after each section was completed.) The five recognition tests were scored dichotomously.

4.4. Procedures

Before collecting data, the materials and instruments were piloted with four Iranian EFL learners with characteristics similar to those of the population under study; consequently, changes were made to some of the instructions

and Farsi translations. Data were then collected during two sessions that were 2 weeks apart:

4.4.1 Session one

If they agreed to participate in the study, participants signed a consent form, after which they were asked to complete the language background questionnaire and the VLT. This session lasted approximately 50 min.

4.4.2 Session two

Participants were told that the objective of this session was to listen to a classic English story and to try to understand it. They were not informed of the vocabulary focus of the study or the vocabulary post-test. However, immediately after they listened to *The Monkey's Paw* (played from a CD), the unannounced vocabulary post-test was administered. This session lasted approximately 75 min.

The control group completed all the abovementioned procedures (i.e., the consent form, language background questionnaire, VLT, and vocabulary post-test), but were not exposed to the listening text.

5 Results

In all the analyses reported below, word recognition was calculated by averaging the scores on the five recognition tests (i.e., recognition tests of spoken form, written form, part of speech, syntagmatic association, and form-meaning connection). Table 2 presents descriptive word recognition statistics for the listening group and the control group.

5.1 Gender

In order to examine the impact of gender on L2 incidental vocabulary acquisition from listening, a two-way between-subjects ANOVA was conducted with group (listening vs. control) and gender (male vs. female) as the independent variables and recognition scores as the dependent variable. The results of the ANOVA yielded a significant main effect for group, $F(1, 95) = 48.72$, $p < 0.001$, partial $\eta^2 = 0.34$, power = 1.0. However, the effects were not significant for gender, $F(1, 95) = 0.05$, $p = 0.83$, or for the interaction between group and gender, $F(1, 95) = 0.68$, $p = 0.41$. Descriptive statistics are presented in Table 3.

Table 2. Descriptive Statistics for Group Scores on the Vocabulary Post-test

Group	*N*	*M*	SD	95% Confidence interval	
				Lower bound	Upper bound
Listening	51	6.24	2.56	5.52	6.95
Control	48	3.10	1.62	2.63	3.57

Note: The maximum possible score is 16.

Table 3. Descriptive Statistics for Males and Females

Group	Gender	*N*	*M*	SD
Listening	Male	28	6.44	2.63
	Female	23	5.98	2.50
Control	Male	29	2.99	1.80
	Female	19	3.26	1.31

Note: The maximum possible score on the vocabulary post-test is 16.

5.2 L2 Vocabulary Knowledge

To investigate the effect of L2 vocabulary knowledge on L2 incidental vocabulary acquisition through listening, the scores on the 2,000, 3,000, and 5,000 word levels of the VLT were combined (M = 45.64, SD = 12.34, range = 23–78, maximum score = 90). The mean was then used as the cut-point to divide the participants into two groups: those who scored at or above 45.64 were classified as having relatively "extensive" vocabulary knowledge, and those who obtained scores below the mean were classified as having relatively "limited" vocabulary knowledge (see Table 4 for descriptive statistics). Next, a two-way between-subjects ANOVA was run with group (listening vs. control) and L2 vocabulary knowledge (extensive vs. limited) as the independent variables and recognition scores as the dependent variable. The ANOVA revealed a significant effect for group, $F(1, 94) = 62.60$, $p < 0.001$, partial $\eta^2 = 0.40$, power = 1.0; for L2 vocabulary knowledge, $F(1, 94) = 9.05$, $p < 0.05$, partial $\eta^2 = 0.09$, power = 0.85; and for the interaction between group and L2 vocabulary knowledge, $F(1, 94) = 8.55$, $p < 0.05$, partial $\eta^2 = 0.08$, power = 0.83. Simple effects analysis indicated a statistically significant difference between limited and extensive vocabulary knowledge in the listening group ($p < 0.001$), but not in the control group ($p = 0.95$).

5.3 Amount of L2 Listening

To examine the impact of the amount of L2 listening on incidental vocabulary gains, participants' number of hours of L2 academic listening in a typical week and number of hours of L2 pleasure listening in a typical week (as reported in their language background questionnaires) were added together. The distribution was skewed and, thus, to divide the participants into two groups, the median (instead of the mean) was chosen as the cut-point (M = 2.95, Mdn = 2.5,

Table 4. Descriptive Statistics for L2 Vocabulary Knowledge

Group	L2 Vocabulary knowledge	*N*	*M*	SD
Listening	Extensive	23	7.43	2.27
	Limited	27	5.07	2.20
Control	Extensive	24	3.12	1.68
	Limited	24	3.08	1.59

Note: The maximum possible score on the vocabulary post-test is 16.
One missing case (a participant's score was an outlier in this analysis and therefore excluded).

SD = 2.32, Range = 0–10). Those whose number of hours of L2 listening in a typical week fell at or above 2.5 hours were classified as doing relatively "extensive" amounts of L2 listening, and those whose number of hours of L2 listening in a typical week fell below the median were classified as doing relatively "limited" amounts of L2 listening (see Table 5 for descriptive statistics). A two-way between-subjects ANOVA was then performed with group (listening vs. control) and amount of L2 listening (extensive vs. limited) as the independent variables and recognition scores as the dependent variable. The ANOVA yielded a significant main effect for group, $F(1, 89) = 50.39$, $p < 0.001$, partial $\eta^2 = 0.36$, power = 1.0. However, the effects were not significant for amount of L2 listening, $F(1, 89) = 2.89$, $p = 0.09$, or for the interaction between group and amount of L2 listening, $F(1, 89) = 1.13$, $p = 0.29$.

5.4 Level of Enjoyment

To investigate the impact of level of enjoyment (experienced while listening to the story) on incidental vocabulary acquisition, the 5-point scale used in the study to measure level of enjoyment was collapsed into a 3-point scale (1 = *disagree*, 2 = *neutral*, 3 = *agree*). This was done in the following way: in response to the statement *I enjoyed the story,* if the participants marked 1 or 2 on the scale, it was regarded as "disagree"; if they marked 3, it was considered "neutral," and if they marked 4 or 5 on the scale, it was considered "agree" (see Table 6 for descriptive statistics). Next, a one-way between-subjects ANOVA was conducted with enjoyment (disagree vs. neutral vs. agree) as the independent variable and recognition scores as the dependent variable. The ANOVA yielded a significant main effect for enjoyment, $F(2, 45) = 4.55$, $p < 0.05$, partial $\eta^2 = 0.17$, power = 0.75. Tukey honestly significant difference (HSD) test revealed a statistically significant difference between "agree" and "disagree" ($p < 0.05$). However, the differences were not significant between "agree" and "neutral" ($p = 0.07$) or "disagree" and "neutral" ($p = 0.84$).

Table 5. Descriptive Statistics for Amount of L2 Listening

Group	Amount of L2 Listening	*N*	*M*	SD
Listening	Extensive	25	6.82	2.65
	Limited	25	5.58	2.37
Control	Extensive	24	3.15	1.83
	Limited	19	2.86	1.32

Note: The maximum possible score on the vocabulary post-test is 16.
Six missing cases (the scores of six participants were outliers in this analysis and therefore excluded).

Table 6. Descriptive Statistics for Level of Enjoyment

Group	Enjoyed the Story	*N*	*M*	SD
Listening	Disagree	12	4.92	2.14
	Neutral	15	5.44	2.69
	Agree	21	7.29	2.35

Note: The maximum possible score on the vocabulary post-test is 16.
Three missing cases (three participants did not provide data).

Table 7. Descriptive Statistics for Level of Comprehension

Group	Understood the Story	*N*	*M*	SD
Listening	Disagree	6	3.77	1.15
	Neutral	8	4.28	1.68
	Agree	34	6.96	2.48

Note: The maximum possible score on the vocabulary post-test is 16.
Three missing cases (three participants did not provide data).

5.5 Level of Comprehension

To examine the impact of level of comprehension on incidental vocabulary acquisition from listening, similar to the previous section, the 5-point scale used in the study to measure level of comprehension was collapsed into a 3-point scale (1 = *disagree*, 2 = *neutral*, 3 = *agree*) in the following way: in response to the statement *I understood the story,* if the participants marked 1 or 2 on the scale, it was regarded as "disagree"; if they marked 3, it was considered "neutral," and if they marked 4 or 5 on the scale, it was considered "agree." Descriptive statistics are presented in Table 7. A one-way between-subjects ANOVA was conducted with comprehension (disagree vs. neutral vs. agree) as the independent variable and recognition scores as the dependent variable. The ANOVA yielded a significant main effect for comprehension, $F(2, 45) = 8.3$, $p < 0.05$, partial $\eta^2 = 0.27$, power = 0.95. Tukey HSD revealed a statistically significant difference between "agree" and "disagree" ($p < 0.05$), and between "agree" and "neutral" ($p < 0.05$), but a non-significant difference between "disagree" and "neutral" ($p = 0.91$).

6 Discussion and Conclusions

6.1 Gender

In this study, males scored higher than females on the vocabulary post-test, but the difference between the two groups was not statistically significant. Hence, it appears that gender had no impact on L2 incidental vocabulary acquisition from listening. Since L2 listening comprehension influences L2 incidental vocabulary acquisition (Vidal, 2003), the lack of gender differences in this study is congruent with studies which have shown that gender does not play a significant role in L2 listening comprehension (e.g., Bacon, 1992; Feyten, 1991; Vandergrift, 2006) as well as studies that have found minimal differences between males and females regarding their self-reported strategy use while listening in the L2 (e.g., Vandergrift, 1997). Hence, although females have generally been considered more successful foreign language learners than males and their greater success is hypothesized to be related to the interaction of neurological, cognitive, affective, social, and educational factors (Rúa, 2006), this superiority does not appear to apply to incidental vocabulary acquisition from listening.

6.2 L2 Vocabulary Knowledge

In this study, learners with a larger L2 vocabulary scored significantly higher on the vocabulary post-test than learners with a smaller L2 vocabulary.

L2 vocabulary knowledge therefore impacts the incidental acquisition of L2 vocabulary through listening. One explanation for this finding is that L2 vocabulary knowledge contributes to L2 listening comprehension (Mecartty, 2000; Stæhr, 2009), and L2 listening comprehension appears to contribute to incidental vocabulary acquisition (Vidal, 2003). In other words, the greater one's L2 vocabulary knowledge and, consequently, L2 proficiency (Stæhr, 2008), the greater the amount of L2 spoken input that can be successfully processed and understood (Vidal, 2003), and thus, the larger the vocabulary gains from that input. Previous reading studies have also shown that L2 lexical proficiency is an important factor in L2 incidental vocabulary acquisition (Elgort & Warren, 2014).

6.3 Amount of L2 Listening

Although learners who reported more L2 listening in a typical week scored higher on the vocabulary post-test than learners who reported less L2 listening, the difference between the two groups was not statistically significant. Hence, amount of L2 listening did not appear to have an impact on L2 incidental vocabulary acquisition from listening. This finding suggests that mere exposure to more L2 listening opportunities in an EFL context does not significantly enhance one's success in incidental vocabulary acquisition from listening. In addition to repeated practice, it appears that instruction and strategy training in L2 listening comprehension and in the use of context are needed. As Vandergrift (2004) states, "students need to 'learn to listen' so that they can better 'listen to learn'" (p. 3). Considering that EFL education in Iranian formal schools and universities focuses heavily on the grammar-translation method and reading comprehension (Farhady, Hezaveh, & Hedayati, 2010; Kiany, Mahdavy, & Samar, 2011), it is not surprising that Iranian EFL learners lack the necessary skills and strategies to take full advantage of their L2 listening and incidental vocabulary acquisition opportunities. This situation exists not only in Iran, but also in other EFL contexts such as Japan (Nishino & Watanabe, 2008). Learners indeed need to "learn to listen" and learn to pay more attention to context, and, in fact, studies have shown improvements in listening comprehension as a result of L2 listening instruction (Goh & Taib, 2006) and improvements in incidental vocabulary acquisition from listening (in terms of word form recognition only) as a result of lexical inferencing training (Chang, 2012).

It should also be noted that retrospective reports of the amount of L2 listening in a typical week (as was the case in this study) may not be very reliable and, thus, these results should be interpreted with caution. Asking learners to keep a daily log or journal of their amount of L2 listening over a specified period of time might be a better option for collecting such data.

6.4 Level of Enjoyment

The degree to which learners enjoy listening to a text appears to affect L2 incidental vocabulary gains. In this study, in response to the statement *I enjoyed the story,* those learners who rated "strongly agree" or "agree," scored significantly higher on the vocabulary post-test than those who rated "disagree" or "strongly

disagree." This result aligns with findings from Ducker and Saunders' (2014) study with intermediate-level Japanese-speaking EFL learners, in which enjoyment and listening comprehension were found to be strongly related. L2 reading studies that have investigated the impact of enjoyment on incidental vocabulary acquisition (e.g., Elgort & Warren, 2014) have reported similar results. These findings suggest that to enhance L2 listening comprehension and incidental vocabulary gains, materials chosen for L2 listening should be interesting and enjoyable to the learners, which indicates the importance of learners self-selecting the topic and text they wish to listen to, where possible. Enjoying the listening material can be so facilitating that it might even compensate, to some extent, for the lack of adequate language proficiency (Waring, 2010).

6.5 Level of Comprehension

In response to the statement *I understood the story,* those learners who rated "strongly agree" or "agree," scored significantly higher on the vocabulary post-test than those who rated "neutral," "disagree," or "strongly disagree." Hence, level of comprehension impacts L2 incidental vocabulary acquisition from listening. This finding supports Vidal's (2003) study, in which she found that incidental vocabulary gains from academic listening appeared to be influenced by learners' degree of lecture comprehension: the higher the level of comprehension, the greater the vocabulary gains. Chang (2012) also found moderate correlations between L2 listening comprehension and incidental vocabulary acquisition. Similar results have also been reported in L2 reading studies (e.g., Elgort & Warren, 2014). These findings indicate the importance of helping learners access texts that are at their appropriate level in order to ensure comprehension and, consequently, incidental vocabulary acquisition.

7 Limitations and Suggestions for Further Research

A few limitations of this study deserve consideration. First, learners' vocabulary knowledge was measured using the VLT, which is a test of orthographic lexical knowledge. Since, in some learners, phonological and orthographic lexical knowledge may be quite different (Milton & Hopkins, 2006), it would have been more suitable to use a test of aural vocabulary knowledge in this study, such as the LVLT (McLean, Kramer, & Beglar, 2015). Second, in order to ensure that learners have knowledge of all the words in the text, the text was run through a word family-based profiler (i.e., the BNC-COCA-25 profiler); however, it should not be assumed that because learners demonstrate knowledge of the base form or most common form on the VLT, they will also have knowledge of all associated derivational forms. Hence, a lemma-based profiler (e.g., the NGSL profiler available at www.lextutor.ca/vp/) might have been more appropriate (see McLean, 2017, for why the lemma or flemma, a word's base form and associated inflectional forms, is likely a more appropriate word counting unit). Third, in this study, because of practical constraints, the vocabulary post-test was mainly in written format; this mismatch between the participants' mode of input and mode of measurement might have placed them at a disadvantage in terms of scores (Alali & Schmitt, 2012) on three of the five tests, that is, tests of part of speech, syntagmatic association, and form-meaning

connection. Finally, in this study, retention of incidental vocabulary gains from listening was not addressed. Initially, this was one of the objectives of this study and, in fact, when collecting data, a delayed vocabulary post-test was administered 3 weeks after the immediate post-test. However, when analyzing the data, it was revealed that the immediate post-test had impacted the scores on the delayed post-test. In other words, because of the presence of testing effects, retention scores had not been accurately measured and therefore could not be used and reported in this study. Future research could employ a different research design in order to avoid possible testing effects (see the research design in van Zeeland & Schmitt, 2013a).

In sum, this study showed that males and females are equally successful at incidentally acquiring vocabulary from listening. Furthermore, the results suggested that simply listening to L2 material may not be adequate to enhance L2 learners' abilities in incidental vocabulary acquisition; explicit instruction might also be needed. Moreover, three facilitating factors for incidental vocabulary acquisition from listening were revealed: L2 vocabulary knowledge, enjoyment from the listening content, and level of comprehension. Hence, learners with a large L2 vocabulary who have access to enjoyable, comprehensible texts are likely to gain more vocabulary from listening. Future research would benefit from identifying other facilitating variables, whether learner-, word-, text-, or task-related. Moreover, in this study, the scores on the five recognition tests measuring the five dimensions of word knowledge were combined. A more detailed analysis, in which each dimension of word knowledge is examined separately, could provide a more in-depth understanding of the process of incidental vocabulary acquisition through listening.

References

Alali, F.A., & Schmitt, N. (2012). Teaching formulaic sequences: The same as or different from teaching single words? *TESOL Journal, 3*(2), 153–180. doi:10.1002/tesj.13

Bacon, S.M. (1992). The relationship between gender, comprehension, processing strategies, and cognitive and affective response in foreign language listening. *The Modern Language Journal, 76*(2), 160–178. doi:10.1111/j.1540-4781.1992.tb01096.x

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language, 20*(2), 136–163.

Chang, L. (2012). *Investigating the relationships between Chinese university EFL learners' metacognitive listening strategies and their comprehension and incidental vocabulary acquisition from listening tasks.* Doctoral dissertation, University of Auckland, New Zealand.

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics, 31*(5), 693–713. doi:10.1093/applin/amq031

Ducker, N., & Saunders, M. (2014). Facilitating extensive listening with non-graded materials in EFL programs. *International Journal of Innovation in English Language Teaching and Research, 3*(2), 201–245.

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning, 64*(2), 365–414. doi:10.1111/lang.12052

Ellis, R. (1994). Factors in the incidental acquisition of second language vocabulary from oral input: A review essay. *Applied Language Learning, 5*(1), 1–32.

Farhady, H., Hezaveh, F.S., & Hedayati, H. (2010). Reflections on foreign language education in Iran. *TESL-EJ, 13*(4), 1–18.

Feyten, C.M. (1991). The power of listening ability: An overlooked dimension in language acquisition. *The Modern Language Journal, 75*(2), 173–180. doi:10.1111/j.1540-4781.1991.tb05348.x

Goh, C., & Taib, Y. (2006). Metacognitive instruction in listening for young learners. *ELT Journal, 60*(3), 222–232. doi:10.1093/elt/ccl002

Hatami, S. (2017). The differential impact of reading and listening on L2 incidental acquisition of different dimensions of word knowledge. *Reading in a Foreign Language*, *29*(1), 61–85.

Kiany, G.R., Mahdavy, B., & Samar, R.G. (2011). Towards a harmonized foreign language education program in Iran: National policies and English achievement. *Literacy Information and Computer Education Journal, 2*(3), 462–469.

McLean, S. (2017). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*. doi:10.1093/applin/amw050

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research, 19*(6), 741–760.

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

Meara, P. (2013). Imaginary words. In C.A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–4). Malden, MA: Wiley-Blackwell.

Mecartty, F.H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning, 11*(2), 323–348.

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

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review, 63*(1), 127–147. doi:10.3138/cmlr.63.1.127

Murphy, J.M. (1991). Oral communication in TESOL: Integrating speaking, listening, and pronunciation. *TESOL Quarterly, 25*(1), 51–75. doi:10.2307/3587028

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

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

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

Nation, I.S.P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston, MA: Heinle.

Nishino, T., & Watanabe, M. (2008). Communication-oriented policies versus classroom realities in Japan. *TESOL Quarterly, 42*(1), 133–138. doi:10.1002/j.1545-7249.2008.tb00214.x

Read, J. (1997). Vocabulary and testing. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 303–320). Cambridge, UK: Cambridge University Press.

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

Rúa, P.L. (2006). The sex variable in foreign language learning: An integrative approach. *Porta Linguarum, 6*, 99–114.

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

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913–951. doi:10.1111/lang.12077

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55–88. doi:10.1177/026553220101800103

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

Stæhr, L.S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition, 31*(4), 577–607. doi:10.1017/S0272263109990039

Vandergrift, L. (1997). The comprehension strategies of second language (French) listeners: A descriptive study. *Foreign Language Annals, 30*(3), 387–409. doi:10.1111/j.1944-9720.1997.tb02362.x

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics, 24*, 3–25. doi:10.1017/S0267190504000017

Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency? *The Modern Language Journal, 90*(1), 6–18. doi:10.1111/j.1540-4781.2006.00381.x

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning, 65*(2), 390–416. doi:10.1111/lang.12105

van Zeeland, H. (2014). Lexical inferencing in first and second language listening. *The Modern Language Journal, 98*(4), 1006–1021. doi:10.1111/modl.12152

van Zeeland, H., & Schmitt, N. (2013a). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System, 41*(3), 609–624. doi:10.1016/j.system.2013.07.012

van Zeeland, H., & Schmitt, N. (2013b). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34*(4), 457–479. doi:10.1093/applin/ams074

Vidal, K. (2003). Academic listening: A source of vocabulary acquisition? *Applied Linguistics, 24*(1), 56–89. doi:10.1093/applin/24.1.56

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning, 61*(1), 219–258. doi:10.1111/j.1467-9922.2010.00593.x

Waring, R. (2010). *Starting extensive listening.* Retrieved from http://www.robwaring.org/el/starting_extensive_listening.htm

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 130–163.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*(1), 33–52. doi:10.1017/S0272263105050023

Wesche, M., & Paribakht, T.S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review, 53*(1), 13–40.

Appendix A

Target Words and Corresponding Non-Words

Target word	Part of speech	Number of occurrences	Frequency band	Non-word
Chair	N.	2		Bartle
Big	Adj.	3	2–5	Scally
Tea	N.	4		Lorey
Smiled	V.	5		Kemble → Kembled
Watched	V.	7		Bamber → Bambered
Warm	Adj.	8	7–10	Turley
Noise	N.	9		Gamage
Window(s)	N.	10		Mollet(s)
Laughed	V.	12		Gummer → Gummered
Living-room	N.	13	12–15	Palote
Afraid	Adj.	14		Alden
Bed	N.	15		Hislop
Old	Adj.	17		Galpin
Asked	V.	18	17–20	Mundy → Mundied
Husband	N.	19		Pegler
Hand	N.	20		Lomax

Note: Only past tense verbs were used in the story.

Appendix B

Vocabulary Post-Test: Description and Examples

Recognition of spoken form

[This measure had an aural multiple choice format; participants heard a recording of the target word and three distracters twice and had 5 seconds to check the box corresponding to the correct spoken form of the target word.]

Example:

Participants heard:

Which pronunciation is correct? Please check the box.

Number one [2sec] A bartle [2sec] B bertel [2sec] C burdle [2sec] D bardel [2sec.]

Number one [2sec] A bartle [2sec] B bertel [2sec] C burdle [2sec] D bardel [5sec.]

At the same time, the participants saw on the test page:

Which pronunciation is correct? Please check (✓) the box.

1. ☐ A ☐ B ☐ C ☐ D

Recognition of written form

[This multiple choice test consisted of the target word and three distracters. The same distracters used for the test of spoken form were used for this test.]

Example:

Which spelling is correct? Please check (✓) the box.

1. ☐ bartle ☐ bertel ☐ burdle ☐ bardel

Recognition of part of speech

[For this test, the target word was presented in three different sentences. Each sentence used the target word as a different part of speech. Only one of the sentences was correct, and the other two were distracters. In order to avoid any learning effects on the tests that follow, sentences were created in such a way that no clues to the meaning of the target words were provided.]

Example:

Which sentence is correct? Please check (✓) the box.

1. bartle ☐ It is a bartle. (Noun)

☐ He is very bartle. (Adjective)

☐ She bartled. (Verb)

Recognition of syntagmatic association

[In this test, the target word was presented followed by four choices; one choice was in a sequential relationship with the target word and the other three choices were distracters. All choices were in the same word class. Because the correct option was a target word in the passage, all the distracters were chosen from the passage as well.]

Example:

Which word is more likely to be used with ***bartle*** in a sentence? Please check (✓) the box.

☐ sit ☐ go ☐ open ☐ stop

Recognition of form-meaning connection

[In this final test, the target word was presented followed by four options: the original real English word which it had replaced in the text and three distracters. The distracters belonged to the same word class. Because the correct option had not been listened to in the passage, all the distracters were chosen from outside the passage as well.]

Example:

Which is the correct meaning for ***bartle***? Please check (✓) the box.

☐ book ☐ chair ☐ food ☐ head

Appendix C

Scales Measuring Comprehension and Enjoyment

[Instructions for these scales were provided orally in Farsi by the researcher.]

The following statements are about "The Monkey's Paw."

