Vocabulary Learning and Instruction

Volume 4, Number 2,

December 2015

doi: http://dx.doi.org/10.7820/vli.v04.2.2187-2759

VLI Editorial Team

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

Vocabulary Learning and Instruction

Volume 4, Number 2, December 2015
doi: http://dx.doi.org/10.7820/vli.v04.2.2187-2759

Table of Contents

Letter from the Editor

Dear Readers,

It is with pleasure that we present to you another regular issue of *Vocabulary Learning and Instruction*. In the following pages you will find articles concerning word association studies, finding useful collocations to teach your students, and a new yes-no test scoring formula. You will also find a commentary on a recent article published in our October 2015 issue.

For any of our readers wishing/needing to upgrade their academic qualifications, we are also pleased to present descriptions of four SLA PhD programs which may be of particular interest to readers of VLI. Three focus on vocabulary: Cardiff University, the University of Nottingham, Victoria University of Wellington; plus one with a reading focus: Carnegie Mellon University.

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.

We hope you will enjoy the papers that follow and wish you a prosperous 2016.

Raymond Stubbe,

Editor, VLI

Vocabulary Learning and Instruction
Volume 4, Issue 2, December, 2015
http://vli-journal.org

Article

A Japanese Word Association Database of English

George Higginbotham[a], Ian Munby[b] and John P. Racine[c]
[a]*Hiroshima Kokusai Gakuin University;* [b]*Hokkai Gakuen University;*
[c]*Dokkyo University*
doi: http://dx.doi.org/10.7820/vli.v04.2.higginbotham.et.al

Abstract

In this paper, two word association (WA) studies are presented in support of recent arguments against the use of native-speaker (NS) norms in WA research. In Study 1, first-language (L1) and second-language (L2) WA norms lists were developed and compared to learner responses as a means of measuring L2 proficiency. The results showed that L2 norms provided a more sensitive measure of L2 lexical development than did traditional NS norms. Study 2 was designed to test the utility of native norms databases in predicting the primary WA responses of Japanese learners to high-frequency English cues. With the exception of only extremely frequent cues, it was shown that native norms were not successful in predicting learner responses. The results of both studies are discussed in terms of cultural and linguistic differences, geographic distance, and dissimilarities in word knowledge between respondent populations. Finally, a proposal is made for the construction of a Japanese WA database of English responses (J-WADE). The methods by which it will be developed, key features, and employment in future research are outlined.

Keywords: word association; database; native norms; non-native norms; L2 learners; vocabulary.

1 Introduction

It has long been held (Fitzpatrick, 2006, 2009; Henriksen, 2008; Meara, 1982; Racine, 2008, 2011a, 2011b) that responses to word association (WA) tests have the potential to provide rich information about the processes within the language learner's mental lexicon. As Meara (1996, p. 14) argues, WA data are particularly useful as it allows us to effectively tap into "two global characteristics: size and organisation". Consequently, databases of WA responses (Kiss, Armstrong, Milroy, & Piper, 1973; Moss & Older, 1996; Nelson, McEvoy, & Schreiber, 1998; Palermo & Jenkins, 1964; Postman & Keppel, 1970) have over the years been compiled and put to various purposes for examining the lexicons of language learners. Kruse, Pankhurst, and Sharwood Smith (1987), for example, used the 1952 Minnesota Norms List (see Jenkins, 1970) against which to compare a group of Dutch learners. In another study, Schmitt (1998) compared learner responses to the Edinburgh Associative Thesaurus (EAT; Kiss et al., 1973) norms list, compiled in the 1970s, as a way of creating an association score. Schmitt's WA score was one of a series of depth of word knowledge scores, used to measure the development of 11 words with three learners over the course of one year. Although Schmitt's findings were inconclusive, his investigation did reveal a general movement toward native-like

Please cite this article as: Higginbotham, G., Munby, I., & Racine, J.P. (2015). A Japanese word association database of English. *Vocabulary Learning and Instruction*, *4* (2), 1–20. doi: 10.7820/vli.v04.2.higginbotham.et.al

responses with increased proficiency, as had Kruse et al.'s (1987) study. While these and other L2 WA studies have presented intriguing results, we will argue below that the decision to utilize native-speaker (NS) response norms as a benchmark for investigating L2 associations is inexpedient.

Besides their use as a standard against which to measure L2 proficiency, native WA norms are also utilized in the selection of productive stimuli for further WA tests intended to measure the organization of learner lexicons (e.g., Higginbotham, 2010, 2014). When investigating lexical organization with WA tests, it is essential to avoid stimuli (such as *black*) that elicit a strong primary response (i.e., *white*), or *king* that usually elicits *queen*. Such stimuli, with excessively strong primary responses, mask individual response characteristics and therefore do not generate useful data. Since the use of such stimuli is unproductive, it is essential that researchers be able to identify them a priori when designing WA studies. Conventionally, native norms lists have been used for this purpose but we will propose below that using norms lists derived from the same community (i.e., more closely related, geographically and temporally) as the learners to be investigated will improve accuracy in identifying useful stimuli.

As we have argued elsewhere (Higginbotham, 2014; Munby, 2012; Racine, Higginbotham, & Munby, 2014), the validity and generalizability of findings in L2 WA studies employing native norms lists, may be called into question for a number of reasons. Our objections include the fact that WA studies involving native English respondents have found that NS responses are not homogeneous and vary over time (Fitzpatrick, 2007). While this may be expected from any population of respondents, this finding fails to support arguments for the use of native data based on its inherent homogeneity and stability across tasks. Importantly, it has also been demonstrated that – in the case of both English learners of Welsh, and Japanese learners of English – as L2 proficiency increases, L2 response profiles become more similar to subjects' own L1 profiles, rather than to NSs' responses (Fitzpatrick & Racine, 2014). This finding suggests that native norms are not the ideal comparative measure for examining L2 proficiency through WA responses. Further empirical evidence negating the utility of native norms comes from the field of language testing where non-native speaker (NNS) respondents have been shown to outperform NS respondents on certain tasks. McNamara (1996, Chapter 7) reports a number of studies involving standardized tests of English proficiency (e.g., IELTS) where NS scores were "neither homogeneous, nor high" (p. 191). As variation in native performance may be attributed to any number of factors – educational level, work experience, application of appropriate study skills, etc. – the author concludes that "reference to the NS as some kind of ideal or benchmark in scalar descriptions of performance on performance tests is not valid" (p. 197). Similarly, on a test of productive vocabulary, Meara (2009, Chapter 4) found that 18 of 48 NNS test-takers were able to outperform certain native respondents. At the same time, only 6 of 48 NS subjects were able to outscore the highest-achieving NNS participant. In at least some cases then, it is inadvisable to consider "native-like" proficiency as the end goal of language learning.