1. I understood the story.

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

2. I enjoyed the story.

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Vocabulary Learning and Instruction
Volume 6, Issue 1, November 2017
http://vli-journal.org

The Impact of Semantic Clustering on the Learning of Abstract Words

Tomoko Ishii
Meiji Gakuin University
doi: http://dx.doi.org/10.7820/vli.v06.1.Ishii

Abstract

It has been repeatedly argued among vocabulary researchers that semantically related words should not be taught simultaneously because they might interfere with each other. However, the types of relatedness that cause interference have rarely been examined carefully. In addition, past studies that have examined this issue disagree, with some providing results showing that semantic clustering does not cause interference and confusion. Reviewing the literature on working memory, a previous paper by the author indicated that psychologists have long seen visual stimulus as an important component of information processing. Researchers of vocabulary learning have also witnessed some evidence that learners do resort to visual imagery when trying to remember new words. Based on such psychological and applied linguistic research, previous research by the author revealed that visually related items may cause confusion despite the lack of semantic connection. Conversely, visually controlled, semantically related items do not seem to cause confusion. This paper presents the follow-up study, examining the learning of semantically related abstract words that do not have concrete visual images. No evidence to indicate any confusion in the learning of such items was obtained. This supports the working hypothesis that the impeding effect of semantic clustering repeatedly reported in the past could partly be due to the shared visual features of semantically similar words.

Keywords: vocabulary, semantic clustering, interference, abstract words, visual imagery.

1. Introduction

The advancement of research in second-language vocabulary acquisition has led to many approaches and beliefs about how words should be taught. Among those is the view that semantically related words should not be introduced together. When the new words that are introduced simultaneously overlap semantically, they are believed to interfere with each other and impede learning (Nation, 2013).

Research on such interferences developed in the context of memory research, and experiments were conducted using the first language of the participants. With a wide range of review of the literature addressing this issue, Waring (1997) cited McGeoch and McDonald (1931), who argued that meaningful relationships among the items to be learned interfere with learning. Higa (1963) also showed that certain types of semantic relationship can have a negative influence

Please cite this article as: Ishii, T. (2017). The Impact of Semantic Clustering on the Learning of Abstract Words. *Vocabulary Learning and Instruction, 6* (1), 21–31. doi: 10.7820/vli.v06.1.Ishii

on learning. These studies, along with others, led psychologists to develop the "Interference Theory" (Crowder, 1976, cited in Tinkham, 1993), which states that people have more difficulty while learning items that share many common elements, than while learning dissimilar items.

Tinkham (1993) applied this theory in the context of second-language learning, having his participants learn words in their first language paired with a pseudoword, modeling the condition where learners of a foreign language attempt to remember the meaning of new words. He compared the efficacy of learning words from two different sets: semantically related and unrelated. In the former, the words were grouped under a semantic category such as "fruits" (*apple, orange,* and *pear*), whereas in the latter, the words were unrelated. Displaying the participants' need for a greater number of repetitions to learn the former, Tinkham (1993) argued that "Interference Theory" is applicable to second-language vocabulary learning as well.

The concept that students should not learn semantically related items simultaneously actually contradicts the intuition of many teachers and learners. Perhaps this contradiction is the reason why this subject has attracted attention. With further empirical studies supporting this concept (Tinkham, 1997; Waring, 1997), the negative influence of semantic clustering on vocabulary learning came to be widely recognized. In his book examining various teacher and learner beliefs about vocabulary learning, Folse (2004) lists "Presenting new vocabulary in semantic sets facilitates learning" as one of seven vocabulary learning myths. Labeling this statement a myth suggests that numerous people believe that grouping words semantically will improve learning efficacy, although it appears to be false. Ishii and Maruyama (2009) conducted a survey with Japanese university students to discover if learners believe this statement or not. The survey revealed that 64% of the 543 participants believed this idea, 12% disagreed, with 24% being unsure. These results suggested that learners generally accept the statement that semantically grouped words are easier to learn; a majority of them expressed their belief in it, with only a small percentage of them explicitly disagreeing with it. It therefore supports that many learners agree with the statement that presenting new vocabulary in semantic sets facilitates learning.

However, regarding the veracity of this statement, as was discussed by Ishii (2015), research in this line has not reached a consensus. Some studies observed the negative effect only among a segment of their participants (Papathanasiou, 2009), while others, particularly those conducted in classrooms rather than laboratories, revealed a positive impact of semantic clustering on learning vocabulary items (Hashemi & Gowdasiaei, 2005; Hoshino, 2010). As Nation and Webb (2011) emphasize, research in this field has not been investigated thoroughly enough in the actual teaching context. Early studies on this issue, both in L1 and L2 contexts, were predominantly conducted under significant time pressure, without providing the participants adequate time to explore their own learning strategies. As such, we have little information about how semantic clustering affects learning in genuine learning conditions. It is therefore particularly informative that some classroom-based studies have displayed different results from previous laboratory-based ones.

Given the diverse results about the impact of semantic clustering on learning words, Ishii (2015) urges the necessity for a detailed examination of the sources

of the confusion between semantically related words. Reviewing the literature on memory studies, the paper indicated that psychologists have long seen visual stimulus as an important component of information processing, offering a new perspective on the issue of semantic clustering. Baddeley and Hitch (1974) proposed a model of working memory, a system that information first goes through when people process it, which included two initial components: the "phonological loop" – processing sound, and the "visuo-spatial sketchpad" – handling visual images. Though multiple revisions of working memory have been proposed over the years, they have always included a visual component (e.g., Baddeley, 2000; Logie, 1995). This consistent inclusion of a visual component suggests that psychologists attribute special importance to visual information when explaining the nature of human memory.

In addition to the literature review on memory studies, Ishii (2015) examined the research materials used in the previous studies on semantic clustering, and revealed a significant overlap in shape among the referents of the words the participants were expected to learn. For instance, "fruit" was a common semantic category employed in the literature, and many of the fruits selected for the studies were round (e.g., *orange, apple,* and *peach*). Similarly, a significant majority of "animals" were four-legged (e.g., *dog, cat,* and *lion*), while "clothes" such as *jacket, shirt*, and *coat* also shared some physical features. Of course, being grouped under a semantic category does not necessarily yield physical commonality among the referents; however, a review of the literature revealed that several semantically grouped items shared common features like shape (see Table 1). For instance, in addition to the examples of fruits and clothes mentioned above, we discerned similarities in shape among kitchen utensils (e.g., *spoon, fork*, and *knife*) and furniture (e.g., *chair, couch*, and *desk* or *table*). Tinkham (1997) exhibited the most extreme case, using metal names (*tin, bronze, iron*, etc.), which are visually challenging to distinguish. Given such shared visual features, as well as the importance attributed to visual information in psychology literature, Ishii (2015) adopted a working hypothesis that the alleged impeding effect of semantic clustering was partly due to the shared characteristics of shapes among the referents of semantically grouped words.

Based on this hypothesis, Ishii (2015) investigated whether a negative impact of semantic clustering was observed when the clustered words were controlled so that their referents had a limited shared visual connection. A second aim of the study was to determine if a semantically unrelated word set, whose referents shared a visual feature (e.g., round objects such as *globe, watermelon,* and *ball*), had a negative influence on the learning of those words. To answer these two questions, the study compared the learning of (1) unrelated, (2) semantically related (but physically dissimilar), and (3) physically related (but semantically unrelated) sets of words. It was reported that, both on the immediate and the delayed posttests, physically related sets yielded lower average scores than the other two sets, and the difference was confirmed to be statistically significant. In contrast, the test scores from semantically related sets were not significantly different from the unrelated sets.

These findings support the hypothesis that the apparent impeding effect of semantic clustering stems from the shared physical features of the referents in

Table 1. Words Employed in Certain Earlier Studies on the Subject of Semantic Clustering

Researchers	Conclusion about semantic clustering	Semantic categories	Words used in the experiment
Tinkham (1993)	Negative	Clothes Fruits	Shirt, jacket, sweater Pear, apple, apricot, plum, peach, nectarine
Tinkham (1997)	Negative	Kitchen utensils Metal	Dish, bowl, plate Tin, bronze, iron, brass, lead, steel
Waring (1997)	Negative	Fruits	Melon, apple, strawberry, grape, peach, orange
Finkbeiner and Nicol (2003)	Negative	Animals Kitchen utensils Furniture Body parts	Cat, cow, dog, elephant, horse, lion, pig, tiger Bowl, cup, fork, frying pan, knife, pot, spoon, stove Bed, chair, couch, desk, dresser, lamp, table, television Ear, eye, foot, hair, hand, leg, nose, toe
Erten and Tekin (2008)	Negative	Animals Foods	Bat, bee, pig, fox, hen, ape, ant, cow, owl, cock, crab, wolf, seal, bear, goat, sheep, eagle, snake, shark, snail Egg, fig, leek, plum, bean, pear, salt, okra, corn, onion, olive, melon, honey, grape, garlic, pepper, carrot, radish, cherry, peanut
Papathanasiou (2009)	Negative among adult beginners, and no difference among intermediate children	Crime Nature Food Synonyms (pairs) Antonyms (pairs)	Smuggling, terrorism, forgery, mugging, trial, proof, jury, verdict, witness, bribery Cape, peninsula, cave, tributary, valley, gorge, stream, estuary, ridge, summit Lamb, herring, veal, ham, cod, trout, prawn, shrimp, squid, lobster Torment, torture/jab, punch/spat, quarrel/gleam, twinkle/boredom, tedium Ebb, flow/gloom, glee/certitude, doubt/loyalty, treason/ poverty, prosperity
Hoshino (2010)	Positive	Various word pairs	Moth, wasp/asthma, diabetes/calf, chick/borough, province/solicitor, astronomer
Hashemi and Gowdasiaei (2005)	Positive	Materials not disclosed	

Note: Some of these studies included categories that were not related to semantic connections (e.g., *homonyms* in Papathanasiou [2009]). Such categories are not listed in this table.

semantically grouped words. A question then naturally emerges – if the source of confusion lies in the visual image of the referents, what happens when no distinct visual image is available? More specifically, what is the impact of semantic clustering in the case of the words without a tangible visible referent? This study was designed to address this question in order to further examine Ishii's (2015) hypothesis.

The following sections of this paper will first discuss the role of visual imagery in the learning of vocabulary items in a second language to provide additional background to the hypothesis. It will then describe the method undertaken to compare the learning efficacy of (1) unrelated and (2) semantically related sets of abstract words, followed by the results of the experiment.

1.1 Role of visual imagery in learning words

In addition to the memory models mentioned above, other psychological studies have also attributed an importance to human memory. For instance, Paivio (1969) proposed the "Dual Coding theory" that argues for the important role visual imagery plays in human memory. According to this theory, we elaborate our learning through the creation of visual images and verbal associations. After several decades, this theory still serves as a basis for the understanding of human cognition (Paivio, 2013). The theory suggests that learning outcomes are better when people successfully create visual images in their minds, as well as when they think about their meanings and create associative networks around them.

In line with this theory, the use of visual imagery has long been supported as a useful mnemonic technique (Nation, 1990). Second-language vocabulary researchers have also witnessed learners resorting to visual imagery when trying to remember new words, of which some examples are provided later.

In his study investigating incidental vocabulary learning through reading, Yoshii (2006) examined the gain in the meaning retrieval with four different glosses: L1 text only, L2 text only, L1 plus picture, and L2 plus picture. As a result, while no difference in vocabulary gain was observed between L1 and L2 text only glosses, the addition of pictures to the glosses contributed significantly to the incidental learning of vocabulary through reading. This suggests that the use of pictures assists the students in forming visual mental representations that facilitate the retention of the word meaning.

Boers, Lindstromberg, Littlemore, Stengers, and Eyckmans (2008) and Boers, Piquer Píriz, Stengers, and Eyckmans (2009) investigated the effect of pictorial elucidation when learning new idiomatic expressions. The studies revealed that learners retain the meanings of newly learned idiomatic items better when they are presented with visual images. Though there was no impact for the word forms, such presentations at least improved the learning of word meanings.

Farley et al. (2012) examined if the meaning recall of words improved in the presence of imagery, and found that only the meaning recall of abstract words improved, while that of concrete nouns did not. A possible interpretation of this finding is that, in the case of concrete nouns, most learners can naturally produce visual images in their mind and use them to remember the words. Therefore, the

additional visual images in the learning material do not affect the learning outcome, since they are already present in their mind. However, in the case of abstract nouns, since it is often difficult for learners to create images independently, the presentation of imagery helps them retain the meaning of the words they are trying to learn.

All these studies, though they focus on different issues, suggest that second-language learners use visual imagery when trying to remember new lexical items. Studies on semantic clustering typically ask the participants to learn new lexical items, and we can expect the learners under such experimental conditions to resort to visual imagery. It is therefore reasonable to assume that the overlap in the shape of the referents of the target words makes it more difficult for the learners to commit the words to memory, as such words lead to similar visual imagery the learners create in their mind.

Investigations into the learning of semantically grouped words most frequently have been conducted using nouns referring to concrete objects, though some studies have included abstract nouns (Papathanasiou, 2009). However, to the best of my knowledge, no study has specifically investigated the influence of semantic clustering while learning abstract nouns that do not have a tangible object as their referents. The current study was designed to examine this issue, and was conducted under the following hypothesis and research question.

Hypothesis:

The apparent negative effect of semantic grouping is partly due to the shared features of the shape of the referents.

Research question:

Is the efficacy of learning of word meaning influenced by grouping abstract nouns into semantic sets?

Based on the hypothesis above, it was predicted that grouping abstract words into semantic sets would not influence the efficacy of learning negatively, as abstract words do not have a referent with a concrete shape. The hypothesis, however, does not allow us to predict whether such clustering has a positive influence on learning. The following sections of this paper will describe the investigation of this hypothesis and will present the results.

2. Method

This study was conducted with 62 Japanese university students, who were first-year Economics majors. The purpose of the study – to contribute to the understanding of how people learn vocabulary – was explained to the participants. They were also aware that they would be learning pairs of Japanese meanings and nonwords. However, no further details about the grouping of the words were disclosed until the completion of the experiment. During the experiment, the participants learned pseudowords paired with meanings (a word given in their native language, Japanese), under two categories: "unrelated" and "semantically related." The former category grouped five abstract words that had no obvious semantic relationships, whereas the latter was a collection of five abstract words grouped under

the following concepts: *personality traits*, *feelings*, *talking*, and *crime*. The English translations of the Japanese words used in the experiment are displayed in Table 2.

Each category contained four sets of five pairs of pseudowords and Japanese words, namely, 20 pairs in each category, totaling 40 pairs that needed to be learned. All the participants learned the words in both categories, and the difference in performance in both categories was compared within subject.

The pseudowords used in this experiment were generated using a computer program called Wuggy (Keuleers & Brysbaert, 2010) that produces pseudowords conforming to English spelling rules. As the software produced pseudowords in accordance with the number of syllables as well as vowel–consonant combinations of the English words entered, it allowed us to control the length and phonological pattern of the pseudowords used in the experiment. The materials were prepared such that the pseudowords included in both categories were balanced in terms of the number of syllables and letters. In addition, in order to account for the possibility that one set of pseudowords might be easier to learn, the pairings were varied.

The participants were asked to memorize one set of five pairs, displayed on a screen, in 40 seconds. This time limit was determined based on the results of a pilot study conducted with three students with a similar background to the participants of this study. Immediately following the 40-second learning session, the participants took a meaning recall test where they demonstrated their memory of the meanings represented by the pseudowords (Test 1). Repeating this learning and testing cycle eight times, they were exposed to all 40 pairs of pseudowords and meanings. They then spent 20 minutes on a class activity that was not part of this experiment, and took another test (Test 2) without any prior notification. Test 2 presented all the 40 pseudowords alphabetically, and the participants were asked to write down the meaning associated with each.

3. Results

Tables 3 and 4 present the results of the two tests in the study. In Test 1, in both the categories, the average score was about 60% (Table 3). However, the score

Table 2. Japanese Meanings Employed in the Study

Category	Nature of the connection	Meaning represented in Japanese
Unrelated	There is no obvious connection among the words	(1) Difference, summary, speed, concern, view (2) Power, play, remaining, help, attention (3) Cause, patience, situation, grammar, determination (4) Memory, variety, completion, role, question
Semantic	Words (meanings represented in Japanese) are categorized under the following themes: (1) personality traits, (2) feelings, (3) talking, (4) crime	(1) Kindness, brightness, honest, bravery, intelligence (2) Happiness, sadness, anger, surprise, worried (3) Conversation, negotiation, instruction, suggestion, agreement (4) Statement, guilt, testimony, arrest, release

Table 3. Results from Test 1 (*N* = 62, Possible max = 20)

	Max	Min	Mean	SD	SEM
Unrelated	20	3	13.00	4.53	0.58
Semantic	20	1	13.42	4.62	0.59

Table 4. Results from Test 2 (*N* = 62, Possible max = 20)

	Max	Min	Mean	SD	SEM
Unrelated	14	0	3.48	3.49	0.44
Semantic	13	0	3.85	3.31	0.42

declined greatly after 20 minutes of distraction, which indicates poor retention of the learned meaning (Table 4).

As the tables show, the observed difference in the mean scores of the two categories was minor in both tests, and paired *t*-tests did not reveal a statistical significance for either of the tests ($t = 1.15$ and $p = 0.25$ for Test 1, and $t = 1.26$ and $p = 0.21$ for Test 2).

4. Discussion and conclusion

The data presented above display no advantage or disadvantage of grouping abstract nouns into semantic categories. The current study was conducted under the hypothesis that the seemingly negative effect of semantic grouping is partly due to shared features in the shape of the referents. The results are consistent with this hypothesis, as the semantic grouping of abstract words, where the referent has no concrete shape, did not exhibit any impeding effect on the participants' memory. The semantic relatedness among the target words did not have a positive impact either.

However, the study presented in this paper has some limitations. First, since this was a laboratory-type research, and not a classroom embedded one, the participants had a tight time limitation to learn the words, and thus were unable to utilize their learning strategies. Second, it employed nonwords as the learning target. This was to avoid the risk of conflating the learning of target items with prior knowledge of the words. However, the disadvantage was that the participants did not perceive the value of memorizing these words, except to know that this study would contribute to knowledge about vocabulary instruction. If a similar study is conducted in a genuine classroom setting, where students learn words they perceive are important, with abundant time to review the target words, different results might be obtained. Third, the study only tested the meaning recall of the learned words. It is quite possible that the interrelationship between the target items have different effects on different sets of skills, such as receptive and productive knowledge. Therefore, it should be noted that investigating different types of knowledge might well provide a different result.

We should also be aware that there are learners with different cognitive styles. Boers et al. (2009) addressed this issue in their study of the influence of pictorial elucidation on the recollection of idioms. They argue that there are "high imagers"

and "low imagers" who resort to mental imagery to different extents. Although they did not reach a solid conclusion, they suggest that pictorial presentation might influence high and low imagers in different ways. In the current study, the extent to which each participant resorted to mental imagery is unknown. Moreover, it is very likely that participants with different cognitive and learning styles perform differently, which is an interesting subject that can be addressed in a future study.

Although we need to be aware of these limitations, this study demonstrates how fragile the concept of the negative impact of semantic clustering is. Namely, it showed how the selection of words to be grouped semantically can greatly affect learning efficacy. In some earlier studies, the concept of "semantic clustering" has been discussed without defining what that meant. As Ishii (2015) previously argued, words can be connected semantically under different types of relationships and to different degrees. For example, under the category of "musical instruments," it is intuitively unlikely that the word *cymbal* is as closely located, in our mental lexicon, to the word *piano* as it probably is to the word *organ*. Thus, the term "semantic clustering" can refer to a variety of connections between words. However, the term has not been defined clearly enough to be considered an operational variable. In order for this line of research to reach a solid conclusion on this subject, future research needs to be designed with detailed consideration about the source and the process of the interference.

The current study, as well as my previous research, was conducted under the hypothesis that the shared features of the shape of the referents of the target words were a source of confusion. Namely, it proposes that some confusion emerges when learners resort to visual imagery while attempting to memorize the meaning of new words. The data obtained in the earlier study (Ishii, 2015) strongly support this hypothesis; also, the results presented in this paper are in accordance with this supposition. Further research, with insight about the cognitive processes of learners, conducted under a variety of conditions, and encompassing different types of knowledge, should reveal more about the nature of the relatedness among words that has an impact on the learning outcome.

References

Baddeley, A.D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417–423. doi:10.1016/S1364-6613(00)01538-2

Baddeley, A.D., & Hitch, G.J. (1974). Working memory. In G.A. Bower (Ed.), *Recent advances in learning and motivation* (Vol. 8, pp. 47–89). New York, NY: Academic Press.

Boers, F., Lindstromberg, S., Littlemore, J., Stengers, H., & Eyckmans, J. (2008). Variables in the mnemonic effectiveness of pictorial elucidation. In F. Boers & S. Lindstromberg (Eds.), *Cognitive linguistic approaches to teaching vocabulary and phraseology* (pp. 189–116). Berlin, Germany: Mouton de Gruyter.

Boers, F., Piquer Píriz, A., Stengers, H., & Eyckmans, J. (2009). Does pictorial elucidation foster recollection of figurative idioms? *Language Teaching Research, 13*(4), 367–388. doi:10.1177/1362168809341505

Crowder, R.G. (1976). *Principles of learning and memory.* Hillsdale, NJ: Lawrence Erlbaum.

Erten, I.H., & Tekin, M. (2008). Effects on vocabulary acquisition of presenting new words in semantic sets versus semantically unrelated sets. *System, 36*(3), 407–422. doi:10.1016/j.system.2008.02.005

Farley, A.P., Ramonda, K., & Liu, X. (2012). The concreteness effect and the bilingual lexicon: The impact of visual stimuli attachment on meaning recall of abstract L2 words. *Language Teaching Research, 16*(4), 449–466. doi: 10.1177/1362168812436910

Finkbeiner, M., & Nicol, J. (2003). Semantic category effects in second language word learning. *Applied Psycholinguistics, 24*, 369–383. doi:10.1017/S0142716403000195

Folse, K.S. (2004). *Vocabulary myths.* Ann Arbor, MI: The University of Michigan Press.

Hashemi, M.R., & Gowdasiaei, F. (2005). An attribute-treatment interaction study: Lexical-set versus semantically-unrelated vocabulary instruction. *RELC Journal, 36*(3), 341–361. doi:10.1177/0033688205060054

Higa, M. (1963). Interference effects of intralist word relationships in verbal learning. *Journal of Verbal Learning and Verbal Behavior, 2*(2), 170–175. doi:10.1016/S0022-5371(63)80082-1

Hoshino, Y. (2010). The categorical facilitation effects on L2 vocabulary learning in a classroom. *RELC Journal, 41*(3), 301–312. doi:10.1177/0033688210380558

Ishii, T. (2015). Semantic connection or visual connection: Investigating the real source of confusion. *Language Teaching Research, 19*(6), 712–722. doi:10.1177/1362168814559799

Ishii, T., & Maruyama, Y. (2009). How mythical are 'Vocabulary Myths' among Japanese learners of English? *The Journal of Rikkyo University Language Center, 21*, 23–31.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*(3), 627–633. doi:10.3758/BRM.42.3.627

Logie, R.H. (1995). *Visuo-spatial working memory.* Hove, UK: Laurence Erlbaum Associates.

McGeoch, A., & McDonald, W.T. (1931) Meaningful relation and retroactive inhibition. *American Journal of Psychology, 43*(4), 579–588. doi:10.2307/1415159

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* New York, NY: Newbury House.

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

Nation, I.S.P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston, MA: Heinle Cengage Learning.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review, 76*(3), 241–263. doi:10.1037/h0027272

Paivio, A. (2013). *Mind and its evolution: A dual coding theoretical approach*. New York, NY: Psychology Press.

Papathanasiou, E. (2009). An investigation of two ways of presenting vocabulary. *ELT Journal, 63*(4), 313–322. doi:10.1093/elt/ccp014

Tinkham, T. (1993). The effect of semantic clustering on the learning of second language vocabulary. *System, 21*(3), 371–380. doi:10.1016/0346-251X(93)90027-E

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research, 13*(2), 138–163. doi:10.1191/026765897672376469

Waring, R. (1997). The negative effects of learning words in semantic sets: A replication. *System, 25*(2), 261–274. doi:10.1016/S0346-251X(97)00013-4

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology, 10*(3), 85–101. Retrieved from http://llt.msu.edu/vol10num3/yoshii/

Vocabulary Learning and Instruction
Volume 6, Issue 1, November 2017
http://vli-journal.org

Responding to Research Challenges Related to Studying L2 Collocational Use in Professional Academic Discourse

Birgit Henriksen and Pete Westbrook
University of Copenhagen
doi: http://dx.doi.org/10.7820/vli.v06.1.Henriksen

Abstract

This study describes the English collocational use of non-native university teachers from two different disciplines lecturing in an English-medium instruction context at the University of Copenhagen (UCPH). The primary focus is on how we addressed the research challenges involved in identifying and classifying collocations used by L2 speakers in advanced, domain-specific oral academic discourse. The main findings seem to suggest that to map an informant's complete collocational use and to get an understanding of disciplinary differences, we need to not only take account of general, academic and domain-specific collocations but also need to cover the full range of both lexical and grammatical collocations.

1. Introduction and Background to the Study

1.1 The Collocational Use of Non-Native Lecturers Teaching in an EMI Context

The past 20 years or so has witnessed rapid growth in internationalisation at institutions of higher education in Europe, especially northern Europe, not least in Denmark. As a result, English has increasingly become the lingua franca of academia, as more and more degree programmes are run through the medium of English. Consequently, university teachers who are non-native speakers of English are asked to lecture, tutor and supervise students in English and are thus expected to perform effectively in professional academic discourse in their L2.

To meet the pedagogical challenges presented by this situation, the University of Copenhagen implemented a language policy based on the parallel language use of Danish and English. To support this internal language policy, as well as to ensure quality in educational programmes and research, the University established the Centre for Internationalisation and Parallel Language Use (CIP) in 2008 as a research, competence development and resource centre. Part of CIP's remit was to provide language training and language certification of tenured academic staff. As a result, the performance-based Test of Oral English Proficiency for Academic Staff (TOEPAS) certification procedure was developed by the Centre and is used for assessing whether university lecturers have sufficient oral proficiency for

Please cite this article as: Henriksen B. and Westbrook P. (2017). Responding to Research Challenges Related to Studying L2 Collocational Use in Professional Academic Discourse. *Vocabulary Learning and Instruction, 6* (1), 32–47. doi: 10.7820/vli.v06.1.Henriksen.Westbrook

coping with the communicative demands of English-medium instruction (EMI) (see http://cip.ku.dk/english/certification/).

This paper reports on a study of university teachers' L2 collocational use when lecturing in an EMI context and is based on data from the TOEPAS certification. Fifteen of the mini-lectures from three different academic domains were transcribed for research purposes. The authors set out on a small exploratory project which aimed to describe the lecturers' *overall collocational use* across all the collocational sub-types that would be expected to be found in academic language, that is, domain-specific, academic and general collocations. The aim of this original study was to test whether there were any parallels between the lecturers' level of English proficiency as assessed in the certifications compared to the frequency and appropriateness of collocational use across the three types of collocations mentioned above. Moreover, we wanted to identify possible similarities and differences in collocational use across different academic domains. By including all three types of collocations in our analysis, we would be able to generate a general score of informants' collocational use which could be used to correlate with other measures like fluency, certification score and vocabulary size measures.

The aim of the present paper is to highlight the research challenges inherent in investigating collocational use in oral, domain-specific language, both in general, but more specifically across the three sub-types mentioned above; an area which seems to have been subject to very little research as far as we are aware. We will present suggestions for dealing with these challenges and the effect of the choices made on the results. Although this is a small-scale study, we believe it can contribute to knowledge about collocational use in academic discourse, particularly on how this could and should be researched.

1.2 The Importance of Domain-Specific, Academic and General Collocations in Academic Discourse

Collocations are frequently recurring two-to-three-word syntagmatic units (e.g. *soft noise, tolerance for*). In the research literature, collocations are defined as a subset of formulaic sequences, distinct from other types of formulaic sequences such as lexical bundles, idioms, and pragmatic phrases (Nattinger & DeCarrico, 1992). Handl (2009) sums up collocations thus: "We can conclude that collocations are *conventionalized* recurring word combinations exhibiting more or less *restrictedness*, more or less *semantic opacity* and a certain degree of *predictability* for native speakers (…). So, two words that collocate are not governed by semantic compatibility, but rather by lexical restriction, that is, by the norms of the language" (Handl, 2009, p. 70, our italics). As examples of this, we can take the combinations *strong coffee* and *powerful car* as the preferred collocates, rather than *powerful coffee* and *strong car.*

Collocations can consist of different grammatical and lexical constituents. A lexical collocation is the type of construction where a verb, noun, adjective or adverb forms a predictable connection with another word from these word classes, as in *completely satisfied* (adverb+adjective), *excruciating pain* (adjective+noun) and *commit suicide* (verb+noun). A grammatical collocation is a

type of construction where for example a verb or adjective must be followed by a particular preposition, as in *depend on* (verb+preposition) or *afraid of* (adjective+preposition).

Mastery of formulaic sequences, including collocations, has been described as a central aspect of communicative competence, enabling the native speaker to process language both fluently and idiomatically and to fulfil basic communicative needs. Nation (2001, p. 318) concludes that "all fluent and appropriate language use requires collocational knowledge." It has also been argued that collocational use is equally important for L2 learners (Barfield & Gyllstad, 2009; Henriksen, 2013). Nevertheless, this is a language phenomenon which is said to be acquired late and which is often not mastered very well even by reasonably competent L2 language learners (Nesselhauf, 2005; Laufer & Waldman, 2011; Henriksen, 2013). For this reason, collocational proficiency may be seen as a quality feature of advanced language use, for example, for academic lecturers like our informants who are operating in a highly demanding professional setting in their L2.

The main reason for focusing on collocations in relation to EMI language use is that collocations typically have a highly referential function (Howarth, 1998), as opposed to the discourse or pragmatic functions of other types of formulaic sequences. Moreover, they tend to be very genre specific. As such, collocations are often seen as characterising technical sub-languages (Ananiadou & McNaught, 1995), that is, languages from different study domains. Similarly, mastery of collocations may be a hallmark of certain types of academic writing which emphasize clarity, precision and lack of ambiguity (Howarth, 1998). As mentioned, even very advanced L2 users seem to have problems with using collocations and, apart from leading to unfortunate misunderstandings, advanced non-native speakers' collocational deviations may signal a lack of academic expertise (Henriksen, 2013, p. 37).

In the research literature, vocabulary is divided into general, academic and technical language (Coxhead, 2000; Hwang & Nation, 1995; Xue & Nation, 1984). However, this research focuses very much on single word items, with little or no research on the same distinctions applied to collocations. Hwang and Nation (1995) found that vocabulary in non-fiction texts can be divided into high frequency (or general service) vocabulary, sub-technical (or academic) vocabulary, technical (or domain-specific) vocabulary, and low-frequency vocabulary (based on Nation, 1990, p. 19). How these different categories of items combine often characterises different kinds of professional academic discourse. Because of the complexity of professional academic discourse found in the certification data, the authors also found it necessary to make the distinction between general, academic and domain-specific collocations in line with the single word item distinctions brought out by Hwang and Nation (1995). A study by Westbrook (2015), who investigated the role of collocations for fluency in the same data set as used in the present paper, found significant differences in results depending on whether domain-specific collocations were included in the calculations or not. This seems to be in line with differences in the density of domain-specific single words between disciplines found by Chung and Nation (2004), and would therefore present a case for also distinguishing between general, academic and domain-specific collocations in our study of academic discourse.

1.3 The Research Issues Addressed

As pointed out by Henriksen (2013), virtually all the previous studies on collocations have dealt with general collocations with only very few, if any, tackling academic and technical, that is, domain-specific collocations. In addition, most research on collocations has focused on written or corpus data. Moreover, these studies have often been limited to one specific type of collocation, typically verb+noun or adjective+noun collocations. By including a range of collocational sub-types used in oral, academic discourse, we have been expanding the range of research focus. The lack of studies dealing with various types of collocations, however, meant that there were very few previous comprehensive research models to draw on in our analysis of our informants' overall collocational use.

The extensive pilot phases, which have been reported at various conferences (Complexity and Idiomaticity, Stockholm University, June 2012; EIE Conference, Copenhagen 2013; SDU SELC Conference, Odense 2013; PhD Applied Linguistics (Lexical Studies) annual conference, Cardiff University, Wales 2014; and AILA World Congress 2014), highlighted a range of methodological problems, both in relation to deciding how to operationalise the distinction between domain-specific, academic and general collocations and how to identify these three types of collocations in the data. Moreover, our preliminary studies showed that the internal structure and complexity of the individual collocations seemed to differ across the three types of collocation, creating analytical challenges in relation to how to deal with more complex, embedded collocations and what to include in the quantification of the individual collocations. In addition, the analysis of oral data created its own challenges, for example, in relation to split collocations, where the distance (span) between node and collocate (Nattinger & DeCarrico, 1992) was in some cases quite considerable (see Section 3.2).

All these research challenges, which needed to be overcome in order to carry out robust research on collocations in domain-specific academic discourse, prompted us to write the current paper. This paper is therefore concerned with identifying the methodological problems and investigating how results might differ according to the methodological choices made. The study includes an analysis of 12 CIP TOEPAS lectures from two academic domains. The first two research questions are related to the research procedure itself:

(1) What challenges are there in trying to describe collocational use across general, academic and domain-specific types?
(2) How might these challenges be met?

On the basis of our suggestions for solving these research challenges, we will present the results of the analysis of the 12 lecturers' collocational use to answer the last three research questions. The focus here is on exploring potential differences across the different collocational types and across the two academic domains:

(3) What characterises academic collocational use across lexical and grammatical collocations?

(4) What characterises academic collocational use across the general, academic and domain-specific categories of collocations?
(5) What characterises collocational use across different academic domains?

2. Previous Research on Collocations

The initial research challenge for any study on collocations is to decide on a method for identifying and delineating the types of collocations to be explored. This question will initially be discussed in a short literature review on collocational research in this section. The final research choices made will be outlined in the actual presentation of the study itself in Section 3.

Two main approaches have been adopted by researchers to identify collocations in a given text or corpus: the *frequency-based approach* and the *phraseological approach*. The frequency-based approach is associated with computer-based searches of large language corpora. These searches involve identifying words that occur within a short span, usually four words, either side of a headword, or "node." If the node occurs together with another word or words within this span "at a frequency greater than chance would predict, then the result is a collocation" (Nattinger & DeCarrico, 1992, p. 20). Thus, collocations are not necessarily contiguous, although they can be. They can also be realised in different lexical combinations. The collocation *strong argument*, for example, can be realised as: *it is a strong argument, he argued strongly for, the argument is a strong one*, and so on. Frequency criteria alone, however, will not necessarily yield all possible collocations. The phraseological approach employs a manual identification based on more intuitive syntactic and semantic analysis of word combinations and is helpful in defining collocations more precisely. Generally, researchers have adopted a combined approach (Barfield & Gyllstad, 2009).

The next fundamental challenge in any study of collocations is deciding what types of word combinations to include as collocations. One question is whether or not to include compounds. Granger and Paquot (2008) argue for excluding them in any analysis of collocations, partly because of their "uncertain status as single or multi-word units" (Granger & Paquot, 2008) (e.g. *good will, good-will, goodwill*), and partly because of their somewhat fixed status (e.g. *black hole, goldfish, blow-dry*). However, domain-specific discourse tends to be compound noun heavy. As Moon (1997) states: "compounds typically denote and have high information content – often because they are technical terms or have specific reference" (Moon, 1997, p. 56). Such compounds tend to be more flexible than Granger and Paquot claim; there is, after all, to take an example from our data, a difference between a "mouth speculum" and an "ear speculum." In addition, they are also included in such collocational reference works as the Oxford Collocation Dictionary (OCD) and in Pearson's Academic Collocation List (ACL). In a study of collocational use including general, academic and domain-specific collocations, the inclusion of compounds would therefore seem to ensure that the full range of collocational types would be represented.

A third point to consider is related to the distinction between lexical and grammatical collocations outlined in Section 1.2. Most research studies on collocations have investigated a limited range of collocational types, often with a focus on lexical collocations (Henriksen, 2013). The broad categories of lexical

and grammatical collocations can be further broken down into different structural sub-types in relation to the different word class constituents they include. As mentioned above, most research on collocations seems to have focused only on the adjective+noun and the verb+noun constructions (Henriksen, 2013). If the aim is to describe informants' overall collocational use, both lexical and grammatical collocations types and the different sub-types would, however, need to be included in the analysis. As will be documented later, academic language includes a range of structural combinations of both lexical and grammatical collocations, for example, adverb+verb (*gradually realize*), noun+preposition (*idea of*), verb+adjective (*be surprised*), verb+noun (*take time*) and verb+preposition (*ask about*).

Apart from the basic structural collocational types mentioned above, there are also other more complex collocational combinations, "nested collocations" (Frantzi & Ananiadou, 1996), which consist of a collocation in combination with additional word class constituents. These complex collocations, which can include both lexical and grammatical sub-types, are typically found in domain-specific discourse but have not been the focus in collocation research in general so far. The types found in our study include the nested collocational combinations shown in Table 1.

The final challenge is related to the distinction between general service (*spend time, good idea, hear about*), academic (*strong argument, reliable data*) and domain-specific (*exponential bounds, be contained*) collocations, which will be one of the main topics for the rest of this paper. Due to the existence of collocations, dictionaries such as the OCD, and with the recent publication of Pearson's Academic Collocations List (Ackermann & Chen, 2013; http://pearsonpte.com/research/academic-collocation-list/), general and academic collocations are relatively simple to identify and delineate in an objective way, whereas identifying and classifying the domain-specific collocations is far more challenging. A few studies have investigated mathematical and medical collocations (e.g. Haag, Heppt, Stanat, Kuhl, & Pant, 2013; Herbel-Eisenmann, 2002; Méndez Cendón, 2004), but it is very difficult to find an objective method for identifying and classifying domain-specific collocations which can be used across different academic domains. Often domain-specific or technical language is associated with the use of specific "technical terms" or "phrases," for example, highlighted in domain-specific dictionaries or terms lists, but no standard method for establishing these inventories have been developed, and many academic fields have not developed or published lists of domain-specific collocations.

Table 1. Lexical and Grammatical "Nested Collocations"

	Collocational combination	Examples from the data
Lexical nested collocations	adjective + collocation	*complicated differential equation*
	adverb + collocation	*purely algebraic definition*
	collocation + collocation	*finitely generated abelian groups*
	collocation + noun	*solutions to this equation*
	proper noun + collocation	*Heisenberg's matrix mechanics*
	verb + collocation	*have a continuous function*
	adjective + collocation +collocation	*compact Hausdorff topological space*
	noun + collocation	*capilliary refill time*
Grammatical nested collocations	collocation + preposition	*continuous functions on*
	preposition + collocation	*in a physical world*

3. Methodology

3.1 Data Source

The TOEPAS test is a high-stakes oral assessment and takes the form of a 20-minute simulated mini-lecture in English, carried out at the University of Copenhagen (http://cip.ku.dk/english/certification/). Teachers are assessed on a 5-point holistic scale based on five dimensions (pronunciation, grammar, lexis, fluency and interaction skills). Scores 3, 4 and 5 are certified, while 1 and 2 are not certified. Teachers come from different faculties and, as part of the test procedure, are given formative feedback as well as their overall score. Each mini-lecture is videoed and CIP now has a databank of around 400 certifications. The data for this paper have been collected from 12 lecturers from two different departments: the Department of Large Animal Science (LAS) and the Department of Mathematics (Maths). The scores for the six LAS informants were 5, 4, 3, 3, 2, 2, and for the six Maths informants 5, 4, 4, 4, 3, 2, respectively.

3.2 Identifying Collocations in Our Data Electronically or Manually?

One possible method for identifying potential general, academic and domain-specific collocations in our data set using the frequency approach is to apply corpus-based tools such as WordSmith or AntConc. Among other things, these tools allow the user to list words in the text in order of frequency, show frequent word partnerships present in the texts, and display concordance lines for a particular word, with the additional option of sorting the lines according to the co-text to the left or right of the node word. Thus, the programmes enable the identification of clusters of word partnerships which form potential (domain-specific) collocations. In addition, if the data set is large enough, both Wordsmith and AntConc have the function to compare the data texts with another (general) corpus in order to identify which word occurs more frequently in the data than in a general data set (the so-called "keyword analysis"). The advantages of using electronic corpus-based tools are that they are objective, apply clear criteria and are less time-consuming. However, utilizing such electronic tools was not a viable option for us, as our corpus of 12 mini-lectures was simply too small and we had no access to a reference corpus for the separate domains. Moreover, an electronic search would not have been able to cope with the long span in "split" or "fragmented" collocations (Pulverness, 2007). These are collocations where the span between the node and collocate is quite long, for example, *the path that a ball rolling down the hill would take.* The varying lengths of the domain-specific collocations (typically from two- to five-word units) would also have made it difficult to work electronically. Finally, we would still need to manually identify and discard the non-meaningful lexical bundles, for example, *and the*, *as we,* etc. (Martinez & Schmitt, 2012) from the potential collocations.

Despite being more subjective and time-consuming, a manual search therefore had the advantage of allowing us to catch all the potential collocations in the data transcriptions, and take account of the problem of "split collocations" as in the example mentioned above. Many of these constructions were found in the oral data, due to the specific features of the oral discourse mode. The manual approach also ensured that we would identify and count two (or more) collocational

pairings on the same headword. These were counted as two separate collocations. For example, *use a different method* was counted as two collocations (*use a method* and *different method*).

Thus, our final approach was based on the manual/phraseological approach. This involved each of the authors identifying all potential collocations manually simply by going through each of the 12 transcripts independently and underlining all the potential collocational units. To ensure that as many potential collocations as possible could be underlined, especially the domain-specific ones that we had no expert knowledge of, any unknown combination which in any way could be construed as a collocational unit was included in the initial identification procedure. We then compared those we had found individually. In most cases, we found the same, but there were also a certain number found by one researcher and not the other, which demonstrates the advantage of having two coders working on the same data, especially in order to "empty" the data in an attempt to identify as many potential collocations as possible. This initial identification stage is described as stage one in Figure 1.

3.3 Categorising and Delineating General, Academic and Domain-Specific Collocations in Our Data

Based on extensive pilot coding, we finally settled on a combined method, adopting a three-tier approach of exclusion: first classifying the academic collocations using Pearson's ACL (stage 2), then classifying the general collocations using the OCD (stage 3) and finally identifying the domain-specific collocations by checking them through a Google search (stage 4). These gradual

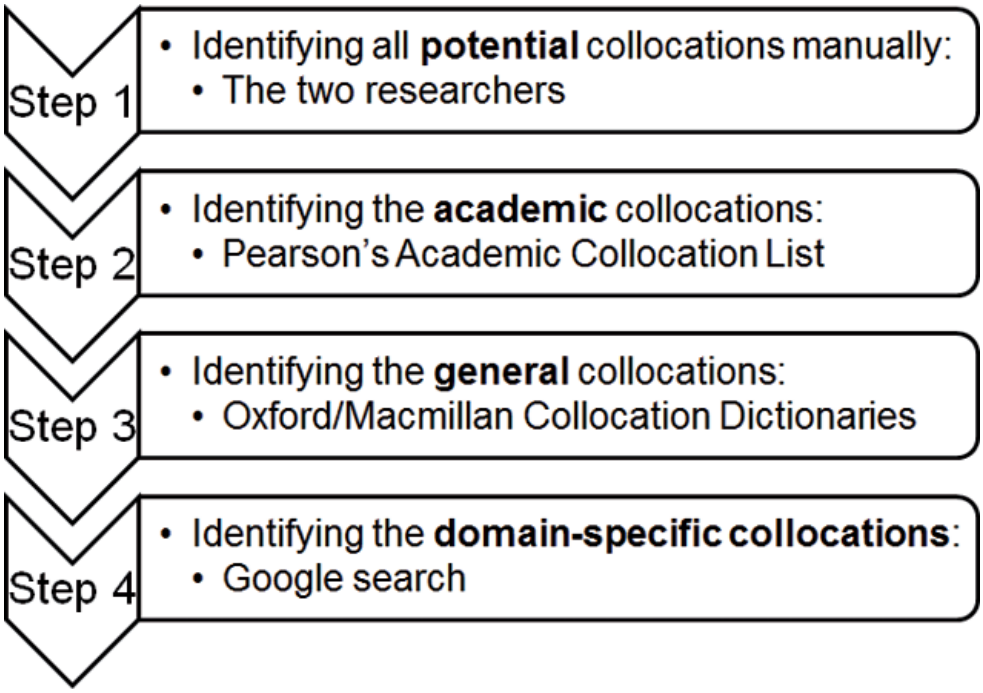

Figure 1. The Four-step Analytical Approach Adopted.

stages of exclusion are illustrated in Figure 1. The more specific issues related to decision stage 4 for identifying the domain-specific types will be outlined in more detail later.

As mentioned above, lists of general collocations (e.g. the OCD) and lists of academic collocations (e.g. Pearson's list) were available tools for identifying and classifying the general and academic collocations. Unfortunately, the same type of research-based tool for identifying domain-specific collocations within our specific domains does not exist. One method of checking for domain-specific collocations could therefore be using expert raters (Chung & Nation, 2004) within a certain academic domain. In an initial pilot, three native speaker informants from LAS were asked to identify potential domain-specific collocations in the data from their own department. However, this procedure proved to be very time-consuming, both in relation to finding the informants, training them to understand the concept of collocations and having them deal with extensive lists of potential collocations. On top of this, the fact that it was also very difficult to obtain consensus from all three native speaker experts was another deciding factor for excluding the use of native speaker raters for the identification of domain-specific collocations. Thus, we decided to investigate other more objective methods of determining domain-specific collocations.

Initial pilot coding of the domain-specific collocations using different technical dictionaries and term lists showed that it would be very difficult to find a reliable procedure that could be applied across academic domains. The quality of the resources and the nature of the lexical entries created differences in the coding that would have seriously biased the results. Some resources primarily included single word items and noun+noun or adjective+noun collocations and very few listed collocations containing verbs or adverbs. The Google search procedure, that is, a frequency-based identification method, was therefore used to identify domain-specific collocations. This was done by putting the potential collocation itself in quote marks and then combining this search with a consistent phrase relating to each of the departments in turn. After several trial runs, it was decided to use the word "maths" for the mathematics informants and "animal science" for the informants from the LAS department, and the cut-off point was set at 5,000 hits. From the resulting combinations found, we checked for lemmas and added those in as necessary; in addition, we excluded any combinations which were clearly not domain-specific collocations (e.g. *background story*).

Finally, any collocations included in stage 1 as potential collocations that were not coded as academic, general or domain-specific were not counted as collocations at all and were excluded from the inventory of collocations found in our data set.

Hwang and Nation (1995), focusing on individual items, have stressed that we cannot operate with clear-cut distinctions between the different vocabulary types, but are dealing with a continuum with fuzzy and arbitrary divisions. This, however, raises the question of how and where the arbitrary dividing line can be most sensibly drawn and how to decide the classification of the individual collocations. Working with this three-tier procedure of gradual exclusion, we still encountered some problems related to the fuzzy boundaries between the

categories. For the collocations coded as general, some of these may be pre-technical (e.g. *complex numbers, relations between*), while others may be crypto-technical, that is, polysemous word with one meaning clearly related to specific fields of study and where the technical meaning is not likely to be known by a lay person (Fraser, 2006), for example, *product of*, *be unique*. This is a well-known issue for "the language of mathematics," that is, the fact that maths words can have multiple meanings, for example, *true* in the general sense and a second more technical meaning of the word. The setting of the boundaries between the categories is also an issue when dealing with the collocations that were coded as domain specific. Some may belong to a larger group of study domains, for example, maths and animal science, the STEM areas: science, technology, engineering and maths (e.g. *do this procedure, subgroup of*), but they are not frequent enough across a range of academic domains to be listed as academic collocations, for example, in the Academic Collocations List.

3.4 Quantifying Our Results and the Issue of "Nested Collocations"

A further research issue to highlight is the question of what counts as one collocation. This is important as it is related to the quantification of the results. Technical language is often characterised by terms that are made up of multi -layered, nested collocations (described in Section 2), where an adjectival specification is added to a technical collocation, for example, *infinite dimensional space* or *symmetric unbounded operators*. Some are only used in the long version, whereas others are used by the same informant in multiple versions. We wanted to ensure that we were not analysing the nested collocations which were only found in the longer version as a combination of two separate collocations. Thus, our acid test was that if a combination only existed as a four-word combination in the informants' lecture data, it was counted as one collocation which was four words long. However, if components of the four-word combination were also used as, for example, two-word combinations, these were counted separately. For example, *infinite dimensional spaces* were found only in this form and were therefore counted as one three-word collocation. Conversely, both *unbounded operators* and *symmetric unbounded operators* were found in the data and were therefore counted as two separate collocations. To check which word combinations should be considered the "node" of the collocation, as a rule of thumb we counted the words in the collocation working back from the back (right) and working left (towards the front) (e.g. *value problem* with 614,000 hits and *initial value problem* which got 89,000 hits). This identified the combination "value problem" as the node for this complex nested collocation which was then coded as adjective+collocation and not as collocation+noun. All collocations were checked electronically (using AntConc) to confirm the number of instances of each collocation in the data.

4. Preliminary Results from Maths and LAS

Following the analytic procedures outlined above, we have so far finished the analysis of data from the 12 informants from LAS and from Maths.

As can be seen in Table 2, the two groups of informants produce more or less the same number of collocations (1776 and 1773). A similar picture emerges when we look at the collocational density per 1,000 words spoken (88 and 95, respectively). Differences, however, emerge when we look at the distribution between lexical and grammatical collocations, with 72% lexical and 28% grammatical collocations produced by the LAS informants, and 66% lexical and 34% grammatical collocations produced by the Maths informants.