A sociolinguistic argument has also been made that English-L1 populations from whom normative response data has been compiled differ too greatly from the NNS populations to which comparisons are to be made (see Racine et al., 2014). These differences are both demographic and cultural/linguistic and render the

employment of NS norms data inadequate for L2 WA research. But even if the evidence and arguments outlined above were to be refuted, and a case made for the continued use of NS norms, it would still be unclear as to which variety of English norms should be adopted. English as a lingua franca (Seidlhofer, 2005), as an international language (Jenkins, 2000), and as a global language (Crystal, 2003), among others, have yet to be clearly defined at the lexical level. These English varieties will surely be distinguished by differences in word and collocation frequencies, spelling, usage, and meaning – as are the various global Englishes and dialects in existence today. To date, most WA norms lists have each been derived from a single variety of English. Traditionally, these have been gathered from British (e.g., Moss & Older, 1996) or American (e.g., Jenkins, 1970) respondents. It is not obvious which variety's norms would be most appropriate for making comparisons with Japanese learners' WA responses.

As further support for the argument against the use of native WA norms, this paper presents two L2 WA studies designed to explore the English associations of Japanese respondents. The first of these (Section 2; from Munby, 2012) involves the use of WA data in measuring L2 proficiency of Japanese learners of English. Study 2 in Section 3 (from Higginbotham, 2014) directly examines native WA norms and their utility in predicting Japanese learner responses. The results of both of these studies, as will become clear, point to the necessity of non-native norms for L2 WA research. Finally, building upon these findings, and in conjunction with the arguments we have made above against native norms in WA research, we will propose the construction of J-WADE in Section 5, outlining the methods by which it will be created and describing its key features.

2 Study 1: WA Norms as Measures of L2 Proficiency

2.1 Background

Given that it has been widely accepted that knowledge of a word's associations is an important aspect of L2 word knowledge (Nation, 2001; Richards, 1976), it would seem logical to predict that developments in a learner's lexical competence would be mirrored in the number and type of associations that a learner could produce in response to a set of stimuli. During the early 1980s, some commentators had assumed that a test could therefore be designed to measure the state of an L2 learner's associational networks which would reflect her level of proficiency. However, the study by Kruse et al. (1987) compared the associations produced by a group of Dutch third-year university students of English with a group of NSs of English in a test which used specially designed software to collect up to 12 responses to each of a set of 9 stimulus words. No significant difference between the two groups was reported when responses were measured against NS associative norms. This study became highly influential as it seemed to show that the free continuous WA test was not a valid proficiency measure.

In two replications of the Kruse et al. study (Munby, 2007, 2008), NSs outperformed NNSs on a multiple response WA test. Further, non-native test scores were found to correlate significantly with standard proficiency tests. The suggestion is that gains in learner proficiency are reflected to a certain extent in the

number and type of associations produced in response to a set of target words under timed conditions. However, the possibility remained that the methodology employed by Kruse et al., and in the two replication studies, did not live up to its potential for two reasons. First, the normative data used to measure both NS and NNS responses for stereotypy (see Jenkins, 1970) seemed inappropriate in this context. One of the issues was that these norms are outdated with the result that, for example, computer-related responses to cues like *memory* (e.g., *memory card*) could not earn points for stereotypy because many of these meanings were not common knowledge or did not even exist at the time the norms list was compiled in the 1950s.

A second issue for these norms – and a central concern of the current study – is the fact that they were derived from the WA responses of NSs while norms from highly proficient NNSs might have been more suitable in this situation. In other words, responses from proficient non-native respondents may provide a more appropriate point of comparison when examining the WAs of learners of English as an L2. The idea here is that the NNSs – in this case, Japanese learners – may be approaching the WA performance of highly proficient Japanese speakers of English, rather than that of NSs, as their proficiency level increases (see Fitzpatrick, 2009; Fitzpatrick & Racine, 2014). Schmitt and Meara (1997) point out that L2 learners will "have different mastery of the various kinds of word knowledge, with formal, grammatical, and meaning aspects probably learned first, and some other aspects, such as collocational behavior and register, perhaps never being mastered at all" (p.18). Collocational competence is one aspect of the ability to produce associations. Thus, if learners – even highly skilled ones – do not demonstrate native-like associational knowledge, it may be more appropriate to measure learner performance against the norms of proficient L2 users.[1]

Based on this reasoning, this study was designed with three key aims. The first was to compile a set of 50 new cue words to gather normative data for a new WA test (hereafter referred to as WAT50). The decision to begin with new cue words was motivated by a desire to limit the pool of candidate cue words to the 0–1K range of the British National Corpus (BNC; see Leech, Rayson, & Wilson, 2001). This would increase the likelihood that the cue words would be known to the non-native participants. The second aim was to compile two separate norms lists for this new set of cues with responses from two groups of participants: a group of NSs of English and a group of highly proficient non-native (L1 Japanese) users of English. In keeping with the tradition of naming norms lists after the locations where they are developed (e.g., Jenkins, 1970; Kiss et al., 1973; Moss & Older, 1996), these lists are now known as the *Sapporo L1 English Norms* and the *Sapporo L2 English Norms* (Munby, 2014). Finally, the third aim of this study was to run WAT50 with a group of learners, using these new norms lists for separate stereotypy scoring.

From these aims, the following research questions were formulated to guide this study:

RQ1 Which norms list, the Sapporo L1 English norms or the Sapporo L2 English norms, yields the best match with learner responses?

RQ2 Which norms lists, the Sapporo L1 English norms or the Sapporo L2 English norms, yields the highest correlations with proficiency?

2.2 Methodology

2.2.1. Cue word selection

In order to elicit an optimal sample of participant associative competence, cue words for WAT50 were selected according to the following criteria:

(1) They were to be known by all subjects. The cues *mutton* and *priest* were unknown to many subjects in Kruse et al.'s (1987) study. Indeed, in a replication of the study (Munby, 2007) many subjects were unable to produce any associations for these cues. Although the list of candidate words was restricted to the 0–1K band of the BNC, some of these words (e.g., *vote*) were unknown to many lower level participants. Personal intuition based on extensive experience of teaching in Japan was used to determine which words were likely to be known and which were not.
(2) They should not produce a dominant primary response, such as adjectives that elicit their antonyms (e.g., *high-low*) or gender-marked nouns (e.g., *king-queen*). Such cue–response pairs (also those described in (3) below) tend to be elicited from the majority of respondents and thus hold little value as a potential measure of proficiency.
(3) They were unlikely to generate responses through highly predicable lexical subset relationships (e.g., *fruit-apple*).
(4) They were not proper nouns. The 0–1K section of the BNC contains some proper nouns such as *Germany* and *America*.
(5) The stimulus was unlikely to elicit proper nouns (e.g., *river-Mississippi, city-Minneapolis, ocean-Pacific*).
(6) The stimulus was not a function word. Prepositions, for example, were eliminated because there was a likelihood that they would generate other function words as responses. Candidate items were reduced to the following word classes: nouns, verbs, adjectives, and adverbs.

After each of the 1,000 words was screened, only 125 candidate cue words fit all the above criteria. The final 50 cue words were chosen at random from the remaining set of 125 and were then screened for *overlap*. Overlap was defined as the phenomenon where a cue word shares, or is perceived to potentially share, an excessive number of responses elicited by another cue word. Also, common responses should not include other cue words. Thus, the final selection criterion was that none of the cue words elicited responses which were also listed as common responses to other cues according to the EAT (Kiss et al., 1973). A common response was defined as a response making up 6% or more of the total responses. For example, *body* elicits the response *soul* on 10% of occasions, which means the cue *heart*, producing *soul* on 7% of occasions, cannot be included in a set of cues that includes *body.* This proved very difficult to realize in practice, so exceptions were made of the following responses: *up* (6 cases), *out* (4), *of* (2), and *me* (2). Since three of these are prepositions (*up*, *out*, and *of*) and *me* is a pronoun, they did not present the risk of semantic overlap that was found with other cues such as *heart* and *body*. The final 50 randomly chosen cue words were replaced continually until all overlaps, with these few exceptions, were filtered out (see Table 1).