Looking at Table 3, there were also a few other interesting findings related to some of the structural types used across the two groups. Considerably more noun+noun constructions were used by the LAS informants (15% compared to 2% for the Maths group), for example, *animal welfare, body mass, contrast effect* (LAS). More adjective+noun constructions were used by the Maths informants (42% compared to 31% for the LAS group), for example, *deep theorem, simple fact, algebraic operations* (Maths) and twice as many adverb+adjective constructions were used by the LAS informants as the Maths informants (10% compared to 5%), for example, *highly efficient, very bad, very noisy* (LAS). Regarding nested collocations, a considerably higher percentage were used by Maths informants (9%) compared to LAS (1%). Examples include *compact Hausdorff space, use the matrix norm, classification theorems for* (Maths).

The inclusion of all structural types made it possible to explore differences across the departments studied. For example, if noun+noun compounds had not been included, a specific feature of collocational usage differentiating LAS (193 instances) from Maths informants (40 instances) would have been missed.

Looking at the distinction between the general, academic and domain-specific collocations, other interesting differences between the two academic domains included in our study were found. As can be seen in Table 4, one in five (22%) of the LAS collocations were deemed to be domain specific, whereas over half of those (52.3%) identified in the Maths data were classified as domain-specific. If the domain-specific collocations had not been included, an important distributional difference in collocational use between lecturers from the two academic fields, LAS and Maths, would have been missed. Overall, both groups produced surprisingly few academic collocations. This is probably due to the use of Pearson's list which has been extracted from a written corpus, and may not reflect the usage of collocations in spoken academic communication. As shown by Dang (2016), an academic word list for oral language for single word items is different from Coxhead's list (2000) developed from a written corpus. Unfortunately, as far as we know, an Academic Collocations List for oral data is yet to be developed.

Table 2. Total Number of Lexical and Grammatical Collocations Used

	Large animal science group (N=6)		Mathematics group (N=6)	
	No. of collocations	% ages	No. of collocations	% ages
Lexical	1281	72	1179	66
Gram	495	28	594	34
Totals	1776	100	1773	100

Table 3. Breakdown in Lexical Collocations Used per Department

Structural types	Large animal science (*N*=6)		Mathematics group (*N*=6)	
	No. of collocations	% ages	No. of collocations	% ages
adj+adv	0	0	1	0
adj+n	396	31	493	42
adj+n+n	1	0	0	0
adj+v	1	0	1	0
adv+adj	132	10	59	5
adv+adv	2	0	0	0
adv+v	7	1	7	1
n+adj	2	0	0	0
n+n	194	15	40	2
n+phr	2	0	1	0
n+v	16	1.5	20	2
phr+n	12	1	5	0.5
pn+n	7	1	25	2
v+adj	184	14	175	15
v+adv	32	3	8	1
v+n	280	22	241	21
v+phr	1	0	0	0
v+v	1	0	0	0
Nested colls	11	1	103	9
Total lexical collocations	1281	100	1179	100

phr = phrasal verb.

Table 4. Collocations Used per 1000 Words Spoken Split into Academic, General and Domain-Specific Collocations

Collocational types	Large Animal Science (*N*=6)		Mathematics (*N*=6)	
	Per 1000 words	% ages	Per 1000 words	% ages
Academic	1.5	2	1	0.7
General	67	76	44	47
Domain-specific	19.5	22	50	52.3
Totals	88	100	95	100

5. Discussion and Perspectives

5.1 Research Questions

Two of the research aims of the project involve identifying the challenges related to studying collocational use in oral academic discourse and establishing a "one-size fits all" research methodology for researching collocations across academic disciplines. At present, the research outlined has revealed a number of research challenges that we have tried to meet by applying a manual procedure of identifying potential collocations and a three-tier analytical approach of mutual

exclusion for classifying collocations as general, academic or domain-specific collocations. On the basis of this procedure of analysis, the data from our two study domains have been analysed and described.

As regards research questions 3, 4 and 5, the above results have revealed interesting differences in collocational use across the two domains investigated. The results highlight the importance of applying an extensive approach which enables us to achieve an overall measure of collocational use based on the inclusion of all collocational types found in the data. The grammatical collocations make up a sizeable proportion of the total, and their distribution varies across the academic domains. If, in line with many previous studies, we had only focused on for example the lexical collocations, we could have missed between a quarter and a third of the collocations used by our informants. Only including the general collocations, which have been the focus of most previous research on collocations, would have yielded a picture where the LAS informants are perceived to use more collocations overall than the Maths informants, which as we have seen is far from the case.

Looking more closely at the distinction between lexical and grammatical collocations, some slight differences between the relative proportions of lexical and grammatical collocations between the LAS and Maths informants were found. Including more informants that would have allowed for statistical analysis of the data may have revealed that these tendencies in fact describe statistically significant differences across domains. Interesting differences in the structural sub-types types used have also been found. For example, excluding noun+noun compounds would have omitted a large proportion of collocations (in the case of the LAS group, 11% of all collocations) and would have hidden another important difference between the profiles of the two groups (as only 2% of the Maths group's collocations were categorised as noun+noun compounds). As can be seen from the tables in Section 4, the distributional difference in the number of general and domain-specific collocations is also striking. Extending the study to include our three IT informants (which completes our data set) or even non-STEM disciplines, that is, the social sciences and humanities (SSH) may reveal further differences.

Looking at the general, academic and domain-specific collocations, it was found that the proportion of domain-specific collocations was considerably higher in the Maths group than in the LAS group, indicating that Maths may be a more "technical" discipline. Our results thus support the idea that clear differences across academic domains may be found in relation to the use of technical vocabulary. Chung and Nation (2003) found that an anatomy textbook contained a significantly larger proportion of fully technical (single word) terms (one in three) than an applied linguistics textbook (one in five). Fraser (2006) even found a figure of 35.9%, that is, more than one in three, technical words in a pharmacology text. In addition, many of the domain-specific collocations found in our data from the Maths group were made up of nested collocations. Both these phenomena seem to be because many of the terms used in Maths are made up of complex names describing particular theorems or modified names of theorems, for example, *Hausdorff space/compact Hausdorff space, C star algebra/concrete C star algebra/finite dimensional C star algebra*. In contrast, the domain-specific terms used by the LAS group tend to be more "penetrable" and stable, typically two-word noun+noun combinations such as *animal behaviour, bird predators, quail chickens,* and *pinyon jay*.

5.2 Future Research and Pedagogical Perspectives

Future research will include covering more disciplines, in order to make more nuanced analyses of potential differences in domain-specific academic language across disciplines. We will apply the same methods to analyse the data from the IT group of informants, and this will give us more comprehensive overall results related to possible differences across domains, perhaps revealing an even more complex picture of cross-disciplinary language use. Moreover, the inclusion of an additional discipline will reveal the robustness of our analytical approach. Finally, we will try to find a method of using expert raters to validate our initial categorisation of the domain-specific collocations which has been based on our Google search procedure.

In describing L2 learners' overall collocational use, Westbrook (2015) revealed that only focusing on general collocations was insufficient to show expected correlations between collocational use and fluency; however, such correlations may conceivably have been in evidence if the other types of collocations had been included. When results for the three domains are in place, we would like to run Westbrook's (2015) fluency measures against our data to find out if, with the inclusion of domain-specific collocations, any correlations can be found between collocational use and fluency. Including all the collocations used by the individual informant has made it possible for us to draw up various measures of collocational use that can be correlated with such fluency measures, but also other proficiency measures in forthcoming studies.

In the future, it would also be important to develop tools that may guide non-native students and researchers, especially for the non-maths domains, where fewer resources (such as technical dictionaries) may be found. This could involve developing lists of domain-specific collocations for various disciplines based on analysis of both L2 and L1 data. It could also be useful to develop sub-lists of clusters of collocations, for example, for frequent de-lexical verbs (have, get, give, make, etc.): HAVE + *a family of curves, the product of, a representation, a unique trace, a norm, a time evolution, real space*, etc. (see Menon & Mukundan, 2010). Finally, it could be useful to draw up a list of grey-zone area collocations, for example,. collocations that have both a general and a crypto-technical sense.

6. Conclusions

This paper has shown that, as with single word findings, there are considerable differences in collocational use across different academic domains, at least as far as our data from the two domains Maths and LAS are concerned. However, on top of the dimensions differentiating single word use between domains, there are also several other dimensions characteristic of collocational use that need to be taken into account in order to fully map collocational use in academic language and across academic domains. In particular, as well as the dimension covering differences between general, academic and domain-specific collocations, such analyses should also examine the lexical/grammatical collocational dimension, different collocational structures within the lexical/grammatical dimension, and nested collocations.

We have suggested possible avenues for solving the many problems inherent in analysing collocations from different domains, demonstrating what we feel is a

robust method to identify collocations and the various dimensions which contribute to an informant's overall collocational use. This method will hopefully be reinforced with the inclusion of IT and other academic domains in our future research.

References

Ackermann, K., & Chen, Y. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes* (Impact Factor: 0.8) *12*(4), 235–247. doi:10.1016/j.jeap.2013.08.002

AILA World Congress 2014, 10–15 August 2014.

Ananiadou, S., & McNaught, J. (1995). Terms are not alone: Term choice and choice terms. *Journal of Aslib Proceedings, 47*(2), 47–60. doi:10.1108/eb051381

Barfield, A., & Gyllstad, H. (2009). Introduction: Researching L2 collocation knowledge and development. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 1–20). Basingstoke: Palgrave Macmillan.

Chung, T. & Nation, I.S.P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language, 15*(2), 103–116. doi:10.1016/j.system.2003.11.008

Chung, T. & Nation, I.S.P. (2004). Identifying technical vocabulary. *System, 32*(2), 251–263.

Complexity and Idiomaticity, Stockholm University, June 2012.

Coxhead, A. (2000). A New Academic Wordlist. *TESOL Quarterly, 34*(2), 213–238. doi:10.2307/3587951

Dang, T. (2016). Investigating vocabulary in academic spoken English: Corpora, teachers, and learners, PhD Thesis, Victoria University of Wellington.

EIE, The Copenhagen conference (19–21 April 2013): The English Language in Europe in Teaching in European Higher Education.

Frantzi, K.T. & Ananiadou, S. (1996). Extracting nested collocations. In Proceedings of the 16th International Conference on Computational Linguistics, 1996. COLING '96, pp. 41–46. Association for Computational Linguistics.

Fraser, S. (2006). The nature and role of specialized vocabulary: What do ESP teachers and learners need to know? *Hiroshima University Scholarly Journals, 2005*, 63–75.

Granger, S. & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds,), *Phraseology: An interdisciplinary perspective* (pp. 27–49). Amsterdam: John Benjamin.

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H.A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction,* 28, 24–34. doi:10.1016/j.learninstruc.2013.04.001

Handl, S. (2009). Towards collocational webs for presenting collocations in learners' dictionaries. In A. Barfield & H. Gyllstad (pp. 69–85). Multiple interpretations. Basingstoke: Palgrave Macmillan, 2009.

Henriksen, B. (2013). Research on L2 learners' collocational use and development – a progress report. In Bardel, Camilla; Laufer, Batia & Lindqvist, Christina (Eds.), *L2 vocabulary acquisition, knowledge and use. New perspectives on assessment and corpus analysis*. Eurosla Monographs Series, 2. EUROSLA.

Herbel-Eisenmann, B.A. (2002). Using student contributions and multiple representations to develop mathematical language. *Mathematics Teaching in the Middle School,* 8(2), 100.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, 19(1), 24–44. doi:10.1093/applin/19.1.24

Hwang, K. & Nation, P. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System, 23*(1), 35–41. doi:10.1016/0346-251X(94)00050-G

Laufer, B. & Waldman, T. (2011). Verb-noun collocations in second-language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647–672. doi:10.1111/j.1467-9922.2010.00621.x

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299–320. doi:10.1093/applin/ams010

Méndez Cendón, B. (2004). *Medical language collocations: The case of the verb perform*. A New Spectrum of Translation Studies. Valladolid: Universidad de Valladolid, pp. 195–208.

Menon, S. & Mukandan, J. (2010). Analysing collocational patterns of semi-technical words in science textbooks. *The Social Science & Humanities, 18*(2), 241–258.

Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt & M. McCarthy (Eds), *Vocabulary description, acquisition and pedagogy* (pp. 40–63). Cambridge: Cambridge University Press.

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* New York: Newbury House.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524759

Nattinger, J.R. & DeCarrico, J.S. (1992). *Lexical phrases and language teaching.* Oxford: Oxford University Press.

Nesselhauf, N. (2005). Collocations in a learner corpus. *Studies in Corpus Linguistics, 14*. doi:10.1075/scl.14. John Benjamins

PhD Applied Linguistics (Lexical studies) annual conference, 10-13th of March, 2014, Cardiff, Wales.

Pulverness, A. (2007). Review of English collocations in use. *ELT Journal, 61*(2). doi:10.1093/elt/ccm014

SDU SELC Conference, Odense 2013.

Westbrook, P. (2015). Talk about mouth speculums: Collocational use and spoken fluency in non-native English-speaking university lecturers. *Studies in Parallel Language Use, C8*, Copenhagen Studies in Bilingualism. University of Copenhagen.

Xue, G., & Nation, I.S.P. (1984). A university word list. Language Learning and Communication*, 3*(2), 215–229.

Vocabulary Learning and Instruction
Volume 6, Issue 1, November 2017
http://vli-journal.org

The Use of Psycholinguistic Formulaic Language in the Speech of Higher Level Japanese Speakers of English

Stephen F. Cutler
Cardiff University
doi: http://dx.doi.org/10.7820/vli.v06.1.Cutler

Abstract

A recent study by Cordier (2013) suggests that psycholinguistic formulaic sequences (multiword units that present a processing advantage to the individual speaker) may be more prevalent in L2 speakers than previously thought. The current study adopts the same identification process to explore the use of psycholinguistic formulaic sequences in the speech of Japanese Speakers of English (JSE).

Eight adult JSE at intermediate or advanced levels of English each performed two speaking tasks: a structured interview and a narration task. Formulaic sequences were identified on the basis of hierarchical conditions applied in strict order. The first condition was fluency and the second condition checked for holisticity (using given diagnostic criteria). For each sample, two measures of formulaicity were calculated: FS% (the percentage of syllables that were part of a formulaic sequence) and ANR (the average number of formulaic syllables per run).

The mean formulaicity of the samples (FS%=34.6%, ANR=1.64) suggests that psycholinguistic formulaic sequences, as defined and identified here, may be a significant feature in the speech of intermediate/advanced JSE. The study also confirms the sensitivity of the results to task, with significantly more formulaic sequences used in the interview task than in the narration. Overall, the identification process was found to be a useful and systematic way of identifying formulaic sequences, but some further refinements of the diagnostic criteria and measures used are also suggested.

1. Background

1.1 Psycholinguistic Formulaic Sequences and L2 Speakers

Formulaic sequences may be defined as prefabricated multiword strings that behave as a single lexical unit. Such sequences are thought to be a significant and ubiquitous feature of native speaker language (Ellis, 2012; Nattinger & DeCarrico, 1992; Pawley & Syder, 1983) and to play a key role in facilitating fluency and automaticity in speech (e.g., Ejzenberg, 2000; Towell, Hawkins, & Bazergui, 1996; Wray, 2002). The central claim is that the sequence as a single holistic lexical item is processed more quickly during speech production than if processing the same sequence online on a word-by-word basis. For this reason, the acquisition and use of an appropriate stock of formulaic sequences would seem to be a useful goal for

Please cite this article as: Cutler, S. F. (2017). The Use of Psycholinguistic Formulaic Language in the Speech of Higher Level Japanese Speakers of English. *Vocabulary Learning and Instruction, 6* (1), 48–60. doi: 10.7820/vli.v06.1.Cutler

L2 learners wishing to develop their fluency. However, the existing research on use of formulaic sequences in L2 speakers suggests that such usage is limited and inconsistent at best. For example, Paquot and Granger (2012) have reviewed the use of formulaic sequences in L2 and found that, even in more advanced speakers, it is marked by the underuse of referential collocations, multiword verb phrases, and idiomatic usage, and by the overuse of some meta-discursive expressions. This research entailed the identification of formulaic sequences in corpora of L2 language using standardized lists or frequency-based methods. There have also been some explorations of individual usage. For example, in a case study on the speech of an intermediate-level Japanese speaker of English, Wood (2009) found that about 12% of her speech consisted of formulaic sequences.

Overall, research into L2 usage of formulaic sequences is limited and has been hampered by a lack of consistency in the definition of formulaic sequences and how they are identified. In particular, the focus has tended to be on sequences that are considered to be formulaic "in the language" (such as idioms and high-frequency multiword units). Wray (2008) makes the distinction between such externally defined sequences and those which may be "psycholinguistic" units in the lexicon of the individual speaker. A number of researchers (e.g., Dahlmann, 2009; Erman, 2007) have shown that these are not necessarily the same, particularly for L2 speakers. For example, an L2 speaker may know of a particular idiom (which is formulaic in the language) but not be able to use it smoothly. At the same time, a nonidiomatic expression (such as "It's a real problem") may become psycholinguistically formulaic for that speaker through frequent repetition. Such formulaic sequences acquired by "fusion" (Peters, 1983; Schmitt & Carter, 2004) may be particularly important for L2 speakers as they represent language that is especially useful and relevant to the individual. Research based on frequency measures or standardized lists within corpora will however tend to miss these (unless it is a corpus of individual usage).

Cordier (2013) defines "psycholinguistic formulaic sequences" as multiword units that present a processing advantage to the individual speaker—either because they are stored holistically or because they are processed automatically as a unit. This definition extends a widely used definition by Wray (2002, p. 9) and facilitates the identification of formulaic sequences (as holistically processed items) on the basis of the spoken output. While there is no direct way to measure the storage or processing of lexical items in the speaker's mind, the study of the temporal features of speech output can give indication of the nature of language processing (Temple, 2000). One key temporal feature that has been widely used for this purpose is fluency (e.g., Lin, 2010; Towell et al., 1993; Wray, 2002). In particular, the absence of disfluency markers (such as pauses, hesitation, and repetition) was used as a criterion for formulaicity in studies by Erman (2007) and Dahlmann (2009), and also in algorithmic approaches such as those of Brooke, Tsang, Hirst, and Shein (2014). A further approach to identification is the use of diagnostic criteria (e.g., Wood, 2009; Wray, 2008). In this approach, a number of different criteria for formulaicity (such as "there is something grammatically unusual about this word string") are listed and the satisfaction of one or more of these is taken to indicate that a sequence is likely to be formulaic.

Combining the fluency and diagnostic criteria approaches to identification, a recent study by Cordier (2013) has suggested that psycholinguistic formulaic sequences

may be more prevalent in the speech of intermediate and advanced L2 speakers than previously thought. She analyzed the speech of five advanced British L2 French speakers who had each undertaken a set of different speaking tasks consisting of interviews, discussions, and story narrations. Formulaic sequences were identified by applying the fluency and diagnostic criteria on a hierarchical basis (following Hickey, 1993) meaning that conditions were applied in a strict order. For a sequence to be declared formulaic, it first had to satisfy the necessary condition of fluency and then also had to satisfy at least one diagnostic criterion to indicate that it showed signs of being a holistic unit. The main results using this methodology were that an average 27.7% of her participants' speech was formulaic. In addition, she found significant differences in observed formulaicity across different tasks, with the story-telling task producing fewer formulaic sequences than the interview or discussion tasks.

1.2 Current Study

The current study used the same identification process and hierarchical criteria to investigate the use of psycholinguistic formulaic sequences in the speech of Japanese speakers of English (JSE). Its aim was to estimate the amount and type of formulaic speech used by a particular group of intermediate and advanced JSE and to check how this compared with the previous research. In undertaking a study of this nature, it is important to recognize that any count of formulaic sequences in an individual's speech depends on how they have been defined and the measurement process used. The practical and theoretical issues associated with investigating formulaicity in the speech of L2 speakers will therefore also be discussed.

The three main research questions are as follows:

RQ1 To what extent do psycholinguistic formulaic sequences feature in the speech of these intermediate/advanced JSE, and how does this compare with results from previous research?

RQ2 How does the nature of the task affect the number of formulaic sequences used?

RQ3 What types of formulaic sequences are used by the speakers and how do these contribute to overall formulaicity in these speakers?

An effective method of identifying formulaic sequences in the speech of L2 learners is important if we are to monitor their acquisition and usage. This study, being the first to apply these hierarchical criteria to JSE, provides an opportunity for testing the methodology as well as giving insight into the prevalence of formulaic sequences by this group of speakers.

2. Method

2.1 Participants

The participants were eight JSE, all of whom were volunteer office workers recruited from companies in Japan that were known to the researcher. The participants were chosen on the basis of availability and to provide a mix of background (in terms

of experience and opportunities to use English) and proficiency levels. There were seven females and one male and their ages ranged from 32 to 55. Four of the participants were from the same company, and three of the participants had similar jobs (associated with book-keeping and accountancy). To provide a point of reference, two native speakers from the United Kingdom also undertook the identical process. Both were working adults with occupations unrelated to teaching English or linguistics.

2.2 Procedure

The participants each undertook two speaking tasks: a structured interview about their work lasting 4–5 minutes, and a story narration based on a picture sequence (around 3–4 minutes). For the story, they had a choice of three picture sequences and were given 2 minutes to prepare. Participants were told the nature and timings of the tasks but not the focus of the research. Informed consent was obtained and they were assured about the anonymity of their contributions. All tasks were recorded and transcribed, with pauses and other relevant disfluency marked. Formulaic sequences were then identified according to a set of hierarchical conditions, following the methodology of Cordier (2013). These conditions were applied in three stages to provide a progressive filtering of the transcribed speech.

2.2.1 Necessary: phonological coherence

The first necessary condition was that of phonological coherence, here operationalized as fluent pronunciation. This has been used as a validation measure in the identification process before (e.g., Dahlmann, 2009; Erman, 2007; Raupach, 1984) but not as an initial necessary condition in a hierarchy of criteria. Signs of disfluency were defined to be:

(1) unfilled pauses > 0.25 seconds
(2) filled pauses (e.g., *er*, *umm*, *ah*)
(3) syllable lengthening > 0.4 seconds
(4) repetition or repair/retracing

The 0.25 seconds cut-off for unfilled pauses follows a standard used frequently in fluency research (e.g., Kormos & Denes, 2004; Lennon, 2000). Filled pauses were taken as nonwords not containing semantic information. For example, lexical fillers (e.g., *you know, yeah*) were not taken as filled pauses since they have a function and may themselves be examples of formulaic sequences. The identification of syllable lengthening follows Dahlmann (2009) and was taken to indicate the end of a run. These disfluency indicators were used to segment the speech stream into fluent runs. For example:

> SACHI: it's // funny because he // I'm working in the office // and it // it's just he and me // so // when he went on business overseas // I just...

2.2.2 Necessary: At least one typical condition showing a holistic dimension

Fluent runs can potentially be quite long stretches of speech and are not necessarily formulaic in themselves. Indeed there may be several formulaic

sequences along with individual words within a fluent run. Therefore, a further way of identifying the formulaic sequences from within the runs was required. The second necessary criterion defined by Cordier (2013) was that there should be at least one typical condition showing a holistic dimension. The diagnostic criteria used here were adapted from those used by Wray (2008) and Wood (2009) and are as follows:

(1) Grammatical or semantic irregularity: The meaning of the sequence is not given by its parts, or the grammar of the sequence is not that typically used to express the meaning. Examples from the current study included: "they lived happily ever after," "bits and pieces," "can't handle."
(2) Regular sequences with semantic or functional unity: These are typically grammatical units, common collocations, proper names, or other sequences with a clear holistic mapping of form to meaning or function. Examples from the current study included: "in charge of," "of course," "on the other hand," "typical day," "Toshima Ward."
(3) Sequences likely to have been learned or used as a whole by the speaker: This was based in the diagnostic criteria from Wray (2008, p. 116): "based on direct evidence or my intuition, there is a greater than chance level probability that this speaker will have used this precise formulation before in communication with other people". Examples from the current study included expressions from the speaker's work experience (e.g., "total administration time," "TOEIC essay contest") or ones that they were likely to have learned before ("on the other hand").

It should be noted that the above criteria are by no means mutually exclusive, and a sequence may satisfy more than one criteria (e.g., "on the other hand" above). This is not surprising since there are a number of potential causal or theoretical links between the criteria. For example, most irregular sequences known to a speaker are likely to have been learned or experienced as a whole. However, evidence of holisticity only requires the satisfaction of one criterion. So, for the purposes of this procedure, no special significance is attached to sequences satisfying multiple criteria.

2.2.3 Graded: frequency

A further graded condition used was that of intra-speaker frequency (i.e., does the speaker use the same term repeatedly). In a small speech sample, it is not possible or desirable to use the repetition of an expression as a necessary criterion. However, when expressions are repeated by a speaker, it adds to the likelihood that they are formulaic (assuming the other conditions are also satisfied). For example, one participant said "I'm very surprised" on three different occasions (even when narrating the past).

2.3 Measures

Two main measures of "formulaicity" were used. For comparative purposes these were identical to the ones used by Cordier (2013):

(1) FS% (Percentage of formulaic syllables): the number of syllables in the speech sample that were part of a formulaic sequence divided by the total number of syllables in that sample.
(2) ANR (Average number of formulaic syllables per fluent run): the number of syllables that were part of a formulaic sequence divided by the number of fluent runs in the speech sample.

The FS% measure gives an overall sense of how much of speech is part of a formulaic sequence, while ANR gives a sense of how they divide up the speech stream. In addition to the formulaicity measures, some standard temporal measures of speech fluency were calculated for each sample in order to explore how formulaicity may vary with fluency. These were the Speech Rate (SR) in syllables per minute, and Mean Length of Runs (MLR) which measures the average length in syllables of a fluent run between disfluency markers (e.g., Kormos & Denes, 2004; Lennon, 2000).

3. Results

Overall, 4,798 words (6,340 syllables) were spoken by the eight participants over the two tasks and 663 formulaic sequence tokens were identified (449 types). These contained 1,685 words (2,285 syllables). The average number of words (syllables) per formulaic sequence was 2.54 (3.56). There were 214 repetitions (22.2%) overall, with 67 tokens (40 types) being repetitions across two or more participants. The most repeated sequences were "for example" (12 tokens across 5 participants), "you know" (11 tokens / 2 participants), and "I think" (9 tokens / 6 participants).

3.1 Types of Formulaic Sequence Used

To explore the different types of formulaic sequence that participants used, sequences were categorized according to a broad typology developed by Cordier (2013). This was chosen to provide a direct comparison with the previous study. In this typology, "Referential sequences" are defined as those predominantly used to refer to entities such as objects, places, times, or ideas. "Meta-discursive expressions" are sequences used to structure, comment on, or engage with the discourse or message, and "Sentence builders" (from Nattinger & DeCarrico, 1992) are the fixed parts of patterns used to build sentences and phrases. The relative distribution of sequences across each category types is given in Table 1 along with examples from the study for each category and subcategory.

3.2 Formulaic Sequence Usage by Task

In order to explore differences in the usage of formulaic sequences across the two tasks, mean values of each formulaicity measure across the participants were calculated. Table 2 shows these values (along with the range for each) for each task and in total. Comparing the two tasks, the results show that more formulaic sequences were used in the first task (the interview

Table 1. Distribution of Formulaic Sequences by Category

Category	Subcategory and examples	No. (%)
Referential	Verb phrase – *have to deal with* Noun phrase – *book stores* Time/place complements – *last year* Adverbials – *on behalf of* Whole clause – *they lived happily ever after*	486 (74)
Meta-discursive	Hedges – *some kind of* Fillers – *you know* Asides – *what do I do?* Discourse structure – *for example*	100 (15)
Sentence builders	*I think __* *I'm not good at __* *It's nothing to do with __*	77 (12)

Table 2. Mean Values (and Ranges) for Both Formulaicity Measures

	Task 1 (Work interview)	Task 2 (Picture story)	Total
FS%	38.2% (33.2–48.1)	31.0% (26.0–38.4)	34.6% (29.6–40.3)
ANR	1.89 (1.03–2.79)	1.39 (0.53–2.64)	1.64 (0.82–2.63)

about their job) than in the second picture narration task. Using a paired *t*-test (two-tailed), these differences were found to be significant (t=3.14, p=0.016 and t=3.62, p=0.009) for both of the formulaicity measures (FS% and ANR). For the combined samples, the mean FS% was 34.6% and mean ANR was 1.64. These mean figures are substantially higher than those found by Cordier (2013) whose five advanced French learners had mean FS% = 27.7% (range 22.1–31.0) and mean ANR = 1.50 (range 0.83 – 1.90) over the five tasks they undertook.

3.3 Formulaic Sequence Usage by Participant

A summary of the quantitative measures of formulaic sequence usage and fluency for each participant are given in Table 3, arranged in order of fluency (SR). Note: participants have been given pseudonyms.

As can be seen from the data, formulaicity as measured by ANR (the average number of formulaic syllables per fluent run) increases consistently in line with fluency (SR). In particular, the two participants (Yayoi and Yoko) who had considerable experience (2 years or more) of living overseas also had the highest fluency and ANR scores. On the contrary, the FS% measure does not show a clear pattern with respect to fluency. For example, the participant Wataru has the highest FS% score (40.3%) but was one of the less fluent speakers (SR=97.0) on the tasks. The two native speakers who did the same tasks and followed the same procedure had considerably higher usage of formulaic sequences than all of the participants (FS%=46.4 and 48.1%, ANR= 3.74 and 4.81) and they were also more fluent (SR=182.0 and 195.7). This provides a good validation of the procedure.

Table 3. Summary Fluency and Formulaicity of Participants Over Both Tasks

Participant	Sex/Age	TOEIC*	FS%	ANR	SR (syll/min)	MLR (syll)
Junko	F-40+	650	30.9	0.81	70.9	2.54
Eri	F-50+	735	29.6	0.84	83.6	2.82
Wataru	M-40+	–	40.3	1.44	97.0	3.50
Sachi	F-40+	865	36.0	1.78	115.7	4.96
Kanae	F-30+	940	35.6	1.58	123.4	4.44
Mami	F-30+	–	33.8	1.81	127.3	5.34
Yayoi	F-40+	975	31.9	2.21	148.3	6.80
Yoko	F-40+	960	38.5	2.63	175.9	6.85

MLR, Mean Length of Runs; TOEIC, Test of English for International Communication.

4. Discussion

4.1 Use of Psycholinguistic Formulaic Sequences

Insofar as they can be reliably measured on the basis of the criteria used here, the FS% figures suggest that psycholinguistic formulaic sequences may be a significant part (e.g., 30–40%) of the speech of the JSE participating in the study. The sequences used were mainly referential (verb phrases, noun phrases, time/place complements), accounting for 74% of all sequences. Within this category, there were few repetitions between or within the individual participant samples, and (as in the previous research) there were few examples of grammatically or functionally irregular sequences found. Meta-discursive and sentence building sequences accounted for a smaller proportion of the sequences overall (15 and 12% respectively), but the majority of repeated expressions (e.g., "I think," "for example," "you know") were from these two categories. The distribution of sequences by category and the mostly standard nature of these matches what Cordier (2013) found with her advanced French learners. Overall, the picture of psycholinguistic formulaic sequence usage that emerges is that of the speakers using a breadth of canonical (transparent and grammatical) referential sequences, each being used only once or twice with almost no overlap across participants. These are then supplemented by a number of useful meta-discursive or sentence building expressions which tend to be repeated more, particularly by the participants with higher degrees of formulaicity in their speech.

Regarding the two different tasks that the participants undertook, there was a significant difference in the formulaicity of samples in them. This was true for both formulaicity measures FS% and ANR, with the interview task producing more sequences than the story-telling in each case. This supports the finding of previous research. For example Cordier (2013) found significant differences between all the task types used, with the more interactive interview and discussion tasks yielding more formulaic sequence usage than the narrative task. In the current study, this may be thought to reflect the familiarity of the topics as much as the tasks themselves. In the work interview task, participants tended to use expressions specifically related to their work and experience (e.g., "procedures for foreigners," "put the cheque in," "test administration," "month end" etc.)

which they have likely used frequently before. In the story narration however, the content was not so familiar to the participants and there were likely to be fewer referential sequences easily available to them. On the other hand, when narrating in general, there are potential opportunities to use common sequences for organizing discourse (e.g., expressions for sequencing time and events such as "last year" or "after that") that the participants could have usefully employed. However, apart from a few examples (e.g., "the next day," "ten years later"), these were not used extensively by most speakers in this study.

While the distribution of sequences by task, category, and regularity is similar to that found in Cordier's study, the formulaicity figures in the current study (for intermediate/advanced JSE) are higher than those found for her advanced British speakers of French. Despite the obvious difference that the texts were in different languages in the two studies, the size and direction of the difference in the FS% scores is perhaps surprising. A possible contributory factor may have been a small difference in the pause cut-off length used (0.25 seconds compared to the 0.2 seconds used by Cordier). However, a follow-up analysis on a sample of the sequences identified as formulaic in the study found that none would have been rejected even if a 0.2 seconds cut-off was applied. A further possibility is that, due to the essentially probabilistic and contextual nature of diagnostic criteria, there may be systematic differences in applying the criteria in the second stage of the identification process. This point is explored further in the next section.

4.2 Identification Challenges

Although a consistent and well-defined process was used, the actual application of the method highlighted particular challenges inherent in identification arising from the nature of formulaic sequences themselves and the necessarily interpretative nature of diagnostic criteria. Three particular challenges were illustrated in the study.

4.2.1 Degree of "fixedness" within the sequence

Formulaic sequences may be either fixed or constructed as frames with slots for variables (Wray, 2002). In addition, they may be subject to expansion (e.g., adding an intensifier within the sequence) or nesting (placing one sequence in the variable slot of another). Deciding which of these options is applicable in individual cases can be challenging, and use of the conditions and criteria may not always be able to resolve this. Such decisions are important however since they may affect which words within the string are taken to be part of the formulaic sequence, thereby affecting the quantitative measures of formulaicity. The following example from the study illustrates this challenge.

(a) YAYOI: it's partially the subcontractor's job to train proctors

The expression in (a) was delivered fluently by the participant and therefore satisfies the first condition for being a formulaic sequence. For the second (holistic) condition, either criterion 2b ("has functional or semantic unity") or 2c ("has been used in the same form to convey the same meaning") may be applicable. However, they may potentially be applied at different levels of abstraction.

For example, it is possible that the whole expression is formulaic as this is a work-related topic which has clearly been discussed before. On the contrary, it could be that the frame "it's someone's job to do something" is formulaic for this speaker, with the (familiar) variables slotted in appropriately and the qualifier "partially" added as an (optional) expansion.

4.4.2 Dynamic nature of formulaicity

The study also provided examples illustrating the potentially dynamic and context-based nature of formulaicity in the individual speaker (e.g., Ellis, 2012). For example, Junko in her interview initially appeared to construct the phrase "PR unit" (as the English translation of her department name) and then subsequently used it in a formulaic way.

(b) JUNKO: My job is a PR- (1) unit? (..) I am in PR unit. […] I think (...) PR unit is very conservative

The phrase "PR unit" does have a semantic unity (Criterion 2b) and is repeated (Criterion 3). So, the two fluent cases of the phrase in the example are taken to be formulaic sequences in the current procedure. However, the evidence of earlier disfluency of the sequence also seems important. For example, here it seems to indicate that the sequence is newly formed and, as such, may only be temporarily available in a holistic form. Other potential indicators of such "temporary formulaicity" may include mixtures of fluent and nonfluent usage of a sequence, or the repetition of a formulaic expression taken from the interviewer's question. Indeed, examples of both indicators were observed in the current study. The extent to which this kind of contextual information should be applied will depend on the needs of the research and how one views the status of newly formed or temporary formulaic sequences. However, since such decision will affect the count of formulaic sequences observed, it is important for any identification process to be explicit about how it deals with these cases.

4.2.3 Use of "multiword" as a defining feature

In most approaches, formulaic sequences are taken to be explicitly "multiword" sequences operating in a unitary way. In such cases, the word is by implication a defining feature of the formulaic sequence. However, as Wray (2014) argues, the concept of the word is not always clear, due to the existence of contractions, polywords, compound nouns, hyphenated words, and so on. While explicit clarifications can be made at the definitional stage (e.g., in this study, contractions, polywords, and hyphenated words are all taken to be multiple words), there were examples from the study that reveal the slightly arbitrary nature of using the word as a defining feature for identification. For example, "test takers" and "a lot of" were included as formulaic but not the single words "examinees" or "many," even though on definitional criteria they are essentially equivalent. This highlights a challenge in applying a multiword criteria as a definitional feature of formulaic sequences, and is another potential source of difference in the identification process,

4.3 The Formulaicity Measures

Two variables, ANR and FS%, were used in this study to provide a measure of the "formulaicity" of the participants' speech samples, and the results show a different pattern across the participants for each. ANR (the average number of formulaic syllables per fluent run) seems to have a close association with fluency, with ANR values increasing in line with increasing SR. However, for FS% (the proportion of syllables that were part of a formulaic sequence), there is not such a clear pattern. For example, one participant, Wataru, had a high value for FS% even though he spoke quite hesitantly (as shown by his fluency measures). At the same time, one of the most fluent speakers, Yayoi, had a comparatively low FS% over her two samples. One way to interpret this is to acknowledge that different measures indicate different aspects of performance and processing. For example, researchers (e.g., Towell et al., 1996) have argued that fluency as measured by MLR (i.e., a greater ability to formulate runs) may be due to greater proceduralization in processing (e.g., in the formulator in Levelt's model of speech production, 1993) and that such proceduralization is facilitated by the use of formulaic sequences. However, how such usage is measured is also important and the results here suggest a possible differentiation of the roles of FS% and ANR.

A case such as that of Wataru, who uses a comparatively high number of formulaic sequences but with high number of disfluent gaps between them (as indicated by low ANR), demonstrates that the proportion of syllables that are formulaic (i.e., FS%) is not necessarily a useful measure of formulaicity to associate with aspects of speech processing such as proceduralization. The FS% figure represents the proportion of speech that is part of a formulaic sequence, but it does not indicate the number and length of sequences or how they fit together into fluent runs (for which ANR may be more appropriate). What this highlights is that although the FS% variable may have intuitive appeal as an apparent measure of how formulaic a speech sample is, it may not be the most appropriate measure for this purpose.

5. Conclusion

This study shows that psycholinguistic formulaic sequences, defined as fluent, semantically or functionally coherent multiword sequences, may be a significant feature in the speech of intermediate/advanced JSE. The results of this first study to use these particular identification criteria on such speakers broadly agree with the main findings of the previous research using the same method, and give some further insight into the prevalence of psycholinguistic formulaic sequences in L2 speakers as well as the practical challenges of identifying these. It also adds further weight to the finding that formulaic sequence usage is sensitive to the kind of task that is used to elicit speech. Overall, the study demonstrates how a systematic hierarchical procedure can be used to identify formulaic sequences in a useful way. In particular, the use of disfluency as an initial criterion provided a clearly quantifiable starting point for identification that can be consistently applied. Examples of sequences used by participants also highlighted some theoretical and definitional aspects of formulaic sequences that will be helpful in making the diagnostic criteria more robust and in interpreting the meaning of formulaicity measures such as FS% and ANR.

At the same time, there are some clear limitations to the study. In particular, this was a small study with specific group of learners which therefore has limited generalizability on its own. In addition, undertaking the procedure highlighted a number of the inherent challenges in identifying formulaic sequences in spoken output. These centered on the dynamic and graded nature of formulaicity and the interpretative nature of diagnostic criteria. Two recommendations for making the process more robust therefore can be proposed. First, ensure that there are explicit, theoretically justified "rules" to cover ambiguous cases (such as when there is a mix of fluent and disfluent examples of the same sequences or when there are multiple interpretations). These help in further standardizing the process. It is also particularly important to use contextual information from the task and from the individual's speech sample as a whole, and to specify how to apply it. However, even with such refinements, it should be recognized that the diagnostic criteria are based on likelihoods and are not always strictly quantifiable on the evidence available. So, a second important recommendation is to utilize multiple judges to make the diagnostic assessments and to have explicit rules and procedures to deal with disputed cases when pooling the results.

Overall, the study supports the suggestion that the use of psycholinguistic formulaic sequences (as measured by ANR for example) is associated with fluency. An observation from the study was that a principal area of difference in formulaic sequence usage between participants with higher and lower ANR (and fluency) was in the use of meta-discursive and sentence starter expressions and their repetition. In particular, higher fluency participants tended to use (and repeat) a greater number of general discursive expressions (sequencers, hedges, and fillers) and longer types of sentence building patterns. This suggests that a useful focus, even for the higher level JSE in this study, may be to support them in becoming fluent in the production of a prioritized set of such formulaic sequences, in order to enhance their output delivery.

References

Brooke, J., Tsang, V., Hirst, G., & Shein, F. (2014). Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 753–761), Dublin, Ireland, August 23–29 2014.

Cordier, C. (2013). *The presence, nature and role of formulaic sequences in English advanced learners of French: A longitudinal study* (Unpublished doctoral thesis). Newcastle University, Newcastle, UK.

Dahlmann, I. (2009). *Towards a multi-word unit inventory of spoken discourse* (Unpublished doctoral thesis), University of Nottingham, Nottingham, UK.

Ellis, N.C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics, 32*, 17–44. doi:10.1017/S0267190512000025

Erman, B. (2007). Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics, 12*(1), 25–55. doi:10.1075/ijcl.12.1.04erm

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (ed.), *Perspectives on fluency* (pp. 25–42). Ann Arbor, MI: University of Michigan Press.

Hickey, T. (1993). Identifying formulas in first language acquisition. *Journal of Child Language, 20*, 27–41. doi:10.1017/S0305000900009107

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145–164. doi:10.1016/j.system.2004.01.001

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 43–60). Ann Arbor, MI: The University of Michigan Press.

Levelt, W.J. (1993). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Lin, P.M.S. (2010). The phonology of formulaic sequences: A review. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 174–193). London, UK: Continuum.

Nattinger, J.R. & DeCarrico, J.S. (1992). *Lexical phrases and language teaching.* Oxford, UK: Oxford University Press.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*, 130–149. doi:10.1017/S0267190512000098

Pawley, A., & Syder, F.H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J.C. Richards & R.W. Schmidt (Eds.), *Language and communication* (pp. 191–225). London, UK: Longman.

Peters, A.M. (1983). *The units of language acquisition*. Cambridge, UK: Cambridge University Press.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp.1–22). Amsterdam, Netherlands: John Benjamins.

Temple, L. (2000). Second language learner speech production. *Studia Linguistica, 54*(2), 288–297. doi:10.1111/1467-9582.00068

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84–119. doi:10.1093/applin/17.1.84

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. Canadian *Journal of Applied Linguistics/Revue Canadienne de Linguistique Appliquée, 12*(1), 39–57.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press

Wray, A. (2008). *Formulaic language: pushing the boundaries.* Oxford, UK: Oxford University Press.

Wray, A. (2014). Why are we so sure we know what a word is? In J. Taylor (Ed.), *The Oxford handbook of the word*. Oxford, UK: Oxford University Press.

Vocabulary Learning and Instruction
Volume 6, Issue 1, November 2017
http://vli-journal.org

Profiling Lexical Diversity in College-level Writing

Melanie C. González
Salem State University
doi: http://dx.doi.org/10.7820/vli.v06.1.González

Abstract

The present paper reports on a study that examined the contribution of lexical frequency to lexical diversity in narrative texts composed by 119 multilingual and monolingual English-speaking students enrolled in first-year college writing courses. The Measure of Textual Lexical Diversity (MTLD) quantified lexical diversity and the BNC-COCA 25 strand in Lextutor's *VocabProfile Compleat* sorted the words according to frequency band. Overall, results from statistical analyses indicated that sample's lexical diversity was not significantly impacted by the use of high-frequency (1,000–3,000 bands) or low-frequency (9,000+ bands) terms. Instead, texts showed greater differences in the mid-frequency (3,000–9,000) bands ($p<0.05$). There were also significant differences between MTLD writers' written productive use of mid-frequency words. Consequently, findings suggest that mid-frequency vocabulary may play a greater role in academic writing quality than the attention it is typically given in the L2 writing classroom.

Keywords: second-language writing; second-language vocabulary; lexical diversity; lexical frequency; academic writing.

1. Introduction

Over the years, vocabulary-related studies in multilingual (ML) writing scholarship consistently cite the positive influence of lexical diversity on college-level writing quality (Crossley & McNamara, 2009; Engber, 1995; Friginal, Li, & Weigle, 2014; Johnson, Acevedo, & Mercado, 2013). Lexical diversity, which refers to the varied use of unique vocabulary words within a text, manifests itself as "the sophisticated use of vocabulary" and "a variety and range of vocabulary" within holistic rubrics that assess the college writing readiness of ML students (Hawkey & Barker, 2004). Such criteria point to the need for ML writers to have a large and diverse lexicon from which to purposefully and strategically select words while writing. However, little is known about what vocabulary items actually contribute to the lexical diversity of a text. As a result, writing instructors aiming to help ML writers to improve their students' productive lexicon often must rely on their intuition to determine which words to target for instructional activities. The present study is a first step toward profiling the lexical diversity of college-level writing.

Please cite this article as: González, M. C. (2017). Profiling Lexical Diversity in College-level Writing. *Vocabulary Learning and Instruction, 6* (1), 61–74. doi: 10.7820/vli.v06.1.González

2. Literature Review

2.1 Lexical Diversity

In much of the research over the years on the intersection of vocabulary and writing, lexical diversity has emerged as a critical component affecting raters' judgment of academic writing quality. In an early study of first language (L1) writing, Grobe (1981) compared the writing scores of 5th, 8th, and 11th grade students with a variety of syntactic, mechanical, and lexical measures. Results indicated that lexical diversity, or the total number of different words, was the strongest predictor of writing score. Linnarud (1986) found a similar result in a study comparing the compositions of advanced ML and monolingual English-speaking (MES) academic writers. Despite their high level of English proficiency, the ML essays lacked lexical diversity and sophistication. In addition, there was a difference in the frequency of individual words used by the MES and ML learners. The MES writers tended to use more low-frequency adjectives and adverbs, thereby expressing more sophistication and adding diversity to their compositions.

Later studies removed the comparison of MES and ML writer texts and focused on one group or the other to determine aspects of writing quality. For example, Engber (1995) examined the specific relationship between lexical proficiency and reader perception of the overall quality of essays written by ML students and found that error-free lexical diversity had the highest significant correlation to writing score. Using a within-subjects design, Schoonen, van Gelderan, de Glopper, Hulstijn, Simis, Snellings, and Stevenson (2003) compared writing performance and fluency between participants' academic compositions written in their L1 (Dutch) and L2 (English). Their results determined that the writers' L1 compositions demonstrated a higher level of lexical diversity than their L2 essays. Considering the participants were able to achieve higher levels of lexical diversity within their L1 compositions, this result may signal that their smaller L2 lexicons inhibited their ability to vary words within their L2 writing samples.

More recently, in an analysis of a corpus of L1 undergraduate essays, McNamara, Crossley, and McCarthy (2009) found that more proficient writers demonstrated a greater amount of lexical diversity in their essays. The essays earning the highest scores rarely repeated words and contained words that occur less frequently in language. This result suggests that high-proficiency writers have a larger lexicon from which to draw rich, diverse vocabulary items to express ideas. Similarly, Crossley, Salsbury, and McNamara (2012) examined 100 essays by ML writers at three levels of proficiency (low, mid, and high) and found that lexical diversity could rather accurately predict students' English proficiency levels. Finally, Friginal et al. (2014) profiled the qualities of 353 highly rated undergraduate and graduate essays and found that those papers that garnered perfect and near perfect scores correlated to greater lexical diversity.

Thus, empirical research has demonstrated lexical diversity closely relates to writing proficiency. Evaluators are more likely to award points to writers who are able to vary their word choice during the composition process. The studies discussed in this section further indicate that there are clear differences in the lexical diversity profiles of ML and MES compositions. One possible explanation

for MES writers' ability to vary lexis in text stems from their relatively large vocabulary size as compared to that of ML learners. Therefore, a closer examination of what vocabulary ML writers ought to possess within their lexicons in order to increase lexical diversity is warranted.

2.2 Lexical Frequency

When investigating the vocabulary ML writers need for college-level writing tasks, there are generally two approaches researchers and instructors can take (Nation & Waring, 1997; Nation & Webb, 2011). The first approach would be to consider how many words exist in the English language. The second could consider how many words their MES peers would likely know. However, the first approach is not a valid measure for vocabulary size because it is extremely unlikely that MES college writers know every word in existence. Therefore, the second approach, to use the words MES writers know and possess within their lexicon is a more feasible measure to use for comparison.

Vocabulary researchers commonly use indices of lexical frequency to operationalize and quantify learners' lexicons for empirical study (Laufer & Nation, 1995; Nation, 2006). Lexical frequency is a word's particular ranking in terms of how frequently it occurs within corpora of natural language production. The rank is represented in bands of 1,000 words groups and ordered by frequency of use. The first 3,000 words indicate high-frequency words, the next 3,000–9,000 words are considered mid-frequency words, and the remaining words in the 9,000 and above comprise the low-frequency category (Cobb, 2007; Schmitt & Schmitt, 2014). This classification of words into 1,000 word family bands by their frequency of use operates under the assumptions that (a) high-frequency words are encountered more often in natural language production and therefore are easier to acquire and are highly useful (Meara & Bell, 2001; Schmitt & Schmitt, 2014; Stæhr, 2008) and (b) the frequency bands can offer a sampling of the words language users possess within their lexicons and are often used in tests of vocabulary size (Nation, 2006; Nation & Begler, 2007; Schmitt, Schmitt, & Clapham, 2001).