Table 1. Final List of 50 Cue Words

AIR	CHOICE	GAS	MEAN	SCIENCE
BEAR	CHURCH	HAPPEN	MOVE	SET
BECOME	CLASS	HEART	NATURE	SHARE
BLOW	CROSS	HOSPITAL	PACK	SORRY
BREAK	CUT	KEEP	PART	SPELL
BOAT	DRAW	KILL	POINT	STAGE
CALL	DRESS	KIND	POLICE	SURPRISE
CASE	FAIR	LEAD	POWER	TIE
CATCH	FIT	LINE	READY	WORLD
CHANCE	FREE	MARRY	RULE	USE

2.2.2. Participants

NSs of English were selected from among the personal friends and colleagues of one of the researchers.[2] As for the group of highly proficient non-native (L1 Japanese) users of English, the greatest challenge was ensuring that members were highly proficient. As prospective participants lived in many parts of Japan and abroad, it was not possible to conduct supervised L2 proficiency testing to justify their inclusion in the study. Instead, a set of criteria based on use of and experience with English was devised. Some had lived, or were living, in English-speaking countries for extensive periods. Others were teaching, or had taught, English. Others had never lived abroad nor taught English but had acquired high degrees of fluency through: (1) using English for academic purposes, such as scientific researchers who publish papers in English, (2) using English professionally in the international workplace, such as EFL publishing representatives, or (3) using English in their daily life (e.g., with English-speaking spouses). The resulting definition of a highly proficient Japanese user of English was a person who:

(1) had lived or was living abroad in an English-speaking country for a year or more, or
(2) was teaching English or had taught English, or
(3) had extensive experience using English socially, in the international workplace, or for academic purposes.

The final pool of 114 L2 subjects was drawn from among friends, colleagues, and professional organizations such as JALT (Japan Association of Language Teachers) and JACET (Japan Association of College English Teachers). A profile of the subject group can be seen in Table 2.

2.2.3. Compiling the norms lists

WA task forms were sent to 114 NSs of English and 114 highly proficient non-native (L1-Japanese) users of English via e-mail attachment. In the task instructions, participants were asked to provide five English responses to each cue, avoiding proper nouns and multi-word responses. They were requested not to worry

Table 2. Profile of Study 1 Participants

		L2	L1
	Total (N = 228)	n = 114	n = 114[a]
Age	Average age	43	47
Gender	Male	38	76
	Female	76	38
Country of residence	Resident in Japan	99	69
	Resident outside Japan	15[b]	45
Highest level of education	University graduates	110	110
	High School graduates	4	4
Dominant occupation	Teacher	70	88
Number of L2 participants who were living or had lived in an English speaking country for a year or more		85	
Mean number of years spent in English-speaking countries		4.8	
Number who were teaching or had taught English		88	
Number who often use English for academic purposes		95[c]	
Number who often use English with family or friends		56	
Number who often use English for business		68	

[a]By nationality, the breakdown of the L1 group was: USA (34), Canada (33), Britain (32), Australia (13), Ireland (1), and New Zealand (1).
[b]The 15 L2 participants living outside Japan live in the following countries: Canada (6), USA (3), Britain (2), Indonesia (1), Brazil (1), Germany (1), and Samoa (1).
[c]Includes those who were teaching English.

about making spelling or typing errors and to refrain from consulting dictionaries, online references tools, or friends. Misspelled items were corrected.

2.2.4. The WAT50 methodology

The participant group was comprised of 82 English majors at a private university in northern Japan. Two proficiency measures were employed: (1) the TOEIC test of listening and reading comprehension and (2) a 50 item cloze test. The same methodology employed by Kruse et al. (1987) was utilized here: subjects entered up to 12 responses for each of the 50 cues and two practice cues. Cues were displayed via the same computer software utilized in the replication studies (Munby, 2007, 2008). This included a timer which allowed participants 30 seconds of thinking time per cue. The timer deactivated while participants were typing so as not to disadvantage those with slow typing speed. Scores were tallied for total number of responses entered for the 50 cues, and for stereotypy. The stereotypy measure was a count of the total number of responses that matched responses on the two Sapporo norms lists. Finally, these scores were then compared with the language proficiency measures.

2.3 Results

RQ1: Which norms list, the Sapporo L1 English norms or the Sapporo L2 English norms, yields the best match with learner responses?

The data in Table 3 indicate that there is a better match between subjects' responses and the Sapporo L2 English norms lists (mean stereotypy score = 184.3), than subjects' responses and the Sapporo L1 English norms (159.3). Results of a one-tailed paired *t* test produced a statistically significant *t* value of 9.45 ($p < 0.0001$). That this difference was significant was in keeping with the fact that every non-native subject scored a higher stereotypy score with the Sapporo L2 English norms list than with the Sapporo L1 English norms list.

RQ2: Which norms list, the Sapporo L1 English norms or the Sapporo L2 English norms, yields the highest correlations with proficiency?

Although it is worth noting that the correlations between stereotypy scores and proficiency were broadly similar, correlations were marginally higher for the Sapporo L1 English norms stereotypy measure (see Table 4). The TOEIC test produced higher correlations with the WAT50 measures than with the cloze test scores.

2.4 Discussion

The main purpose of the study was to design an improved multiple response WA test, the WAT50, by rectifying weaknesses apparent in the original probe by Kruse et al. (1987). In doing so, the aim was to establish the optimal conditions for the new association test to reflect level of proficiency with adult Japanese learners of English. The new norms lists clearly represented an improvement on Jenkins's (1970) list in terms of their utility when making comparisons to contemporary responses. As expected, a large number of learner-generated responses (in both the L1 and L2 norms) were related to contemporary consumer items or fashions which did not even exist when Jenkins's list was first published in 1952 (e.g., *call-cell-phone, pack-ziplock, tie-dye, pack-CD, use-computer*). The findings from the measure of stereotypy outlined in Table 3 provide support for our argument above (see also Racine et al., 2014) that the utility of comparing norms lists to learner data is directly related to the relative proximity of the populations from which the data were derived. Proximity between respondent groups may be measured in terms of geographical, cultural, and linguistic differences or, as seen here, in terms of temporal ones.