Studies of lexical frequency in writing have resulted in a number of useful conclusions for ML writers and their instructors. First, lexical frequency has significant correlations with writing proficiency. ML writers at lower levels of language proficiency tend to use more high-frequency words than writers at higher levels of proficiency or MES students who more often insert words from a wider range of frequency bands into their texts (Crossley & McNamara, 2009; Johnson et al. 2013; Laufer & Nation, 1995; Stæhr, 2008). Therefore, studies generally recommend that instructors of ML writers explicitly include vocabulary instruction in the L2 writing classroom in order to stretch ML students' productive lexicons to include words from the lower frequency bands in their writings.

Second, lexical frequency offers a way to evaluate the utility of words based on the percentage of text coverage the various bands achieve. Corpus research has concluded that high-frequency words can cover up to 80–90% of written English texts, with mid-frequency vocabulary contributing an additional 5–8% of text coverage, and low-frequency words making up the remaining percentages (Nation, 2006). Such thresholds hold important implications for ML students' abilities to

comprehend and produce vocabulary within the written domain. Research has shown that 1,000 to 9,000 word family bands are critical to obtaining 98% text coverage, the minimum figure needed to comprehend academic texts (Laufer & Ravenhorst-Kalovski, 2010). While a similar threshold is yet to be identified for writing proficiency, it can be inferred that the vocabulary up to the 9,000 level also facilitates academic writing (Stæhr, 2008).

Finally, lexical frequency is also a main indicator of lexical sophistication, a term often utilized in academic writing rubrics such as the TOEFL, IELTS, and English as a second language (ESL) Composition Profile that assess writing proficiency and academic writing readiness. Read (2000) defines lexical sophistication as "the use of technical terms and jargon as well as the kind of uncommon words that allow writers to express their meanings in a precise and sophisticated manner" (p. 200). This definition is based on the operationalization that a sophisticated lexical item is one that does not occur frequently in use. In more pedagogical terms, lexical sophistication is often coded as "big words," "academic words," "technical terms," or "tier 3" words (Beck, McKeown, & Kucan, 2002; Calderón, 2007; Coxhead, 2000; Schmitt, 2010). These rubrics are based on studies that have found that college-level writing tends to draw more heavily from the mid- and low-frequency bands than more informal texts (Daller, Van Hout, & Treffers-Daller, 2003; Hawkey & Barker, 2004). Studies comparing the lexical frequency profiles between ML and MES college texts have indicated that MES writers produce more lexically sophisticated texts (Crossley & McNamara, 2009).

2.3 The Intersection between Lexical Frequency and Lexical Diversity

In terms of the intersection between lexical frequency and lexical diversity, there is an assumption within the field that a text that produces more low-frequency words would earn a higher score of lexical diversity than a similar text that uses more high-frequency terms (Daller et al., 2003). However, there is scarce, but preliminary, evidence that lexical frequency does not always correlate well with lexical diversity. Laufer (1994) unearthed no significant relationship between learners' increase in the use of lower frequency terms and their lexical diversity scores. Johnson et al.'s (2013) study, however, indicated that the use of lower frequency words, albeit from the 4,000–5,000 frequency bands, did facilitate writing score. In a study of advanced ML and MES college writers, González (2013) found there was only a moderate correlation between lexical frequency and diversity, suggesting that lexically diverse texts do not always require the use of low-frequency, sophisticated words.

These findings suggest that academic writers do not necessarily need to draw from the low-frequency bands in order to achieve the lexical diversity necessary for proficient L2 writing. However, further study of which frequency bands facilitate lexical diversity is needed in order to validate these results.

3. Research Questions

Therefore, the examination of college writers' written lexical frequency profiles and how they intersect with lexical diversity has the potential to lend insight

into possible gaps in their productive vocabulary knowledge. Furthermore, looking at what frequency bands are represented in multilingual (ML) texts as compared to their monolingual English-speaking (MES) peers can help to answer the question on what words to spend instructional time on.

Given the lack of studies profiling lexical diversity in academic compositions, this study took a first step in the hope of filling this gap via the following two research questions:

(1) How do the lexical frequency profiles of advanced multilingual writers' college-level compositions compare to those of their MES peers?
(2) What frequency level(s) is a significant contributor to lexical diversity in college-level compositions?

3. Methods

The research goal sought to examine the lexical frequency profile of the lexical diversity present in authentic college-level writing. In order to meet this goal, a corpus of college student-authored texts from four first-year college writing courses was gathered and analyzed using two computerized textual profilers. Regression and difference in mean statistical analyses targeted the answers to the two research questions. The following sections provide further detail on the methods employed.

3.1 Corpus Collection

In order to ensure that the corpus included texts from both multilingual (ML) and MES writers, first-year writing classes that contained both populations of writers at four universities where the researcher had professional contacts were targeted. Student writers were given a demographic survey in order to identify them as a ML or MES writer. The resulting corpus contained 377 texts.

Given that the vocabulary of a text is highly dependent on the genre, prompt, and length of a text, the corpus was culled according to these criteria in order to control for students' word choice and subsequently the lexical frequency and diversity profiles. First, the corpus was restricted to the final drafts of narrative essays with the topic of "Myself as a Writer," which was the first writing assignment required in the targeted writing courses to provide insights about the students' abilities as a writer to the instructor. From this narrowed corpus, texts that were a minimum of 500 words in length were selected. This process yielded 119 texts (from the original 377) for analysis that controlled for the genre, prompt, and text length. Of these 119 texts, 65 were written by MES students and 54 were written by ML students. Overall, 17 first languages were represented within the corpus (see Table 1).

3.2 Textual Profiling

In order for the two computerized textual profilers to recognize and tag the words accurately, minor mechanical errors (such as spelling, extra spaces, and consecutive

Table 1. First Languages Represented in the Corpus

First language	*n*
Arabic	1
Amharic	1
Chinese	21
English	64
French	1
Hindi	2
Hmong	4
Indonesian	1
Japanese	2
Kannada	1
Korean	2
Portuguese	1
Russian	1
Spanish	10
Tagalog	1
Vietnamese	6

repeated words) within each text in the corpus were corrected. These fixes were necessary in order to ensure that mechanical errors would not be mistaken as off-list words and consequently skew the profiles of lexical frequency and lexical diversity. Since the texts were final drafts (and thus inferred to have undergone some proofreading on the part of the authors), the number of mechanical errors needing correction were minimal.

The texts were then entered into the BNC-COCA 25 profiler within the *VocabProfile Compleat* tool available in Lextutor (Cobb, n.d.), which counted the number of words into 1,000 word family bands up to the 25,000 word family and off-list bands. The resulting figures were further classified as high-, mid-, or low-frequency vocabulary using Schmitt and Schmitt's (2014) broad band categorizations with high-frequency terms falling within the first 1,000 to 3,000 word family bands, mid-frequency terms in the 3,000 to 9,000 word family bands, and low-frequency terms lying in the 9,000 word family band and above range and off-list word families. These totals created the lexical frequency variable.

The Measure of Textual Lexical Diversity (MTLD) quantified the lexical diversity of the texts (McCarthy & Jarvis, 2010, available in the Coh-Metrix, Graesser, McNamara, Louwerse, & Cai, 2004). A critical limitation of indices measuring lexical diversity is that scores are greatly impacted by text length. In other words, as the total number of words in a text rises, so does the likelihood of word repetition, thereby reducing lexical diversity. In validation studies, the MTLD has shown to account for this fatal flaw by sampling strings of texts forward and backward multiple times in order to provide a more accurate measure of lexical diversity (McCarthy & Jarvis, 2010). MTLD scores typically range between 70 and 120, where the higher scores equal greater lexical diversity. Although the present study's corpus was restricted to texts with similar word counts, there is nonetheless some variation in text length and as such, the MTLD was chosen to measure the lexical diversity.

The specific quantitative analyses run on the two variables and are discussed in conjunction with the results of these analyses in the subsequent section.

4. Results

Descriptive results of the corpus (n=119) revealed that, on average, text length ranged between 500 and 600 tokens in length with a mean token count of 557. ML writers' (n=54) texts averaged 556 tokens in length while MES writers' (n=65) texts had a mean token count of 559. In terms of lexical frequency, ML writers averaged a total of 546 high-frequency words, 7 mid-frequency words, and 2 words from the low-frequency bands. The MES writers' texts utilized an average of 541 high-frequency words, 14 mid-frequency words, and 3 low-frequency words (see Table 2).

4.1 Research Question 1

The first research question targeted the mean differences between ML and MES writers' profiles of lexical frequency and lexical diversity. A one-way analysis of variance (ANOVA) revealed that ML writers significantly used more high-frequency words than their MES peers ($F_{2,117}$=54.13, p<0.00, η^2=0.57). In contrast, MES writers used significantly more mid-frequency words within their texts than ML writers ($F_{2,117}$=15.12, p<0.00 η^2=0.27). There was no significant difference between ML and MES texts' use of low-frequency words in their compositions. For lexical diversity, MES writers' texts exhibited significantly greater lexical diversity than the texts composed by ML writers ($F_{2,117}$=5.06, p<0.05, η^2=0.11; see Table 3).

4.2 Research Question 2

The second research question sought to determine which broad lexical frequency bands contributed to lexical diversity. Multiple regression analysis revealed that lexical frequency explained about 27% of the variation in lexical

Table 2. Descriptive Data by Language Designation

	Designation	M	SD
High frequency	ML	546.43	8.26
(K1–K2)	MES	541.20	8.29
	Total	544.06	12.46
Mid-frequency	ML	7.09	6.34
(K3–K8)	MES	14.45	6.01
	Total	10.51	6.81
Low frequency	ML	2.39	2.1423
(K9+)	MES	3.20	3.6648
	Total	2.77	3.0460
MTLD	ML	69.54	17.35
	MES	79.95	12.90
	Total	74.38	15.84

MES, monolingual English-speaking; ML, multilingual.

diversity score ($F_{2,117}$= 4.75, p<0.05). The regression model correctly classified 105 of the 119 essays' lexical diversity based on lexical frequency. The regression analysis also revealed that mid-frequency vocabulary was the only significant predictor of lexical diversity (beta=0.42, p<0.05; see Table 4). In other words, as lexical diversity score increased, there was an uptick in likelihood of a mid-frequency word being deployed within the composition.

5. Discussion

The findings that ML college writers' texts (when composing in the same genre and on the same topic) employed more high-frequency words, fewer mid-frequency words, and exhibited less lexical diversity than texts composed by their MES peers are in line with previous studies with similar conclusions (see Crossley & McNamara, 2009, 2012). However, the result that there was no difference between both groups' use of low-frequency words is noteworthy. It suggests that there may be a gap in ML writers' productive knowledge of mid-frequency terms that in turn affects the lexical diversity of their texts, which, as results indicate, is greater when mid-frequency words are present. The result that MES writers used, on average, twice the number of mid-frequency terms in their texts provides some evidence to support the unique contribution of these words to the lexical diversity of a text.

Table 3. One-Way Analysis of Variance of Lexical Frequency Levels and Diversity by Language Designation

	Source	SS	df	MS	F	Sig.	η^2
High frequency	Between groups	3710.49	2	3710.49	54.13	0.00	0.57
	Within groups	2810.25	117	68.54			
	Total	6520.74	119				
Mid-frequency	Between groups	579.97	2	579.97	15.12	0.00	0.27
	Within groups	1572.78	117	38.36			
	Total	2152.74	119				
Low frequency	Between groups	7.00	2	7.00	0.75	0.39	0.02
	Within groups	382.68	117	9.33			
	Total	389.67	119				
MTLD	Between groups	1158.43	2	1158.43	5.06	0.03	0.11
	Within groups	9382.19	117	228.83			
	Total	10540.62	119				

MTLD, Measure of Textual Lexical Diversity.

Table 4. Multiple Regression Predicting Lexical Diversity

Model		B	SE	Beta	t	Sig.	95% CI	
							LB	UB
1	(Constant)	65.7	122.00		0.54	0.60	-181.07	312.48
	High frequency	-0.01	0.25	-0.01	-0.038	0.97	-0.51	0.49
	Mid-frequency	0.93	0.415	0.42	2.24	0.03	0.09	1.77
	Low frequency	1.24	0.75	0.24	1.65	0.11	-0.28	2.75

There are a few potential explanations as to why mid-frequency vocabulary facilitates lexical diversity. First and foremost, mid-frequency words that lie between the 3,000 and 9,000 word family bands contribute 5–8% to the 95–98% written textual coverage, a key threshold in the ability to comprehend academic texts (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006). In addition, word lists of academic vocabulary such as Coxhead's (2000) Academic Word List and Gardner and Davies' (2014) Academic Vocabulary List contain a good deal of words that lie within the mid-frequency vocabulary range and are utilized across academic subject areas. Moreover, many mid-frequency terms (e.g., *endow, bestow*) are synonyms, hypernyms, and hyponyms for high-frequency (e.g., *give*) and low-frequency words (e.g., *bequeath*) that can add sophistication, specificity, and diversity to a text (Ferris, 1994). Consider the following two paragraphs extracted from this study's corpus in which the mid-frequency vocabulary has been bolded and low-frequency words italicized for emphasis:

Excerpt 1

> Writing is an important tool to have in life and can be **applicable** to various occupations. Writing is a way of formulating your thoughts and observations on paper with no requirement on the subject matter. In the past I have been exposed to multiple writing experiences that required that I be clear and **concise** in my *diction.* I have had some encouraging and frustrating experiences, but overall I am determined to master the art of composition.

Excerpt 1 is composed by a MES writer and had an overall lexical diversity score of 95.46. In this 71-token snippet, the writer used 42 high-frequency words, two mid-frequency words, and one low-frequency word.

Excerpt 2

> I believe I have the potential of being a good writer. I could not speak or understand *English* when I entered school in the US, but I was determined to become **fluent** in this new language. I was previously nervous about writing, but as I had many writing opportunities, I continued to improve each time. Who I am as a person describes who I am as a writer. I have always been a determined type and that can show in my writing.

Excerpt 2, composed by a ML writer, has an overall lexical diversity index of 87.85. Of the 67 tokens present, 45 are high-frequency words, one is a mid-frequency word, and one is a low-frequency word (the proper noun of English). These two excerpts from the study's corpus exemplify where using a mid-frequency word in comparable sentences can potentially add diversity and sophistication to a text: the MES writer states, "I have been exposed to multiple writing experiences that required that I be clear and **concise**"; whereas their ML peer writes: "but as I had many writing opportunities, I continued to improve each time." It can be argued that the term *concise* more specifically refers to a positive quality of writing, a term that is more specific than the general but positive verb *improve* that the ML writer employs.

It should also be noted that although the number of low-frequency terms present in the two excerpts are the same (*diction* in Excerpt 1 and *English* in Excerpt 2), more qualitative judgments might perceive the use of *diction* over *English* as a more sophisticated term despite both being categorized as low-frequency. This example points to the need for further qualitative analysis to be included in future studies.

6. Applications to Pedagogy

In terms of pedagogy, the general recommendation has been for ESL instructors to explicitly teach the high-frequency vocabulary (the first 2,000–3,000 word families) and to use more implicit methods such as exposure through extensive input to provide ML students with the mid- and low-frequency terms needed to expand their lexicon. However, as Schmitt (2010) states, "it is clearly not realistic for learners to acquire the lexis beyond the 2,000 level without a great deal of help" (p. 70). Yet, when instructors are asked to identify words outside of high-frequency terms, it is often the low-frequency items that make into lesson plans, which often relates more to students' comprehension of the topic at hand. Such approaches ignore the wide range of mid-frequency vocabulary that comprises a good deal of the academic language research has shown is necessary for college-level literacy events and practices. Furthermore, research has shown that in order for vocabulary to not be a problem for ML students, the mastery of many knowledge aspects of the words within the frequency bands up to the 9,000 word family level is critical for productive use (Schmitt & Schmitt, 2014). In other words, students with only receptive or partial knowledge of mid-frequency words are unlikely to deploy them in their compositions.

Consequently, more research has been advocating for the explicit instruction and reinforcement of all types of vocabulary, including mid-frequency terms, in order to equip ML students with the lexical tools they need for English-medium college coursework (see Folse, 2004, 2008; Schmitt, 2010; Schmitt & Schmitt, 2014). This trend is particularly evident in the "tiering" approach to vocabulary instruction gaining popularity in the U.S. primary and secondary education research base that recommends targeting "tier 2" word for instruction (see Beck et al., 2002 and Calderón, 2007). The conclusions of this study validate a similar, but more frequency-based approach to vocabulary instruction within the L2 writing classroom and advocate for explicit attention to and fortification of mid-frequency vocabulary in order to begin filling in the gaps within ML writers' productive lexicons.

In addition to targeting mid-frequency vocabulary for instruction, the skills involved in diversifying lexis during the composition process also deserve explicit focus in the L2 writing classroom. Learners can be made aware of the role lexical diversity plays in their writing quality through the analysis of model texts as well as instructor think-alouds or talk-alouds demonstrating and modeling the decision-making involved when choosing words. Furthermore, instructors can design exercises and activities that practice rephrasing and identifying synonyms, hypernyms, and hyponyms that add diversity and sophistication to the text. ML students can also utilize various online lexical profilers (such as Lextutor) to obtain

a visual of the spread of their vocabulary during the proofreading, editing, and revision process to target overly repeated words that could be replaced. Finally, it cannot be assumed that students know how to make use of a thesaurus or the synonyms feature in word processing software that could provide assistance in identifying appropriate synonyms. Instructors ought to model, demonstrate, and engage in practice with these tools.

7. Limitations and Future Research

There are some cautions that apply to the present findings. The corpus was relatively small (n=119) and limited to the narrative genre and the topic of "Myself as a Writer." Had the study included a larger and wider range of different genres and topics, findings may shift as vocabulary is highly dependent on these two variables. In addition, the corpus was compiled from first-year writing courses at U.S. universities. As such, conclusions may not be as applicable to writings produced in other contexts, such as in elementary, secondary, or other L2 classrooms that target lower proficiency levels, as these contexts may contain different expectations for vocabulary in writing. Finally, any study using corpus tools for analysis is beholden to the corpus it utilized. The lexical frequency variable was based on British National Corpus and Corpus of Contemporary American English, which includes a wide range of informal and formal topics as well as utilizes the word family as the unit of counting. If frequency analysis had used a different corpus or the lemma as a counting unit, the lexical frequency profiles might shift.

In order to address these limitations as well as continue to tease out the contribution of lexical diversity to writing quality, there are a few avenues for further study. First, the conclusions from the present study could be strengthened through qualitative analyses such as inviting writing instructors to rate the lexical quality of a corpus and interviewing them regarding what aspects impact their ratings, and listing which words they feel contribute most to the lexical quality of a text. Further quantitative analyses are also warranted. For example, including independent measures of the general receptive and productive vocabulary sizes of student writers could isolate the impact of mid-frequency vocabulary on lexical diversity as well as illuminate if in fact ML writers do possess gaps in this range of their lexicons. Finally, experimental methods such as manipulating the lexical frequency and/or diversity in texts and asking writing instructors to judge and rate the resulting quality may provide further insight into the vocabulary needed to perform college-level writing.

8. Conclusion

The present study investigated the influence lexical frequency imparts to lexical diversity in multilingual (ML) and MES first-year college student writing. Quantitative analyses indicate that mid-frequency vocabulary has a greater impact on the lexical diversity of a text and that there are significant differences between ML and MES writers' use of these words. These conclusions support the explicit teaching of mid-frequency vocabulary words and lexical diversity skills in the L2 writing classroom.

References

Beck, I., McKeown, M.G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary development*. New York, NY: Guilford.

Calderón, M. (2007). *Teaching reading to English language learners: Grades 6–12*. Thousand Oaks, CA: Corwin Press.

Cobb, T. (n.d.). *Compleat Lexical Tutor* [Online Software]. Available from http://www.lextutor.ca/

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology, 11*(3), 38–63. Retrieved from http://llt.msu.edu/vol11num3/cobb/

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

Crossley, S. & McNamara, D.S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35*(2), 1–21. doi: 10.1111/j.1467-9817.2010.01449.x

Crossley, S.A., & McNamara, D.S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18*(2), 119–135. doi:10.1016/j.jslw.2009.02.002

Crossley, S.A., Salsbury, T., & McNamara, D.S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29*(2), 243–263. doi: 10.1177/0265532211419331

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics, 24*(2), 197–222. doi:10.1093/applin/24.2.197

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4(2), 139–155. Retrieved from https://www.journals.elsevier.com/journal-of-second-language-writing/

Ferris, D. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*(2), 414–420. doi:10.2307/3587446

Folse, K. (2004). *Vocabulary myths*. Ann Arbor, MI: University of Michigan Press.

Folse, K. (2008). Myth 1: Teaching vocabulary is not the writing teacher's job. In J. Reid (Ed.), *Writing myths: Applying second language research to classroom teaching* (pp.1–17). Ann Arbor, MI: University of Michigan Press.

Friginal, E., Li, M., & Weigle, S.C. (2014). Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing, 23*(1), 1–16. doi:10.1016/j.jslw.2003.09.001

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327. doi:10.1093/applin/amt015

González, M.C. (2013). *The intricate relationship between measures of vocabulary size and lexical diversity as evidenced in non-native and native speaker*

academic compositions (Doctoral dissertation). Retrieved from http://stars.library.ucf.edu/etd/2633. (CFE0004852)

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers, 36*, 193–202. Retrieved from http://129.219.222.66/Publish/pdf/McNamara_Graesser_Coh-Metrix.pdf

Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English, 15*(1), 75–85. Retrieved from https://eric.ed.gov/?id=EJ242214

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing, 9*(2), 122–159. doi:10.1016/j.asw.2004.06.001

Johnson, M., Acevedo, A., & Mercado, L. (2013). What vocabulary should we teach?: Lexical frequency profiles and lexical diversity in second language writing. *Writing & Pedagogy, 5*(1), 82–103. doi: 10.1558/wap.v4i5.1

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal, 25*(2), 21–33. Retrieved from http://journals.sagepub.com/doi/pdf/10.1177/003368829402500202

Laufer, B., & Nation, I.S.P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307–322.

Laufer, B., & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30. Retrieved from http://files.eric.ed.gov/fulltext/EJ887873.pdf

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Malmö, Sweden: Liber Förlag Malmö.

Meara, P. & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect, 16*(3), 5–19. Retrieved from http://www.lognostics.co.uk/vlibrary/meara&bell2001.pdf

McNamara, D.S., Crossley, S.A., & McCarthy, P.M. (2009). Linguistic features of writing quality. *Written Communication, 27*(1), 57–86. doi:10.1177/0741088309351547

McCarthy, P.M., & Jarvis, S. (2010). MTLD, voc-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*, 381–392. doi:10.3758/BRM.42.2.381

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59–82. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/2006-How-large-a-vocab.pdf

Nation, I.S.P., & Beglar, D. (2007). A vocabulary size test. The Language Teacher, 31(7), 9–13. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/2007-Beglar-TLT.pdf

Nation, I.S.P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N. Schimitt & M. McCarthy (Eds.), Vocabulary: Description, acquisition

and pedagogy (pp. 6–19). Cambridge: Cambridge University Press. Retrieved from http://www.lextutor.ca/research/nation_waring_97.html

Nation, I.S.P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston, MA: Heinle-Cengage.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press. Retrieved from http://www.cup.es/servlet/file/store6/item2466355/version1/item_9780521627412_frontmatter.pdf

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* New York, NY: Palgrave MacMillan.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484–503. doi:10.1017/S0261444812000018

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55–88. doi:10.1177/026553220101800103

Schoonen, R., van Gelderen, A., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning, 53*(1), 165–202. doi:10.1177/13670069030070010201

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

Vocabulary Learning and Instruction
Volume 6, Issue 1, November 2017
http://vli-journal.org

i-lex v1 and v2: An Improved Method of Assessing L2 Learner Ability to See Connections between Words?

Ian Munby
Hokkai Gakuen University
doi: http://dx.doi.org/10.7820/vli.v06.1.Munby

Abstract

Knowing a word's associations is considered an aspect of word knowledge. It follows that L2 learner ability to see connections between words may improve with gains in vocabulary knowledge. Word association tests (WATs) may measure not only learner ability to see links between words, but they may also assess the degree of organization of L2 learner lexical knowledge which plays a role in the development of lexical competence. The aim of this study is to develop a new WAT wherein learners are presented with the three most common associates of a cue word. The task is to supply the missing cue word. Following this format, a test was developed using sets of three cue words chosen from the five most common associates to 50 target words (TWs) listed in the Edinburgh Associative Thesaurus, or EAT. Results of an initial study (i-lex v1) showed that, on average, a group of native speakers outperformed an experimental group of Japanese learners of English ranging in level from elementary to upper intermediate. Further, both in the initial study and a follow-up study (i-lex v2), significant and positive correlations were found among nonnative i-lex scores and a translation test. In i-lex v2, significant and positive correlations were also found among nonnative i-lex scores and the New Vocabulary Levels Test. These results indicate that the ability of these groups of participants to see links between highly frequent English words is related to their vocabulary knowledge.