Table 3. Mean Scores, Standard Deviations, Highest & Lowest Scores and Maximum for all Scoring Methods of the WA Test and Proficiency Measures ($N = 82$)

	Mean	*SD*	High	Low	Maximum
No. of responses	269.4	107	578	89	600
L1 Stereotypy	159.3	51.4	300	60	600
L2 Stereotypy	184.3	58.8	377	71	600
TOEIC	539.2	137	935	300	990
Cloze	18.5	7.2	40	5	50

This finding begs the question: Why do learner responses on this test yield more matches with the L2 norms than the L1 norms list? This finding (Table 3) is especially puzzling in view of the fact that the L1 lists feature a larger total number of different responses to 37 of the 50 cue words. While the prompt *ready* elicited exactly the same number of different responses from each normative group (150), the L2 group produced a larger number of different responses to only 12 of the cue words. One reason for this is that there are a number of responses on the L1 norms list that were not elicited from either the learners or the highly proficient L2 respondents. Animal-related responses to *pack*, such as *wolf* or *mule*, are examples of this class of exclusively native response and reflect the breadth of the native lexicon as well as the heterogeneity of NSs' responses seen in prior WA studies (see Fitzpatrick, 2007). Conversely, many responses elicited from the learners appear on the L2 norms lists but not on the L1 lists. For example, in response to *spell*, the form-based response *misspelling* appears on the L2 lists, but not on the L1 response lists. Further, although a large number of the native L1 participants (69 out of 114) were living in Japan and are familiar with the culture and language, they appeared not to respond in a Japanese-like way. For example, the response *typhoon* to the cue *blow* was often provided by Japanese learners of English, and was listed among the L2 norms, but did not appear in the L1 norms. In this way, there may be some truth in the claim made by Kruse and his associates that the WA test is influenced by "problems such as . . . the effects of cultural background knowledge" (1987, p. 153). This too points to the necessity for reliable non-native norms generated from the same community of respondents as those whose L2 proficiency researchers wish to examine.

The correlation scores between the norms lists and the proficiency measures (Table 4), on the other hand, appear not to support the arguments we have made above. One reason for this contradictory finding may stem from differences in the nature of these particular measures of proficiency and the kind of word knowledge elicited in WA tasks. The TOEIC test for example, is a receptive/passive test of English language proficiency. The reading and listening sections administered here are quite different from tasks involving productive vocabulary knowledge such as that required in making WAs. Likewise, cloze tests may be useful in measuring a learner's ability to map orthographic form to meaning in sentence-reading exercises. However, this too differs from the kind of lexical ability that may be tapped through associative measures. Indeed, had an association-based measure of proficiency (e.g., Read, 1993, 2004) been employed in this study, the resulting

Table 4. Pearson Correlations between WA Test Scores and Proficiency Measures

	CLOZE	TOEIC
No. of responses	.389**	.433**
L1 stereotypy	.562**	.601**
L2 stereotypy	.523**	.563**

1-sided *p*-value: Significant at $^{**}p < 0.01$.

correlations between the norms and proficiency scores may have been stronger and the scores more meaningful in their support for the current proposal.

3 Study 2: Predicting Learner Responses From Native WA Norms

3.1 Background

An essential element of WA studies examining the learner's lexical organization is the selection of 'productive' cue words. That is, stimuli intended to generate responses that accurately portray the response profiles of the respondents (see Fitzpatrick, 2007, 2009; Higginbotham, 2010). As explained above (see also Meara, 1982), the problem is that some cues are strongly associated with just one other word. *Hard*, for example, will generally produce the response *soft*, regardless of the age, language proficiency, or educational background of the respondent. Similarly, the stimulus *cat* is likely to elicit *dog*. Such stimuli are unhelpful for researchers attempting to use WA responses to determine specific characteristics of subjects' response profiles. The problem that needs to be addressed is how to separate productive cues from these other words.

Given that many of the unproductive stimuli are highly frequent words, one solution is to select prompt words from less-frequent bands. Fitzpatrick (2006), for example, did precisely this in choosing words from the Academic Word List (AWL; Coxhead, 2000). The AWL excludes highly frequent items such as those in West's (1953) General Service List and therefore consists of mid-frequency and semi-technical words. As Fitzpatrick's respondents were high-ability learners, choosing cues in this manner proved to be an effective method of dealing with the cue-selection problem. If researchers are interested in the association behavior of learners with low or moderate ability, however, then it may be necessary to select cues from higher frequency bands. As the majority of learners fall into this category, many researchers continue to face the issue of how best to select productive cues for their WA studies. While some choose to run time-consuming pilot studies to determine which cues are most appropriate for use in further research, cross-referencing potential stimuli against an established WA database remains the simplest method of determining the primary response strength of potential cues. As we have outlined above, however, the vast majority of WA databases, whether online (e.g., Kiss et al., 1973; Nelson et al., 1998) or in print format (e.g., Moss & Older, 1996; Palermo & Jenkins, 1964; Postman & Keppel, 1970), are based on the responses of native English-speaking respondents. Taking the above arguments against the use of native norms and the results of Study 1 into account, it seems that the use of these norms lists is not the ideal method of determining which cues would be most productive in studies involving Japanese learners of English. The current study was conducted to test this assumption directly. If it could be shown that native norms databases accurately predicted unproductive cue words in WA tests, then the time-consuming process of running pilot studies to eliminate these cues would be dispensed with, and the hypothesis that NS WA norms were useful in this regard would be supported.

3.2 Methodology

Within the context of a larger study of lexical organization (Higginbotham, 2014), the WA responses of 30 Japanese learners of English (19 female, 11 male; TOEIC scores ranging from 550 to 750) were compared to the NS norms in the EAT database (Kiss et al., 1973). Two lists of cues were assembled, each containing 40 adjectives selected from the BNC (see Leech et al., 2001). The first list of prompt words (PWL1) was selected from the most frequent 1000 words of the BNC, while the second list (PWL2) was selected from the 1500–2000 range. All prompts were thus relatively frequent words, ensuring that the majority would be understood by the participants. Results of the Vocabulary Levels Test (Nation, 1990) verified this assumption as subjects achieved a mean score of 91.4% ($SD = 8.2$) on the 2000-level of the test, and 72.4% ($SD = 18.2$) on the 3000 level of the test. While the VLT is not a direct measure of knowledge of words in the BNC, there is considerable overlap with the word list it was based on. Analyzed by way of an online lexical profiler (Cobb, 2014), it was found that all 60 of the items from the 2000 section of the VLT were found within the BNC's most frequent 3000 words. Ninety-five percent of the sixty items in the VLT's Section 3000 fell within the first 4000 BNC words. The remaining five percent (i.e., three words) appeared less frequently in the BNC (see Table 5). As with the TOEIC test employed in Study 1, it may be noted that the VLT is a measure of 'receptive' vocabulary knowledge and therefore perhaps not an ideal measure to compare with a productive free WA test. In the absence of a widely accepted and standardized test of productive vocabulary, however, Nation's VLT was deemed a suitable indicator of learners' ability to respond to the words used in the WA tests.

For this study, *unproductive* prompts were operationalized as any cue words for which the primary response constituted more than 25% of all the responses to that cue. With that criterion in mind, the learners' primary responses were compared to those in the database of native norms. If the same cues were considered unproductive on the basis of the primary responses in the norms list and in the learner data, we could say that the norms list effectively 'predicted' that the cues were unproductive and need not be included in the larger study of lexical organization. Further, if enough unproductive cues were correctly identified in this manner, then the use of native norms lists could still be considered a valid tool for this purpose.

Table 5. Percentage of VLT Test Items within the First Five Frequency Bands of the BNC (Derived from Leech et al., 2001)

	BNC band				
VLT items	K1	K2	K3	K4	K5
Section 2000	18	60	22	–	–
Section 3000	7	23	35	30	5

For a prediction about a given cue to be considered *correct*, two criteria had to be fulfilled:

(1) the primary response on the native norms list had to match the learners' primary response, and
(2) this primary response had to be categorized as either *productive* or *unproductive* (constitute more or less than 25% of the total responses) according to *both* the norms list and the learner data.

This is exemplified by the cue *possible* (see Table 6) which elicited the primary response *impossible* from both the EAT respondents and from the Japanese learners. In both cases *impossible* constituted more than 25% of responses from their respective respondents and were thus categorized as unproductive cues. Predictions were scored as *partially correct* if either of these two criteria were met. That is, the primary response matched, but it was shown to be productive in one set of data, but unproductive in the other (e.g., the cue *equal* in Table 6); or the primary responses did not match, yet both were either productive or unproductive (e.g., *social* in Table 6). Predictions were considered *incorrect* if neither of these two conditions were met (e.g., *used* in Table 6).

3.3 Results

As seen in Figure 1, the norms list was able to quite accurately predict the learner responses to the prompts in PWL1 (41% correct; 52% partially correct). However, it should be noted that the PWL1 cues represent only the most frequent words of the English language (from the first 1000 words of the BNC). The results for the PWL2 cues (from the 1500 to 2000 range of the BNC) show a notable decrease in accuracy. While these prompts are slightly less frequent than those in PWL1, they are still considered to be very frequent by most language learning standards. This suggests that only the most rudimentary cue words will elicit the same primary responses from Japanese learners as from native-speaking respondents. These findings imply that native norms lists are of very little utility in predicting Japanese learner responses. Indeed, as can be seen on the right side of Figure 1, many of the responses generated by the PWL2 cues did not even appear in

Table 6. Example Prediction Scores of Productive and Unproductive Cues

	Primary response		Percentage of total responses		
Cue	Native norms[a]	Japanese learners	Native norms[a]	Japanese learners	Prediction score
possible	*impossible*	*impossible*	38 (unproductive)	55 (unproductive)	Correct
equal	*same*	*same*	15 (productive)	33 (unproductive)	Partially correct
social	*science*	*network*	22 (productive)	18 (productive)	Partially correct
used	*car*	*old*	19 (productive)	32 (unproductive)	Incorrect

[a]From the EAT (Kiss et al., 1973).

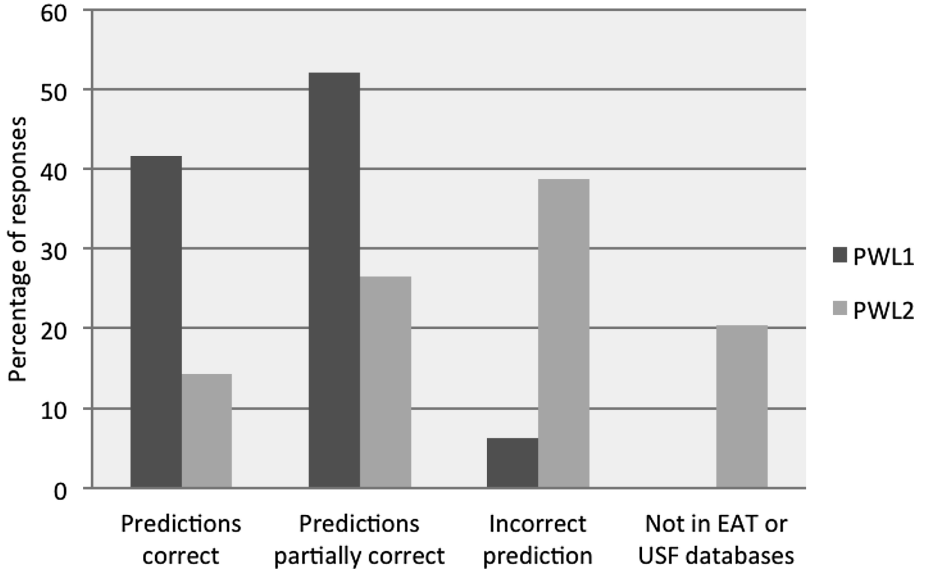

Figure 1. Prediction scores for two groups of prompt words.

the EAT database or the University of Southern Florida (USF; Nelson et al., 1998) norms examined here.

3.4 Discussion

The main findings of the study (Figure 1) confirmed the suspicion that native norms lists were inadequate for the purposes of identifying useful prompt words. There are two issues in particular with these sets of native norms that appear to make them inaccurate predictors of non-native responses. First, these norms lists are outdated. Three of the top 20 responses for the cue *help* in the EAT norms were *beatles*, *beatle*, and *beetle*. Presumably, such responses were based on knowledge of the Beatles song and movie entitled "Help!" which was popular at the time the EAT data was collected. Almost 50 years since the Beatles disbanded, it is unlikely that these would still be such popular WA responses today.

The age of these norms lists notwithstanding, a second and more fundamental problem is that they appear to lack validity as useful measures in investigations of learner responses in EFL (English as a Foreign Language) contexts. Students studying English in Japan (as were the respondents in this study) receive quite a different exposure to the language than do ESL (English as a Second Language) students who learn the language while living in English-speaking countries. Such exposure to the target language differs not only in quantity but also in terms of the types of language encountered – EFL English is often limited to English in classrooms, while ESL typically offers both classroom English and real-life language experience. This point is illustrated in responses to the cue *cup*. While the word was likely to have been known by all the L2 learners in this study, none of

them associated it with the word *saucer*. This is contrary to what might be predicted from an examination of the EAT norms – or from more recent databases such as Hirsch and Tree (2001). While *cup and saucer* is a common collocation in the UK where the EAT norms were collected, it may be inappropriate to judge the development of a learner's lexicon based on knowledge of this phrase in other contexts. This has been illustrated in a recent cross-cultural study (Son et al., 2014) in which French respondents more frequently associated *rice* with concepts related to foreign countries, foreign cultures, and travel, while Asian respondents tended to associate it with agricultural products and necessary food items. Likewise the iconic status of the Beatles in British musical history makes it more likely that UK respondents will reply with the group's name to a cue like *help*. This link is less salient in the minds of Japanese respondents, as observed above.

In conclusion, it would appear that currently used native norms lists suffer from both temporal and cultural mismatches when used as a standard by which to evaluate contemporary learner data. That said, researchers may wish to continue using them as a very rough guide for determining productive stimuli for WA studies. It should be clear from the current study however, that their utility is limited to only very frequent words. The screening process for selecting mid-and low-frequency cues should not rely exclusively upon native norms lists like those currently employed. Researchers may however choose to continue using native norms lists in a two-step process (e.g., Higginbotham, 2014) where norms lists are first used to filter out cues with extremely strong primary responses (e.g., $>50\%$) and then subsequent testing is employed to further identify unproductive prompts.