1. Introduction

1.1 Knowledge of Word Associations

One complicating factor with research into L2 word associations is that some researchers have viewed word associations purely as an aspect of individual word knowledge. This line of thinking can be traced back to an influential paper by Richards (1976). He included knowing the "network of associations between that word and other words in language" (p. 81) as the sixth of eight word knowledge categories. It is interesting that his conclusion on this assumption focuses on the way words are stored in the mind "according to associative

Please cite this article as: Munby I. (2017). i-lex v1 and v2: An Improved Method of Assessing L2 Learner Ability to See Connections between Words?. *Vocabulary Learning and Instruction, 6* (1), 75–94. doi: 10.7820/vli.v06.1.Munby

bonds" (p. 87), or how words are learned and remembered, rather than whether or not knowing a word's associations actually constitutes word knowledge. This is a vital consideration since associational knowledge differs from other forms of word knowledge in that it is not usually declarative knowledge, such as orthographic or morphological knowledge. It is for this reason that Meara (1996b) found this sixth assumption to be an exception to the set since it is not "driven exclusively by the concerns of descriptive linguistics, rather than by psycholinguistic concerns" (p. 3). Further, he believes associational knowledge has the potential to explain how word knowledge is acquired, or integrated into existing knowledge, while most descriptive knowledge does not.

While knowledge of word associations in L2 has since become established as an aspect of both productive and receptive word knowledge (Nation, 2001, p. 27), there appears to be a division of views among researchers regarding the degree of importance of associative links in the mind of the learner. On the one hand, associative knowledge is regarded as part of depth of individual word knowledge (e.g., Haastrup & Henriksen, 2000). On the other hand, while Meara and Wolter (2004) do view individual word knowledge as being important, they describe the breadth/depth model as "unfortunate," preferring a "size/organization" model. The notion of organization is linked to a lexicon-focused or lexical network-oriented perspective on L2 lexical competence. This is based on the view that a learner's L2 lexical competence is dependent not only on the number of words known but also on the state or quality of their interdependency in the mind of the learner. Underpinning this notion of lexical organization is the metaphor of the lexical store as a network, described by Aitchison (1994) as a "gigantic multi-dimensional cobweb" of words (p. 84). Wray (2002) also views lexical networks as central to the retention and production of language. These networks may play a role in determining lexical processing efficiency with both second language input (listening and reading) and output (speaking and writing). In other words, the dynamics of a learner's lexical networks may tap into the core of an L2 learner's overall proficiency and ability to make gains in the future.

This said, this view of L2 lexical competence has not gained much traction in the literature in the past two decades. For example, the section on word association in the famous 477-page, otherwise comprehensive book titled *Learning Vocabulary in another Language* (Nation, 2001) barely spans three pages, implying a limited role for word association studies. Similarly, at the end of a 7-page section on word associations in his book *Vocabulary and Language Teaching*, Schmitt (2000) concluded that research in this subfield has produced little that can inform the teaching of vocabulary.

1.2 Word Association Tests

Existing word association tests (WATs) in L2 vocabulary research seek to measure learner ability to make associations between words using a variety of formats. These can be broadly divided into productive WATs and receptive WATs. Given that associative knowledge cannot be classified as declarative knowledge, problems immediately surface in designing free, productive WATs.

A productive WAT typically involves inviting participants to supply a single response (in a discrete or continued WAT) or multiple responses (in a continuous WAT) to a set of stimuli with no restrictions placed on the type of response. In one strand of word association research, researchers such as Schmitt and Meara (1997), Schmitt (1998a), Schmitt (1999), Wolter (2002), Zareva (2005), and Higginbotham, Racine, and Munby (2015) have elicited responses from cue words and measured them against lists of associative norms generated from native speakers for scoring purposes.

While the results of these studies have generally indicated a link between performance on WATs and standard language tests or measures of overall language proficiency, there is some evidence to suggest that results depend on the cue words used to elicit responses and the norms list used to measure them for native-like stereotypy. For example, Kruse, Pankhurst, and Sharwood-Smith (1987) found no correlation between learner performance on a multiple response WAT and cloze test scores. They concluded there was no link between ability to produce word associations and proficiency. However, in a replication of this study using carefully selected cue words and new purpose-built norms lists in a WAT known as WAT20, Fitzpatrick and Munby (2014) found that learner WAT scores correlated significantly and positively with three different proficiency measures: a cloze test, a translation test based on Webb (2008), and the TOEIC (Test Of English for International Communication) test of listening and reading comprehension. These results indicate that with gains in proficiency, learner word associations become more native-like. Nevertheless, the issue raised by Schmitt (1998b) of what constitutes a native-like response is a concern that still challenges the validity of this approach to measuring associative competence. Indeed, Munby (2011) finds several instances of both native and nonnative partic-ipants providing the same nonnorms listed, and therefore nonscoring, responses to some cue words in WAT20.

With receptive WATs (e.g., The Word Associates Test, Read, 1993, 1998; V_Links, Meara & Wolter, 2004), different challenges to validity emerge. For example, with Read's "classic" 1998 test format, the task is to choose associated words to a TW from among two sets of four words each, one containing synonyms and the other containing collocates. As Read (2012) admitted, this format allows testees to guess word associations through elimination processes without knowing why they are associated. A validation study by Schmitt, Ng, and Garras (2011) provided evidence to support this weakness in the form of interviews with participants to examine test-taking strategies. These interviews sometimes revealed discrepancies between their successful answering of items and actual knowledge of the words being tested. A further issue highlighted by Read (2012) is that items are constructed through careful thought and dictio-nary study with the result that they bear no psycholinguistic reality, or are not based on data drawn from lists of associative norms. Finally, since many of the items in this test are low-frequency items, the test does not have the potential to test the ability to see connections between words with lower level learners. However, even if a test has weaknesses, it does not mean that it is not useful, and solutions can usually be found to most weaknesses in test design. Mean-while, there is also a strong case for experimenting with new test formats. This certainly applies to WATs.

1.3 Aims and Research Questions

The aim of this study is therefore to pilot a new WAT inspired by Meara (1994), who mused upon possible uses of a Spanish word association norms list. He suggested presenting learners with the three most common associates of a cue word and asking them to supply the missing word. He added: "This is a task which native speakers find very easy, but one which turns out to be very difficult for non-native speakers" (p. 8). The author therefore decided to design a WAT with this format. The perceived strengths are firstly that the WAT is productive, with limited opportunity for guessing answers as with receptive formats, such as Read's Word Associates Test. Second, there is no need to measure word association responses with norms lists for scoring purposes. Finally, it is appropriate for learners of all levels.

The instructions of this new WAT, i-lex, are as follows: What word is associated with the following sets of three words? Example: *drink*, *red*, *glass* > w _____ [4]. The first letter, "w," is given for you and the word has four letters [4]. The answer is *wine*. In order to assess the validity of this WAT, the following four research questions were formulated to guide two versions of the test hereafter referred to as i-lex v1 and i-lex v2:

RQ1 Does i-lex v1 distinguish between native and nonnative speakers?

RQ2 Is there a significant, positive correlation between learner i-lex v1 and i-lex v2 scores and vocabulary test scores?

RQ3 Do i-lex v1 and i-lex v2 demonstrate internal reliability?

RQ4 Are nonnative speaker i-lex v2 results consistent between test and retest?

With reference to RQ4, note that a retest was not planned for i-lex v1 because it was not certain that the learner performance on the WAT would yield significant and positive correlations with scores on a standard vocabulary test.

2. Methodology

The following describes the methodology adopted in i-lex v1 and i-lex v2. The latter appears with the answer key in Appendix 1. Note that a brief report of the first version of i-lex (i-lex v1) appears in Munby (2013).

2.1 Participants

In i-lex v1, conducted in 2012–2013, the participants comprised a control group of 25 native speakers of English and 99 Japanese EFL (English as a Foreign Language) students who ranged in level from low elementary to upper intermediate. They were from four different universities, and were in their first, second, and third year. In the study on i-lex v2, conducted in 2016, there was no control group, but the experimental group of 164 Japanese participants was similar in range of level to the first study, and was drawn from first- and second-year EFL classes at three of the four universities where data from i-lex v1 were collected. These three

universities are classified nationally as high, middle, and low-ranking. All participants had studied English for at least 6 years.

2.2 Test Materials and Procedures

The three cue words (CWs) in each test item are chosen from the five most common associates on the Edinburgh Associative Thesaurus, or EAT (Kiss et al., 1973). The following criteria for item selection were established:

(1) To minimize the possibility of words used in the test being unknown to lower level testees, all the TWs and CWs must be from the BNC (British National Corpus, 2007) 1K range, with words from this range not commonly known to low-level learners (e.g., *accept*, *account*, and *achieve*) also excluded according to the author's classroom experience.
(2) To lend maximum transparency to the associations being tested, each set of CWs must include the most common associate on the EAT, listed first.
(3) To ensure that dominant primary responses, including polar opposites such as *dark*> *light*, for example, do not excessively facilitate task success, the first CW must not account for more than 50% of the responses to the TW on the EAT.
(4) To avoid priming, the TW must not also appear as a CW in another item.
(5) To avoid conceptual repetition, the CWs in each set must not be part of the same word family, for example, *tell, teller* > *story*.
(6) To avoid the need to accept alternative correct responses, all TWs which are verbs must not have a past tense form with the same number of letters, for example, *lose*/*lost*.
(7) To ensure that the task is purely linguistic, TW–CW relationships must not require cultural knowledge such as *match, cricket* > *test*.
(8) To reduce task difficulty, TWs must not be function words or noncontent words such as *through*.

For scoring purposes, one point was awarded for each correct TW supplied, whether it was spelled correctly or incorrectly, for example *club, up, together* >*joyn* for the TW *join*, but only if the specified number of letters was provided. A total of three trial versions of i-lex v1 were conducted with two groups of learners (n=22 and n=25) in order to: (1) estimate appropriate time limits for a 50-item test, (2) identify and remove problematic items, such as items which no participants could answer, and (3) sort the items in order of difficulty from the easiest to the most difficult in order to limit incidences of lower level test-takers becoming stuck early on in the test and to facilitate consistency among items in a split-half reliability test.

Before beginning the test, participants were told that if they could not answer an item, they should leave it and move on to the next item. After completing i-lex v1 (25 minutes), the nonnative group completed a translation test of controlled productivity adapted from Webb (2008) in 20 minutes (see Appendix 2). The answer key appears in Appendix 3. The task is to write English translations for a series of 160 single words of varying levels of word frequency written in L1(Japanese). In the original version of this translation test there were 180 items with a sample of 60 items from each of the following frequency bands in the BNC:

701st–1,900th, 1,901st–3,400th, and 3,401st–6,600th. In the original version of this translation test there were 180 items, but it was decided to shorten this to 120 items (40 in each band) because there are 36 loan words that are easily translated as transliterations of English loan words. Seventeen of these appear in the third band (3,401st–6,600th). Further, since a number of participants scored close to the maximum score in a previous study (Munby, 2011), it was felt that this test may not have the power to adequately differentiate the higher level students from their lower level peers. For this reason, it was decided to include an additional column of 40 Japanese words, with 10 each from the 6–7,000 K, 7,000–8,000 K, 8,000–9,000 K, and 9,000–10,000 K levels of the BNC. For scoring purposes, what Webb terms as "soft scoring" is applied and misspelled responses are accepted.

Following scoring of i-lex v1, a problem emerged, undetected during trialling, with one of the most challenging items: *boy, face, girl* > *baby* [4]. While 9 of the 99 nonnative participants supplied the target word successfully, 19 provided *body.* Among the 25 native participants, 8 responded successfully, but since 5 also provided *body*, it was decided to replace the item with a new set: *tree, fire, forest* > *wood* [4] in i-lex v2. Following i-lex v1, items were once again sorted for difficulty from easiest to most challenging according to test results and following a further trial with the new item with a group of nonnative participants (n=25).

The procedure for i-lex v2 was also modified slightly. The time allowed was reduced from 25 minutes to 20 minutes following an eleventh-hour conclusion that generally there was very limited pencil activity in the final 5 minutes of the test. Further, a second counter-proficiency measure, the New Vocabulary Levels Test (McLean & Kramer, 2015a), was administered following the translation test. This test was preferred to the Vocabulary Levels Test (Nation, 1990; Schmitt, Schmitt, & Clapham, 2001) for reasons described in McLean and Kramer (2015b), for example, there is no section which tests the first 1,000 word frequency band. Note that 30 minutes were allowed to complete Sections 1, 2, 3, 4, and 5 of this test. These sections test receptive knowledge of the first five frequency bands of the BNC/COCA corpus (Nation, 2012) with 20 multiple-choice items in each. Due to time constraints, it was decided to leave out Section 6 of the test which contains 30 items drawn from the Academic Word List (Coxhead, 2000). All three tests were completed in one 90-minute session during class time. A retest of i-lex v2 was conducted 2 weeks later to examine its reliability, and scores for participants who failed to attend both sessions were eliminated from the data set.

3. Results

In this section, with a view to answering the first two research questions concerning the validity of i-lex v1 and i-lex v2, the descriptive statistics for these two studies are presented in Table 1. In addition, native and nonnative performances are represented in a scatterplot for comparison in Figure 1. Correlational analysis is presented in Table 2, together with a scatterplot representation of i-lex v2 and translation test scores in Figure 2. To address the final two research questions concerning the reliability of i-lex v1 and i-lex v2, the results of the split-half reliability estimates for the learner scores are reported. Finally, test–retest reliability measures in learner i-lex v2 performance are presented.

Table 1. A Comparison of the Means and Standard Deviations of All Test Scores for All Participants

	Mean	SD	High	Low	MPS	*R*
i-lex 1 (ns, *n*=25)	41.76	4.28	49	33	50	0.96
i-lex 1 (nns, *n*=98*)	25.00	7.23	42	6	50	0.99
Translation test	96.26	20.2	145	44	160	0.99
i-lex 2 (nns, *n*=164) T1	21.54	7.98	40	4	50	0.99
i-lex 2 T2	24.86	8.52	42	5	50	0.99
Translation test	89.83	22.76	152	31	160	0.99
New VLT	68.45	20.25	116	20	120	0.99

*Note that scores for one nonnative subject were removed because her i-lex score of 4 was an outlier.

High = Highest score achieved; Low = Lowest score achieved; MPS = Maximum Possible Score; R = Reliability coefficient; T1 = The first test; T2 = Retest after 2 weeks.

Table 2. Pearson Correlations among Scores for i-lex v1 and i-lex v2 and Proficiency Countermeasures

	Translation	New VLT
i-lex v1	0.729**	–
i-lex v2	0.804**	0.749**
New VLT	0.827**	–

Pearson 1-sided *p*-value: **$p<0.01$.

3.1 RQ1 Does i-lex v1 Distinguish Between Native and Nonnative speakers?

With reference to RQ1, the results in Table 1 indicate that, on average, native speakers outperform nonnatives on i-lex v1. A one-tailed unpaired *t*-test confirms that the difference between i-lex v1 scores for the two groups is significant at $t=4.199$ ($p<0.0001$).

Figure 1 features a comparative representation of the distribution of i-lex v1 scores for the two subject groups: natives and nonnatives. One nonnative speaker scored above the native mean, but none of the native speakers scored below the nonnative mean. Note that, as Bachman (1990) points out, native speakers neither perform uniformly well, nor uniformly better than nonnatives in tests designed for L2 language learners.

3.2 RQ2 Is There a Significant, Positive Correlation between Learner i-lex v1 and i-lex v2 Scores and Vocabulary Test Scores?

Pearson correlations among all sets of scores are reported in Table 2. In both i-lex v1 and i-lex v2, these results indicate that the ability of these groups of participants to see links between highly frequent English words is related to their breadth of vocabulary knowledge. See Figure 2 for a scatter plot representation comparing learner performance on i-lex v2 and the translation test.

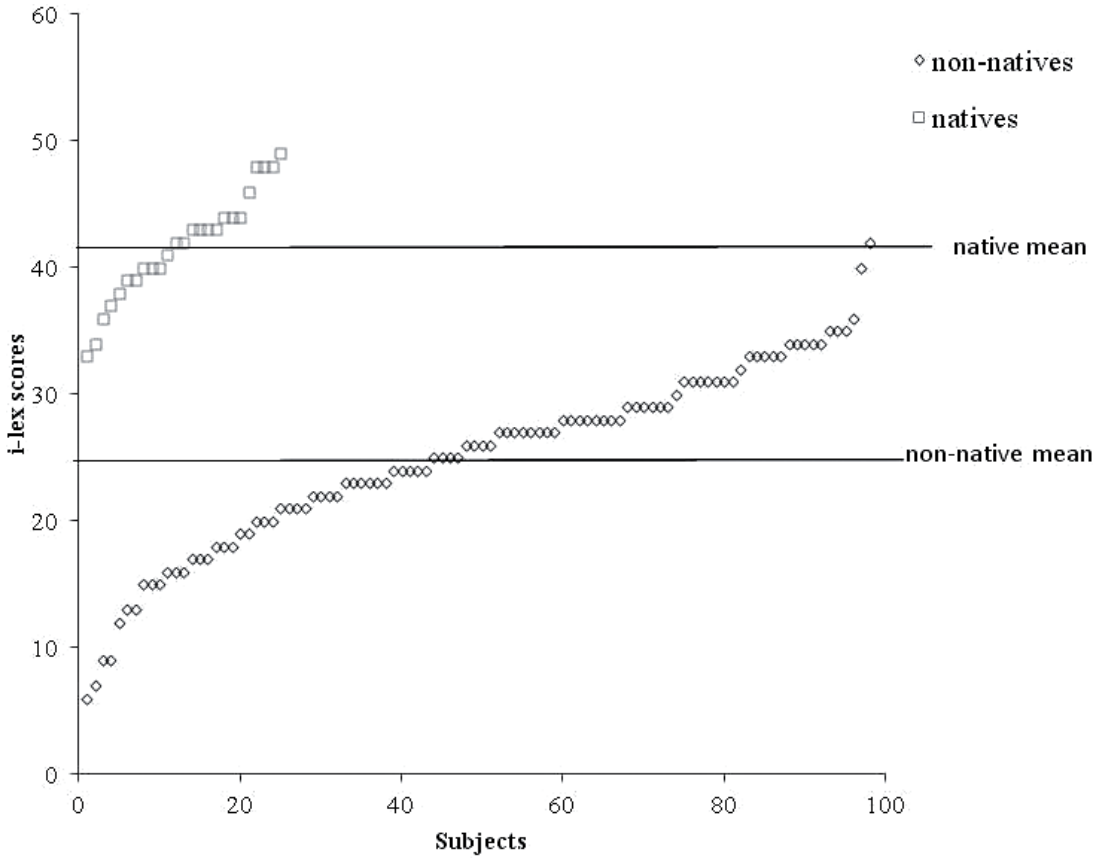

Figure 1. Distribution of nonnative and native speaker scores for i-lex v1.

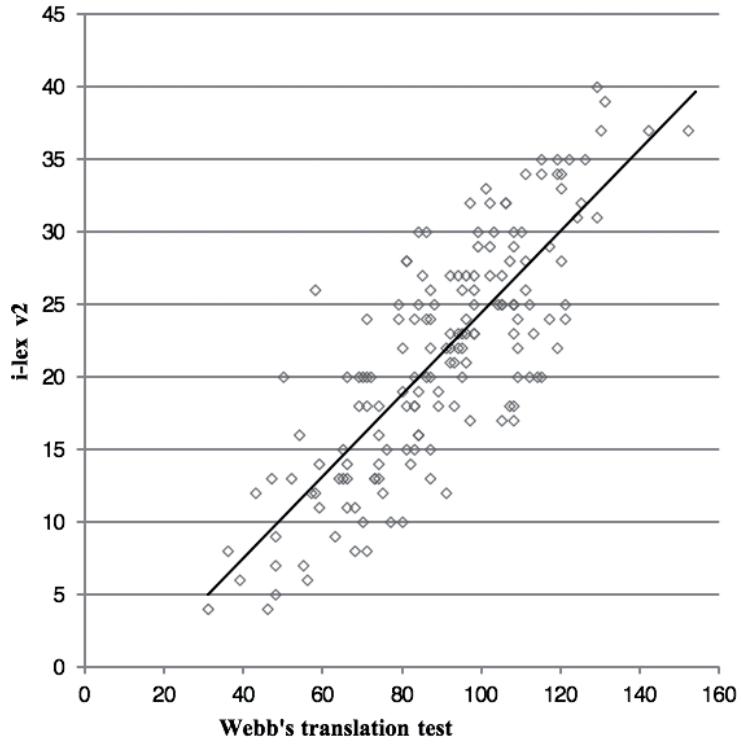

Figure 2. Scatter plot representation comparing learner performance on i-lex v2 and the translation test.

3.3 RQ3 Do i-lex v1 and i-lex v2 Demonstrate Internal Reliability?

In order to rule out the possibility of the participants producing correct responses in much greater quantity for some items than for others, affecting the balance of the set, a split-half reliability test was performed to check results for internal consistency. In this analysis, following Bachman (1990, p. 173), two parallel sets were constructed for both i-lex v1 and i-lex v2. The "odd" set consisted of an analysis of the number of correctly answered items for the whole group for the odd-numbered sets (first, third, fifth, etc.). The even set comprised an analysis of scores for the even-numbered items (second, fourth, sixth, etc.). With reference to Table 3, the correlations for both i-lex v1 and i-lex v2 indicate that the items in each sub-test set of 25 items were assessing ability to see connections in a similar way. Further, results of a *t*-test do not indicate a significant difference between the means of the two sets in both versions of i-lex.

3.4 RQ4 Are Nonnative Speaker i-lex v2 Results Consistent between Test and Retest?

Correlations between i-lex v2 scores at T1 and the T2 retest after 2 weeks are $r = 0.871$ ($p < 0.01$) which suggest a satisfactory level of test–retest reliability of the WAT. As a further reliability check, a paired *t*-test between the pairs of means of the test at time 1 and time 2 was conducted, and a significant difference in i-lex scores was found ($t = 10.0967$, $p < 0.0001$). This indicates that these gains are consistent due to a practice effect that benefitted the majority of nonnative participants. See Figure 3 for a scatter plot representation of test and retest performance.

4. Discussion

This section presents an evaluation of the results in the light of the research questions and correlational analysis. It continues with discussion of the potential limitations of i-lex and suggests avenues for future research

Regarding RQ1, in i-lex v1, the native control group outperforms the nonnative control group since, as Meara (1994) predicted, learners often find this kind of task more difficult than native speakers. With regard to RQ3, both i-lex v1 and i-lex v2 demonstrate internal consistency through a split-half reliability check. With regard to RQ4, with i-lex v2, a test–retest confirms that the WAT has yielded reliable results. Finally, concerning the key research question (RQ2), the results of these two studies indicate a relationship between learner ability to

Table 3. Means and Standard Deviations of Nonnative Participants Scores for Odd-Numbered and Even-Numbered Items in i-lex v1 and i-lex v2, Pearson Correlations between the Two Sets, and One-tailed Paired *t*-test

	ODD Mean (SD)	EVEN Mean (SD)	Correlations	*t*
i-lex v1 (*n*=98)	51.08 (28.3)	47.00 (25.0)	0.879**	1.5118 ns
i-lex v2 (i=164)	70.76 (39.7)	70.84 (46.0)	0.733**	0.9949 ns

Pearson 1-sided *p*-value: Significant at **p<0.01; ns= not significant.

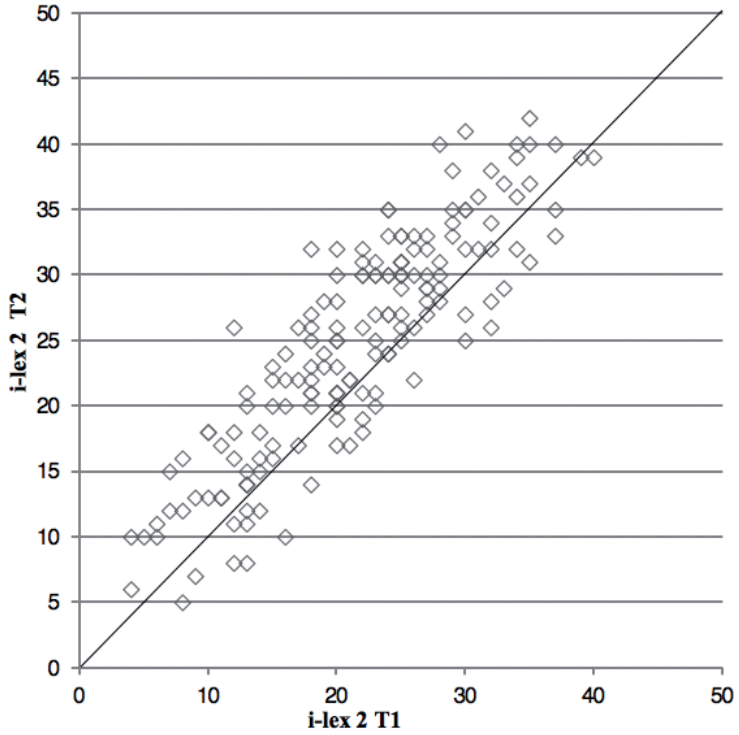

Figure 3. Comparison of i-lex v2 test–retest scores (r= 0.871 [p<0.01]).

see connections between words and breadth of vocabulary knowledge. However, caution is required in claiming the relationship is strong on account of the high correlations between both versions of i-lex and the countermeasures. As Brown (2005) explains, there is a potential problem with correlational analysis since: "If a tester chooses to base a correlational analysis on a sample that is made up of fairly homogenous language proficiency levels … the range of talent may have been restricted, and such a restriction will tend to make any resulting correlation coefficients much lower" (p. 161). Indeed, previous researchers (e.g., Wolter, 2002) have found low, but positive correlations when comparing results of their WATs with proficiency countermeasures in groups of nonnative participants of similar level. Unfortunately, their interpretation of their correlational analyses led to premature conclusions that the weak links were not promising for WA research. Conversely, if a broad range of proficiency levels is used, as in these two studies, the resulting correlation coefficients will tend to be much higher with the suggestion that the tests are essentially measuring the same kind of lexical knowledge.