4 Summary

We have presented two very different WA studies designed to examine the utility of NS norms in L2 WA research. Study 1 (Section 2) involved the development of the WAT50, a 50-cue WA test intended for use as a measure of L2 proficiency. For the purposes of comparison, two norms lists were created (Munby, 2014): one from a group of NSs of English (L1) and another from a group of highly proficient non-native (i.e., Japanese) users of English (L2). Results showed that both of these norms lists performed better than a traditional native norms list (from Jenkins, 1970) when investigating responses elicited from language learners today. This finding was attributed to the fact that many contemporary responses were derived from word knowledge – in particular, collocational knowledge – of expressions that either had not existed or were not common knowledge when the Jenkins norms were collected. At the same time, the new L2 norms better matched the learner data than had the new L1 norms (Table 3). An analysis of the responses showed that certain responses were only produced by native-speaking respondents (e.g., *pack-wolf*). These appear to reflect very specific aspects of semantic knowledge for these cues. Other responses were elicited only from the Japanese respondents (e.g., *blow-typhoon*). These too appear to reflect the salience of specific types of word knowledge (perhaps influenced by their L1) or *world* knowledge reflecting their geographical and temporal location.

In Study 2 (Section 3 above), native norms (from Kiss et al., 1973) were tested directly for their ability to predict primary responses in L2 learner data. With the

exception of only extremely frequent cue words, it was found that the native norms were unable to accurately predict these responses. Two reasons for this were postulated. First, it appeared that the database of native norms was outdated. Many responses to the cue *help*, for example, were related to the long-disbanded group The Beatles. Second, as observed in Study 1, responses seem to reflect word knowledge specific to the context in which respondents reside. NSs and learners in ESL contexts are exposed to substantially more, and different, types of English than are learners in EFL contexts such as the Japanese learners of English examined here. Clear conceptual differences found across cultures in internationally-conducted association studies (e.g., Son et al., 2014) provide support for this conclusion.

The results of these two studies, in conjunction with the arguments we have made against the use of native WA norms here and elsewhere (Fitzpatrick & Racine, 2014; Racine et al., 2014), make it clear that linguistic researchers are in need of a comprehensive set of L2 learner norms for WA research purposes. As demonstrated above, the better match between L2 learners' responses and the associative norms of their higher-level peers, may prove invaluable in tracing changes in learners' associative behavior over time. In this way – through longitudinal studies – L2 normative data may shed more light on the development of the L2 learner lexicon than could norms lists drawn from NSs. This is one of the many potential advantages to the use of L2 WA norms. It is with this thought in mind that we present the proposal below for a word association database of English responses elicited from high-ability Japanese learners.

5 A Japanese Word Association Database of English (J-WADE)

The design and construction of J-WADE will involve four basic steps to be implemented over a period of at least three years:

(1) the selection of stimuli,
(2) the creation of an online survey page and data entry website,
(3) data collection by the authors (and solicitation of broader participation), and
(4) the creation of a results site (and publishing the results).

The first consideration for the J-WADE project is the decision as to which stimuli to utilize as association cues. We have outlined here and elsewhere (e.g., Fitzpatrick & Munby, 2014; Higginbotham, 2010; Racine, 2013) some of the considerations for the selection of appropriate cues for various populations of respondents. With these considerations in mind, it is likely that the first rounds of data collection will be from one of the new general service lists (Brezina & Gablasova, 2015; Browne, 2014). Selection from these high-frequency words would insure that the majority of learner-respondents would already be familiar with the cues. Cues will also be selected from the most frequent bands of the New Academic Vocabulary List (Gardner & Davies, 2014). Ongoing examination of respondent data would yield lists of the most frequent responses from which new cues could also be selected. Responses gathered from these 'responses' may yield a richer picture of

participants' lexical networks. In total, approximately 5,000 to 10,000 words will be utilized as cues over the course of at least three years of data gathering. This figure brings the scope of the project in line with prior WA databases such as Kiss et al. (1973; 5,019 cues) and Nelson et al. (1998; 8,400 cues). At least 100 responses are to be collected for each cue word and approximately 100 cues will be presented to each participant.

In the early stages of the project, two websites will be created. The first website will be a simple web-based WA survey form that will allow native-Japanese respondents – predominately university students and other adults – to go online wherever they are to take part in the survey. Cues will be presented in accordance with the principles of psycholinguistic research: presenting each word on screen individually and in random orders across subjects to avoid the influence of priming and order effects. Responses to the cues and demographic data will be collected automatically and uploaded to the database directly from the website. The second website to be developed will allow manual data entry of responses and demographic information from respondents who have completed the WA task in paper-based format. Introducing survey forms on paper, in addition to the online survey form, will allow the researchers and their colleagues to administer the forms to classes of university student-learners and groups of other respondents en masse. Data collection will begin in the first year and continue throughout the course of the project. The researchers will travel to various universities in Japan to solicit the cooperation of teachers and researchers. Cue lists will be continuously updated and responses will be collected wherever willing participants are encountered.

In the second and third years of the J-WADE project, data collection will remain ongoing. The researchers will continue to travel, soliciting participants wherever they can and publicizing preliminary results as they become available. An assistant will be hired to begin data entry of the many paper-based response forms that will have accumulated by this time. One of the final steps in the completion of the project will be to create a third website that will allow other researchers to access the findings. The site will be fully searchable, allowing interested researchers to search results via cue words and responses. Results will be further filterable by a variety of linguistic factors (e.g., grammatical class, number of orthographic neighbors) and in terms of demographic information (e.g., age, gender, language proficiency). With the results site online, the researchers will continue to publicize the findings and encourage other researchers to use the website and results for the purposes of their own research.

6 Conclusions and Further Research

The creators of the native norms lists employed in the experimental studies presented here had likely intended their lists to be used in research into native language development. They may never have anticipated that their data might someday be put into service by L2 acquisition researchers as they were here. Long after its completion, we may also find that J-WADE has been employed for purposes that we cannot now anticipate. One possible offshoot from its construction would be the development of separate norms lists for English language learners from a variety of different L1 backgrounds. One can imagine, for example, the

creation of 'D-WADE' consisting of the responses from Dutch learners of English, a line of research building on the Dutch L1 WA work of De Deyne and Storms (2008) among others. A comparison of various -WADE databases may reveal universal properties of English lexical acquisition observed across a variety of first languages. Conversely, comparisons of J-WADE with responses collected from Japanese learners of other foreign languages such as French (J-WADF) or German (J-WADG) may reveal properties of L2 acquisition and lexical organization characteristic of Japanese learners specifically. It may also prove fruitful to compile a database of Japanese (L1) responses to English (L2) prompt words. Results of a comparison of these responses to J-WADE norms may contribute to our knowledge of how the L1 and L2 networks relate in the bilingual lexicon of Japanese learners.

But first things first. Many aspects of the J-WADE project plan are contingent upon the availability of research funding (e.g., professional computer programming, solicitation of respondents from across Japan). The authors are currently seeking a grant through the Japan Society for the Promotion of Science for this purpose. Although the funding period and the plan outlined here involve three years of research it is likely that data collection will continue for an indefinite period of time thereafter. It is expected that this will result in a very broad set of data that will prove useful for the findings it uncovers and for its utility in further research. Due to the breadth of this plan, we wish to encourage interested Japanese users of English to contact the authors about becoming participants in this study. Likewise, we would appreciate the cooperation of all instructors teaching Japanese learners of English. Please make contact concerning how to get your students to participate in this project.

Notes

1. While still not widely accepted in contemporary WA research, this observation – that it may be more appropriate to adopt non-native norms than native ones when examining learner data – was made by Meara more than 30 years ago: "Teaching a language aims to produce people who are bilingual, not mere replicas of monolingual speakers. It would, therefore, be more appropriate to compare the associations of learners with those of successful bilingual speakers, and not with native speakers" (1982, p. 31).While this quotation from Meara has at its base second language pedagogy, it should be noted here that the utilization of the norms discussed in this study and those of the database proposed below will not typically occur in the language classroom. It is the aim of the authors to construct a norms database that will find its utility among applied linguists and language testing researchers. Those investigating the assessment of second language word knowledge, in particular, may find the most value in this project.

   For this reason, it should also be noted that, by nature, the level of L2 proficiency of respondents from whom normative data is to be collected will necessarily be determined by the levels of available participants. One can foresee, then, that certain learners' levels may exceed those of participants whose data make up the norms lists to which their responses are to be compared. This is not an indictment of L2 norms data, but one of the practical issues surrounding their use. The existence of 'incorrect' or 'non-proficient' responses among norms data from otherwise proficient L2 learners does not mean that learners should attempt to emulate such mistakes. WA norms are not pedagogical tools. They are to be used as a yardstick by which proficiency may be measured. It may very well be the case that extremely proficient test-takers' responses will differ from the normative data to be accumulated here, to the same degree as would responses from participants with very low levels of proficiency. That the norms might predict this occurrence is an endorsement for their adoption.

2. It should be acknowledged here that – while building an argument against the utilization of traditional native norms – the current studies, by necessity, involve the collection of native norms data. It should also be noted, however, that at least some of the arguments against traditional native norms lists (e.g., that they were gathered from populations of solely UK or US residents) have been addressed here. This is evident in the profile of NS participants in Table 2.

References

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, *36*(1), 1–22. doi:10.1093/applin/amt018

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction*, *3*(2), 1–10. doi:10.7820/vli.v03.2.browne

Cobb, T. (2014). *Compleat lexical tutor (v.8)*. Retrieved December 2, 2014, from http://www.lextutor.ca/

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

Crystal, D. (2003). *English as a global language* (2nd ed). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511486999

De Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, *40*, 198–205. doi:10.3758/BRM.40.1.198

Fitzpatrick, T. (2006). Habits and rabbits: Word associations and the L2 lexicon. *EUROSLA Yearbook*, *6*, 121–146. doi:10.1075/eurosla.6.09fit

Fitzpatrick, T. (2007). Word association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics*, *17*(3), 319–331. doi:10.1111/j.1473-4192.2007.00172.x

Fitzpatrick, T. (2009). Word association profiles in a first and second language: Puzzles and problems. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second language learners* (pp. 38–52). Bristol, UK: Multilingual Matters.

Fitzpatrick, T., & Munby, I. (2014). Knowledge of word associations. In J. Milton & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge* (pp. 92–105). Basingstoke, UK: Palgrave Macmillan.

Fitzpatrick, T., & Racine, J. P. (2014). *Using learners' L1 word association profiles as an alternative to native speaker norms*. Paper presented at the AILA World Congress, Brisbane, Australia.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*, 305–327. doi:10.1016/j.system.2010.06.010

Henriksen, B. (2008). Declarative lexical knowledge. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Vocabulary and writing in a first and second language* (pp. 22–62). Basingstoke, UK: Palgrave Macmillan.

Higginbotham, G. (2010). Individual learner profiles from word association tests: The effect of word frequency. *System*, *38*(3), 379–390. doi:10.1016/j.system.2010.06.010

Higginbotham, G. (2014). *Individual profiling of second language learners through word association* (Unpublished doctoral dissertation). Swansea University, Swansea, UK.

Hirsch, K. W., & Tree, J. T. (2001). Word association norms for two cohorts of British adults. *Journal of Neurolinguistics*, *14*(1), 1–44. doi:10.1016/S0911-6044(00)00002-6

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford, UK: Oxford University Press.

Jenkins, J. J. (1970). The 1952 Minnesota word association norms. In L. Postman & G. Keppel (Eds.), *Norms of word associations* (pp. 1–38). New York, NY: Academic Press.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In J. Aitken, R. W. Bailey, & N. Hamilton Smith (Eds.), *The computer and literary studies* (pp. 153–165). Edinburgh, UK: Edinburgh University Press.

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

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*. Harlow, UK: Longman.

McNamara, T. F. (1996). *Measuring second language performance*. London, UK: Longman.

Meara, P. (1982). Word associations in a foreign language. *Nottingham Linguistic Circular*, *11*(2), 28–38.

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

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

Moss, H., & Older, L. (1996). *Birkbeck word association norms*. East Sussex, UK: Psychology Press.

Munby, I. (2007). Report on a free continuous word association test. *Gakuen Ronshu*, *The Journal of Hokkai-Gakuen University*, *132*, 43–78.

Munby, I. (2008). Report on a free continuous word association test. Part 2. *Gakuen Ronshu*, *The Journal of Hokkai-Gakuen University*, *135*, 55–74.

Munby, I. (2012). *Development of a multiple response word association test for learners of English as an L2* (Unpublished doctoral dissertation). Swansea University, Swansea, UK.

Munby, I. (2014). *Sapporo word association norms lists*. Retrieved October 11, 2014, from http://sapporowordassociationnormslists.wordpress.com/

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

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

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* Retrieved from http://www.usf.edu/FreeAssociation

Palermo, D. S., & Jenkins, J. J. (1964). *Word association norms: Grade school through college.* Minneapolis, MA: University of Minnesota.

Postman, L., & Keppel, G. (Eds.). (1970). *Norms of word association*. New York, NY: Academic Press.

Racine, J. P. (2008). Cognitive processes in second language word association. *JALT Journal*, *30*(1), 5–26. Retrieved from http://jalt-publications.org/jj/issues/2008-05_30.1.

Racine, J. P. (2011a). Grammatical words and processes in the L2 mental lexicon: A word association perspective. *Studies in Foreign Language Teaching*, *29*, 153–197.

Racine, J. P. (2011b). Loanword associations and processes. *OTB – The Tsukuba Multi-lingual Forum, 4*(1), 37–44.

Racine, J. P. (2013). The history and future of word association research. *Dokkyo Journal of Language Learning and Teaching*, *1*, 55–73.

Racine, J. P., Higginbotham, G., & Munby, I. (2014). Exploring non-native norms: A new direction in word association research. *Vocabulary Education and Research Bulletin*, *3*(2), 13–15.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355–371. doi:10.1177/026553229301000308

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 209–227). Amsterdam, The Netherlands: John Benjamins.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*(1), 77–89. doi:10.2307/3585941

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

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*(1), 17–36. doi:10.1017/S0272263197001022

Seidlhofer, B. (2005). English as a lingua franca. *ELT Journal*, *59*(4), 339–341. doi:10.1093/elt/cci064

Son, J.-S., Do, V. B., Kim, K.-O., Cho, M. S., Suwonsichon, T., & Valentin, D. (2014). Understanding the effect of culture on food representations using word associations: The case of "rice" and "good rice". *Food Quality and Preference*, *31*, 38–48. doi:10.1016/j.foodqual.2013.07.001

West, M. (1953). *A general service list of English words.* London, UK: Longman.