Nevertheless, the following four potential limitations in i-lex are identifiable. These limitations concern the construct validity of the test defined by Daller, Milton, and Treffers-Daller (2007) as a question of "whether the test measures the skill or construct it is meant to" (p. 16). First, it is by no means certain that all nonnative participants had receptive knowledge of the meanings of all the CWs. It is also possible that they knew the L1 equivalent of the TW, but were unable to supply it due to lack of productive knowledge of the TW in L2. In other words, with some items, with lower level participants, i-lex may also be testing receptive and productive L2 vocabulary knowledge rather than pure ability to see associative links between words. A second potential problem is that it may be possible for

participants to answer items successfully without seeing links between all three CWs as intended. Taking the example of item 28 (*money, account, book> bank*), it is not inconceivable that testees could guess the answer from the first CW *money*. The third issue concerns test-taking strategies. Although test-takers are advised to move on if they are unable to answer an item in order to avoid getting stuck and losing precious time required to answer other items, there is no way of knowing if they took this advice. Alternatively put, low scores may indicate an unwillingness to abandon time-consuming struggles to solve a limited number of problematic items rather than inability to see associative links between words in general. The fourth issue concerns test format familiarity, a threat to validity mentioned by Daller et al. (2007, p. 17). Since the test format of i-lex is nonstandard, and the participants are unlikely to have completed any similar test, performances may be affected by unfamiliarity with the task of supplying a target word from its associates. Admittedly, the high test–retest correlations appear to rule this out to some extent. However, varied performance in i-lex may indicate varied need for practice with task type rather than varied ability in seeing connection between words. Regarding the effects of task familiarization, Akamatsu (2008) finds that a 7-week training program of once a week tests in L2 word recognition with a group of 49 first-year Japanese university students resulted in significant gains in both reaction times and response accuracy. It is therefore possible that participants could make similar gains with training in the i-lex format.

Turning to avenues of future research, there is no question that the issues described above concerning the validity of i-lex require addressing. One way to investigate the first issue of whether the CWs and TWs are known to the nonnative participants would be to give a translation test to a sample of lower level participants. In order to investigate the second issue of guessing TWs, it would be necessary to interview participants post-test to determine how they arrived at their successful answers following an approach adopted by Schmitt et al. (2011) in their validation of the Word Associates Test. Regarding the third potential limitation of test–taking strategies, it may be useful to give a parallel version of i-lex to learners who perform relatively poorly and adopt think aloud procedures to investigate whether or not poor test–taking strategies influence performance. Finally, replicating the approach taken by Akamatsu with i-lex could shed light on the extent to which format familiarity influences results.

It should not be forgotten that because the correlations are positive and significant, this apparent link between ability to see connections between high-frequency words and breadth of vocabulary knowledge requires explanation and further investigation. Of particular interest is whether or not participants who perform relatively well on i-lex are displaying above average lexical processing skills. Longitudinal studies to track learner development over time would be necessary to examine this possibility. In addition, studies involving a battery of tests combining i-lex with other WATs which investigate similar constructs such as WAT20 (Fitzpatrick & Munby, 2014) may shed light on whether similar patterns in learner performance emerge across tests.

In order to support the view that the efficiency of L2 learner lexical processing may be determined by L2 lexical networks, it is necessary to consider language learning from theories and findings in two other fields: foreign language

(FL) aptitude theory and neuroscience. To begin with FL aptitude theory, according to Wen, Biedroń, and Skehan (2017), the work of John Carroll in the 1950s and 1960s has proved enduring to this day. Carroll (1962) viewed specific talent for learning foreign or second languages as dependent on phonetic coding ability, grammatical sensitivity, inductive language learning ability, and associative memory. Clearly, it is the latter, or learner capacity to form associative links in memory, that should interest researchers in the field of L2 vocabulary learning. It may also be useful to draw parallels between associative networks as a metaphor for vocabulary learning in an L2 and evidence from studies in neurological science within the field of FL aptitude. For example, Wen et al. cite empirical studies by Li and Grant (2015) which found that "brain connectivity networks can indeed serve as a reliable predictor for learning both L2 novel words and (artificial) grammar/sequence rules in different learning contexts (natural, virtual etc.)" and that there was clear evidence of "more efficient and more flexible brain connectivity detected among more successful learners as opposed to their less successful counterparts" (p. 14).

Further, recent research conducted by neuroscientists Huth, Heer, Griffiths, Theunissen, and Gallant (2016) found that an "atlas" of the brain can be created by examining activity in the cerebral cortex. The technology employed (voxel-wise modelling of functional MRI) allows for visual imaging of how single words in stories light up spots in different areas of the brain of the listener. These combine to illuminate networks of spots representing how meanings of individual words may be constructed. The shape and form of these networks were found to differ from subject to subject. Judging from what we know about L2 word associations, these networks are likely to be less stable or more tenuous in learners than in native speakers of the language (Meara, 1983). In view of this recent research, lexical networks may not simply be confined to metaphorical representations of L2 language learning. In a similar vein, during the course of a presentation by Anthony, Schmitt, and Nation (2016), Anthony claimed that words are known by the company they keep. Although this comment was related to how studies in corpora can reveal relationships between words, the same claim can be applied to words in associative networks in the minds of language learners.

During the same presentation, in reference to word association research, Schmitt commented that most association responses were idiosyncratic and so we could not learn much about the learners. However, the proportion of idiosyncratic to nonidiosyncratic responses on norms lists of word associations depends on a number of factors such as the CWs selected, whether single or multiple responses are being elicited, and the number of participants. With many CWs, most responses are nonidiosyncratic. Further, the following two points need to be borne in mind. First, considerable gains in our understanding of WA have been made in the last three decades not only in this particular strand of WA research, as evidenced in the work cited in this paper, but also in a wealth of research left uncited here. See Meara (2009) for examples. Second, supported by newly available corpora, such as the BNC, and recent frequency profiling tools, L2 vocabulary research in the last three decades has been primarily motivated by a clearly identifiable need to investigate, for example, the readability of texts for L2 learners and what words need to be learned rather than by a desire to discover how they

are learned. If minor skirmishes regarding the influence on results of guessing in multiple-choice vocabulary size tests do not continue for too much longer, might it not be time to begin a new "post-size" era and to focus more on questions like "why do some learners learn more words than others?"

5. Conclusions

The present study investigated two versions of a new WAT called i-lex. Although conclusions remain tentative pending further validation, it was revealed that the ability of learners to make connections between highly common English words appears to be dependent on the number of words they know. The more words they know, the more connections they are able to identify. At present, it is not known whether this ability to make connections is a cause or a result of knowing the meanings of more words, or if it is a combination of both. Hopefully the next three decades will draw WA research out of the shadows, and it will receive more attention in the field of L2 vocabulary acquisition. Indeed, it is also hoped that new avenues shall be explored that focus more deeply on what it means to know a word and the role of lexical retrieval and memory in L2 lexical processing. At present, to its detriment, the field of L2 vocabulary studies remains remarkably insular. With new research techniques becoming available, this will surely change.

References

Akamatsu, N. (2008). The effects of training on automatization of word recognition in English as a foreign language. *Applied Psycholinguistics, 29*, 175–193. doi:10.1017/S0142716408080089

Aitchison, J. (1994). *Words in the mind.* Oxford: Blackwell.

Anthony, L., Schmitt, N., & Nation, I. (2016). Unanswered Questions in L2 Vocabulary Acquisition and Research Agendas for the Next 10 Years. Speeches presented at Vocab@tokyo, Tokyo, September 14th.

Bachman, L.F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Brown, J.D. (2005). *Testing in language programs*. New York: McGraw-Hill.

Carroll, J.B. (1962). The prediction of success in intensive foreign language training. In R. Glaser (Ed.), *Training research and education* (pp. 87–136). Pittsburgh, PA: University of Pittsburgh Press.

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

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

Fitzpatrick, T., & Munby, I. (2014) Word associations and the L2 lexicon. In J. Milton & T. Fitzpatrick. (Eds.), *Dimensions of vocabulary knowledge*. Basingstoke: Palgrave Macmillan, 92–105.

Haastrup, K., & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics, 10*(2), 221–239. doi:10.1111/j.1473-4192.2000.tb00149.x

Higginbotham, G., Munby, I., & Racine, J.P. (2015). A Japanese word association database of English. *Vocabulary Learning and Instruction, 4*(2), 1–20.

Huth, A.G., Heer, W.A., Griffiths, T.L., Theunissen, F.E., & Gallant, J.L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453–458. doi:10.1038/nature17637

Kiss, G.R., Armstrong, C., & Milroy, R. (1973). *An associative thesaurus of English*. EP Microfilms, Wakefield.

Kruse, H., Pankhurst, J., & Sharwood-Smith, M. (1987). A multiple word association probe. *Studies in Second Language Acquisition, 9*(2), 141–154. doi:10.1017/S0272263100000449

Li, P., & Grant, A. (2015). Second language learning success revealed by brain networks. *Bilingualism: Language and Cognition, 19*(4), doi:10.1017/s1366728915000280

Meara, P.M. (1983). Word associations in a second language. *Nottingham Linguistics Circular, 11*, 28–38. doi:10.1177/026553229301000308

Meara, P.M. (1994). Word associations in Spanish. *Vida Hispánica 10*, 12–22. Retrieved from www.lognostics.co.uk/vlibrary/meara1994.pdf

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

Meara, P.M. (1996b). *The vocabulary knowledge framework*. Retrieved from http://www/swan.ac.uk/cals/calsres.vlibrary/pm96d.htm revised 1999

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

Meara, P.M., & Wolter, B. (2004). V_Links: Beyond vocabulary depth. Angles on the English *Speaking World, 4*, 85–97.

McLean, S., & Kramer, B. (2015a). *The new vocabulary levels test*. Retrieved from http://www.lextutor.ca/

McLean, S., & Kramer, B. (2015b). The creation of a new vocabulary levels test. *Shiken,* 19(2), 1–11. Retrieved from http://www.lextutor.ca/tests/levels/recognition/nvlt/paper.pdf

Munby, I. (2011). Development of a multiple response word association test for learners of English as an L2. Unpublished PhD thesis, University of Wales, Swansea.

Munby, I. (2013). I-lex: An improved method of assessing L2 learner ability to see connections between words. *Vocabulary Education & Research Bulletin, 2*(2), 11–13.

Nation, I.S.P. (1990). *Teaching and learning vocabulary*. Rowley, MA: Newbury House.

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

Nation, I.S.P. (2012). *The BNC/COCA word family lists*. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Information-on-the-BNC_COCA-word-family-lists.pdf

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*(3), 355–371.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah, NJ: Erlbaum.

Read, J. (2012). Piloting vocabulary tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 287–299). London: Routledge.

Richards, J. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10*, 77–89. doi:10.2307/3585941

Schmitt, N. (1998a). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48*, 281–317. doi:10.1111/1467-9922.00042

Schmitt, N. (1998b). Quantifying word association responses: What is native-like? *System, 26*, 389–401. doi:10.1016/S0346-251X(98)00019-0

Schmitt, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing, 16*(2), 189–216. doi:10.1177/026553229901600204

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N., & Meara, P.M. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. Studies in Second Language *Acquisition, 19*, 17–36. doi:10.1017/S0272263197001022

Schmitt, N., Ng, J.W., & Garras, J. (2011). The word associates format: Validation evidence. *Language Testing, 28*(1), 105–126. doi:10.1177/0265532210373605

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55–88. doi:10.1177/026553220101800103

The British National Corpus, version 3.2 (BNC XML Edition). (2007). *Distributed by Oxford University Computing Services on behalf of the BNC Consortium*. Retrieved from http://www.natcorp.ox.ac.uk/

Webb, S. (2008). Receptive and productive vocabulary sizes Of L2 learners. *Studies in Second Language Acquisition, 30*(1), 79–95. doi:10.1017/s0272263108080042

Wen, Z., Biedroń, A., & Skehan, P. (2017). Foreign language aptitude theory: Yesterday, today and tomorrow. *Language Teaching, 50*(1), 1–31. doi:10.1017/s0261444816000276.

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

Wray, A. (2002). *Formulaic language and the Lexicon*. Cambridge: Cambridge University Press.

Zareva, A. (2005). Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System, 33*, 547–562. doi:10.1016/j.system.2005.03.005

Appendix 1. i-lex v2

Word associations puzzle

Name _____ Student number _____

What word is associated with the following sets of 3 words?
Example: drink, red, glass >>>> w i n e [4]
The first letter, "w", is given for you and the word has 4 letters [4]. The answer is "wine"
Write the answers on the spaces below. If you don't know the answer, leave it and go on to the next one.

1. coffee, cup, time	t_____ [3]	26. study, trees, mother	n_____ [6]
2. club, up, together	j_____ [4]	27. round, left, over	t_____ [4]
3. eat, drink, good	f_____ [4]	28. money, account, book	b_____ [4]
4. post, write, box	l_____ [6]	29. hand, shopping, paper	b_____ [3]
5. work, money, employment	j_____ [3]	30. teach, read, book	l_____ [5]
6. science, painting, music	a_____ [3]	31. see, hospital, come	v_____ [5]
7. tree, fire, forest	w_____ [4]	32. leader, me, behind	f_____ [6]
8. country, city, village	t_____ [4]	33. run, talk, over	w_____ [4]
9. truth, down, tell	l_____ [3]	34. game, birthday, playing	c_____ [4]
10. bed, door, space	r_____ [4]	35. god, think, trust	b_____ [7]
11. long, head, cut	h_____ [4]	36. now, later, near	s_____ [4]
12. minute, day, time	h_____ [4]	37. car, fast, light	s_____ [5]
13. man, people, individual	p_____ [6]	38. play, stage, now	a_____ [3]
14. street, car, way	r _____[4]	39. fun, game, drink	p_____ [5]
15. thought, good, mind	i_____ [4]	40. see, people, girl	m_____ [4]
16. time, clock, out	w_____ [5]	41. past, future, here	p_____ [7]
17. big, large, small	s_____ [4]	42. push, out, hard	p_____[4]
18. wrong, left, away	r_____ [5]	43. of, piece, time	p_____ [4]
19. go, home, out	l_____ [5]	44. red, test, cut	b_____ [5]
20. hear, to, radio	l_____ [6]	45. paper, reader, good	n_____ [4]
21. for, stop, bus	w_____ [4]	46. center, end, east	m_____ [6]
22. front, back, top	s_____ [4]	47. boy, teacher, college	s_____ [6]
23. good, fortune, bad	l_____ [4]	48. shopping, words, names	l_____ [4]
24. stick, football, game	m_____ [5]	49. over, out, fast	m_____ [4]
25. fish, ball, drop	c_____ [5]	50. out, kind, of	s_____ [4]

Answer key.
1. Tea 2. Join 3. Food 4. Letter 5. Job 6. Art 7. Wood 8. Town 9. Lie 10. Room 11. Hair 12. Hour 13. Person 14. Road 15. Idea 16. Watch 17. Size 18. Right 19. Leave 20. Listen 21. Wait 22. Side 23. Luck 24. 25. Match 26. Nature 27. Turn 28. Bank 29. Bag 30. Learn 31. Visit 32. Follow 33. Walk 34. Card 35. Believe 36.Soon 37. Speed 38. Act 39. Party 40. Meet 41. Present 42. Pull 43. Part 44. Blood 45. News 46. Middle 47. School 48. List 49. Move 50. Sort

Appendix 2. Translation Test

Write (single word) English translations for the following Japanese words.

Band 1	Band 2	Band 3	Band 4
1. 島 i_____	1. 薬 m_____	1. 霧 f_____	1. 2カ国語 b_____
2. 動物 a_____	2. 宮殿 p_____	2. 糊(のり) g_____	2. 赤道 e_____
3. 住所 a_____	3. 英雄 h_____	3. 靴下 s_____	3. 化石 f_____
4. 地球 e_____	4. 冬 w_____	4. 青あざ b_____	4. エビ s_____
5. 医者 d_____	5. 耳 e_____	5. 誓う s_____	5. いやみ s_____
6. 安い c_____	6. 嫌う h_____	6. 頬 c_____	6. 迷信 s_____
7. 弱い w_____	7. 空 s_____	7. ふくろう o_____	7. 歯磨き粉 t_____
8. 安全 s_____	8. 隠す h_____	8. 潜る d_____	8. カメ t_____
9. 農場 f_____	9. 歌手 s_____	9. 葬式 f_____	9. 滝 w_____
10. 深い d_____	10. 屋根 r_____	10. 泡 b_____	10. やなぎ w_____
11. 鉄砲 g_____	11. 牛 c_____	11. 雷 t_____	11. 弁護士 l_____
12. 科学者 s_____	12. 砂 s_____	12. 抱き合う h_____	12. ウナギ e_____
13. 雨 r_____	13. 野菜 v_____	13. はしご l_____	13. 天皇 e_____
14. 柔らかい s_____	14. 甘い s_____	14. 数学 m_____	14. 絶滅 e_____
15. 椅子 c_____	15. 草 g_____	15. 天井 c_____	15. 湿気 h_____
16. 女王 q_____	16. 骨 b_____	16. 熟した r_____	16. 熟語 i_____
17. 秘密 s_____	17. 親戚 r_____	17. まくら p_____	17. いかだ r_____
18. 機械 m_____	18. 葉 l_____	18. 豚肉 p_____	18. おみやげ s_____
19. 切符 t_____	19. たんぱく質 p_____	19.訳す t_____	19. 日光浴 s_____
20. 雑誌 m_____	20. 妊娠 p_____	20. 蝶 b_____	20. マグロ t_____
21. 鳥 b_____	21. 大使 a_____	21. 知らない人 s_____	21. 人類学者 a_____
22. 寝る s_____	22. 公害 p_____	22. 階段 s_____	22. 天文学 a_____
23. 穴 h_____	23. 冗談 j_____	23. 絨毯 c_____	23. 簿記 a_____

Appendix Table continued.

Write (single word) English translations for the following Japanese words.

24. 橋 b_____	24. 競馬の騎手 j_____	24. 襟 c_____	24. まつげ e_____
25. 夢 d_____	25. お面 m_____	25. 誘拐 k_____	25. 遠足 e_____
26. 鍵 k_____	26. 式 c_____	26. 香味料 s_____	26. タコ o_____
27. 温かい w_____	27. 城 c_____	27. 苺 s_____	27. 薬剤師 p_____
28. 燃える b_____	28. 茹でる b_____	28. 酸っぱい s_____	28. 学期 s_____
29. 人口 p_____	29. 鏡 m_____	29. 虹 r_____	29. 株式仲買人 s_____
30. 科学 s_____	30. 胃 s_____	30. かつら w_____	30. 竜巻 t_____
31. 約束 p_____	31. 黄色 y_____	31. 浅い s_____	31. 水槽 a_____
32. 気象 w_____	32. 認可 l_____	32. 垂直の v_____	32. 雪崩 a_____
33. 箱 b_____	33. 宝石 j_____	33. 巣 n_____	33. 学生寮 d_____
34. 二月 F_____	34. 俳優 a_____	34. 百合(ゆり) l_____	34. キリン g_____
35. 風 w_____	35. おもちゃ t_____	35. 怠け者 l_____	35. 旅程 i_____
36. 角 c_____	36. 塔 t_____	36. 動物園 z_____	36. ヒョウ l_____
37. 台所 k_____	37. 双子 t_____	37. 奴隷 s_____	37. サソリ s_____
38. 王 k_____	38. 盗む s_____	38. 奇跡 m_____	38. 継母 s_____
39. 明日 t_____	39. ささやく w_____	39. ろうそく c_____	39. ひまわり s_____
40. 電車 t_____	40. 隣人 n_____	40. 液体 l_____	40. 台風 t_____

Appendix 3. Translation test answer key

1. 島 island	1. 薬 medicine	1. 霧 fog	1. 2カ国語 bilingual
2. 動物 animal	2. 宮殿 palace	2. 糊(のり) glue	2. 赤道 equator
3. 住所 address	3. 英雄 hero	3. 靴下 socks	3. 化石 fossil
4. 地球 earth	4. 冬 winter	4. 青あざ bruise	4. エビ shrimp
5. 医者 doctor	5. 耳 ear	5. 誓う swear	5. いやみ sarcasm
6. 安い cheap	6. 嫌う hate	6. 頬 cheek	6. 迷信 superstition
7. 弱い weak	7. 空 sky	7. ふくろう owl	7. 歯磨き粉 toothpaste
8. 安全 safe	8. 隠す hide	8. 潜る dive	8. カメ turtle
9 農場 farm	9. 歌手 singer	9. 葬式 funeral	9 滝 waterfall
10. 深い deep	10. 屋根 roof	10. 泡 bubble	10. やなぎ willow
11. 鉄砲 gun	11. 牛 cow	11. 雷 thunder	11. 弁護士 lawyer
12. 科学者 scientist	12. 砂 sand	12. 抱き合う hug	12. ウナギ eel
13. 雨 rain	13. 野菜 vegetable	13. はしご ladder	13. 天皇 emperor
14. 柔らかい soft	14. 甘い sweet	14. 数学 mathematics	14. 絶滅 extinct
15. 椅子 chair	15. 草 grass	15. 天井 ceiling	15. 湿気 humidity
16. 女王 queen	16. 骨 bone	16. 熟した ripe	16. 熟語 idiom
17. 秘密 secret	17. 親戚 relative	17. まくら pillow	17. いかだ raft
18. 機械 machine	18. 葉 leaf	18. 豚肉 pork	18. おみやげ souvenir
19. 切符 ticket	19. たんぱく質 protein	19. 訳す translate	19. 日光浴 sunbathe
20. 雑誌 magazine	20. 妊娠 pregnant	20. 蝶 butterfly	20. マグロ tuna
21. 鳥 bird	21. 大使 ambassador	21. 知らない人 stranger	21. 人類学者 anthropologist
22. 寝る sleep	22. 公害 pollution	22. 階段 stairs	22. 天文学 astronomy
23. 穴 hole	23. 冗談 joke	23. 絨毯 carpet	23. 簿記 accounting
24. 橋 bridge	24. 競馬の騎手 jockey	24. 襟 collar	24. まつげ eyelash
25. 夢 dream	25. お面 mask	25. 誘拐 kidnap	25. 遠足 excursion
26. 鍵 key	26. 式 ceremony	26. 香味料 spice	26. タコ octopus
27. 温かい warm	27. 城 castle	27. 苺 strawberry	27. 薬剤師 pharmacist
28. 燃える burn	28. 茹でる boil	28. 酸っぱい sour	28. 学期 semester
29. 人口 population	29. 鏡 mirror	29. 虹 rainbow	29. 株式仲買人 stockbroker
30. 科学 science	30. 胃 stomach	30. かつら wig	30. 竜巻 tornado
31. 約束 promise	31. 黄色 yellow	31. 浅い shallow	31. 水槽 aquarium
32. 気象 weather	32. 認可 license	32. 垂直の vertical	32. 雪崩 avalanche
33. 箱 box	33. 宝石 jewel	33. 巣 nest	33. 学生寮 dormitory
34. 二月 February	34. 俳優 actor	34. 百合(ゆり) lily	34. キリン giraffe
35. 風 wind	35. おもちゃ toy	35. 怠け者 lazy	35. 旅程 itinerary
36. 角 corner	36. 塔 tower	36. 動物園 zoo	36. ヒョウ leopard
37. 台所 kitchen	37. 双子 twins	37. 奴隷 slave	37. サソリ scorpion
38. 王 king	38. 盗む steal	38. 奇跡 miracle	38. 継母 stepmother
39. 明日 tomorrow	39. ささやく whisper	39. ろうそく candle	39. ひまわり sunflower
40. 電車 train	40. 隣人 neighbor	40. 液体 liquid	40. 台風 typhoon