Vocabulary Learning and Instruction
Volume 4, Issue 2, December 2015
http://vli-journal.org

Article

On Using Corpus Frequency, Dispersion, and Chronological Data to Help Identify Useful Collocations

James Rogers[a], Chris Brizzard[a], Frank Daulton[b], Cosmin Florescu[c], Ian MacLean[a], Kayo Mimura[a], John O'Donoghue[d], Masaya Okamoto[e], Gordon Reid[a] and Yoshiaki Shimada[f]
[a]*Kansai Gaidai University;* [b]*Ryukoku University;* [c]*University of New England;* [d]*Osaka Board of Education;* [e]*University of Manchester;* [f]*State University of New York at Albany*
doi: http://dx.doi.org/10.7820/vli.v04.2.rogers.et.al

Abstract

This study analyzed corpus data to determine the extent to which frequency, dispersion, and chronological data can help identify useful collocations for second language learners who aim to master general English. The findings indicated that although various analysis levels of frequency and dispersion data are largely effective, the analyses could not identify useful collocations reliably. The findings also indicated that chronological data analysis is not as useful as dispersion analysis due to the amount of time it took versus the improvements that resulted from it. Ultimately, it was found that a manual analysis of data using native speaker intuition is unavoidable. This study highlighted the value and reliability of certain types of corpus data analysis, and also the necessity of labor-intensive, native speaker analysis for identifying useful collocations.

Keywords: corpus; frequency; dispersion; collocations; multi-word units; formulaic sequences.

1 Introduction

Comprehending and producing collocations is an essential skill for native-like fluency (Durrant & Schmitt, 2009; Wray, 2002). Knowledge of collocations helps the language learner sound more native-like and process language more efficiently (Nation, 2001a; Snellings, van Gelderen, & de Glopper, 2002). However, research has shown that many second language learners have significant trouble achieving collocational fluency due to a number of persistent hurdles (DeCock, Granger, Leech, & McEnery, 1998; Kallkvist, 1998). These include the complexity of how collocations function (Hill, Lewis, & Lewis, 2000), and also the lack of emphasis by teachers and material developers (Gitsaki, 1996; Nesselhauf, 2005). Adding to the problem is the fact that there are still very few studies that identify which are the most frequent (Durrant & Schmitt, 2009), and there is a lack of agreement on what criteria should be used to achieve this.

Thus, many questions persist regarding how students, teachers, and researchers should approach collocations. Corpora can help us tackle the multifaceted and

Please cite this article as: Rogers, J., Brizzard, C., Daulton, F., Florescu, C., MacLean, I., Mimura, K., . . . Shimada, Y. (2015). On using Corpus frequency, dispersion, and chronological data to help identify useful collocations. *Vocabulary Learning and Instruction*, *4* (2), 21–37. doi: 10.7820/vli.v04.2.rogers.et.al

complex issues that must be resolved to help students develop their collocational fluency. However, many questions still remain in regard to how to use corpus data to accomplish this. For instance, the ideal frequency cut-off in corpus data for identifying high-frequency collocational co-occurrence is still unknown, in that previous research has examined as low as two occurrences per million tokens (Liu, 2003) and as high as 40 occurrences per million tokens (Biber, Conrad, & Cortes, 2004). Furthermore, studies often disregard a collocation's distribution among a wide variety of genres, or *dispersion*. For instance, although Liu's (2003) study was quite comprehensive in regard to its frequency cut-off, it did not consider dispersion as Biber et al. (2004) did. In addition, considering whether a collocation occurs regularly over a number of years instead of sudden surges or a decline in frequency (*chronological data analysis*) has likewise been neglected. No previous collocation identification studies have used this criterion to our knowledge.

Thus, this paper examined the extent to which frequency, dispersion, and chronological data from corpora helped identify useful collocations to directly teach. Such items will have a good cost/benefit value for these learners; they will be worthwhile to study because they occur frequently in a wide variety of texts and will be chronologically stable.

2 Research Topic Background

2.1 What is a Collocation?

This paper defines collocations in the traditional sense by frequency of co-occurrence (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Shin, 2006), while at the same time counting co-occurrence from a more modern perspective as lemmatized *concgrams*, as per the methodology set forth in Rogers, et al.'s (2014) study. Cheng, Greaves, and Warren (2006) define concgrams as "all the permutations of constituency and positional variation generated by the association of two or more words" (p. 411). *Constituency variation* (AB, ACB) involves a pair of words not only co-occurring adjacent to one another (*lose weight*) but also with a constituent (*lose some weight*). *Positional variation* (AB, BA) refers to counting total occurrences of two or more particular lexical items that include occurrences on either side of each other. Thus, *provide you support* and *support you provide* would both be included in the total counts for a multi-word units (MWU) concordance search for the lemma *provide* and *support*. This study also does not limit itself to including only collocations with high predictability such as *crux/matter*. This is due to the unreliability of statistical measures of association, discussed further below. In addition, this paper includes semantically transparent lemma pairs as collocations. Grant and Bauer (2004) refer to such word combinations as *literals*, or MWUs which are both compositional (the meaning of the whole can be deduced from its parts) and non-figurative. For example, with the MWU *eat breakfast* you literally *eat breakfast.* On the other end of the spectrum would be core idioms, non-compositional but also non-figurative MWUs, such as *kick the bucket* (you literally do not *kick* anything, there is no actual *bucket*, and it is impossible to deduce the meaning from the MWU's parts). There are clear rationales for considering such items, which are discussed further in this paper as well.

2.2 The Lack of Resources

Knowledge of collocations and formulaic language that have frequent co-occurrence is of obvious value to the language learner. Such knowledge has been referred to as a decisive factor in developing fluency (Almela & Sanchez, 2007). However, despite growing recognition of collocational fluency, few resources are available to guide collocation selection. Resources do exist, but they often only have hundreds of items and thus are far from being considered as comprehensive resources for helping learning to master the collocations of high-frequency vocabulary. In addition, very large dictionary-like resources with tens of thousands of items also exist, but clearly such massive contents are not practical in regard to direct instruction. One exception would be Shin (2006), but this study only aimed at created a collocation inventory for beginners and only involved L1-L2 congruency analysis with Korean.

In addition, many of the studies that have been conducted have been referred to as flawed in some aspect or lacking comprehensiveness (Durrant & Schmitt, 2009). Collocation can be viewed as intertwined with formulaic language, depending on one's definition of collocation. However, many formulaic language studies limit their scope to a specific type of multi-word unit. For instance, Biber et al. (2004) only found 172 "lexical bundles", limiting themselves to a cut-off of 40 occurrences per million and only considering four-word sequences. Such findings pale in comparison to the estimated collocations of native speakers, which have estimated to be in the hundreds of thousands (Hill, 2000).

2.3 Criteria for Identifying Useful Collocations

There are various ways to identify useful collocations. The simplest and most common involves frequency data from a corpus (Biber et al., 1999; Shin, 2006). While setting a frequency cut-off is unavoidably "arbitrary" (Nation, 2001a, p. 180), for teaching a cut-off must be set in regard to the practical limitation of how many items can be directly taught during limited classroom time. Nation (2001a, p. 96) suggested 2,000 word families as "practical and feasible" in regard to direct teaching, while Nation (2001b) suggested a limit of 3,000 word families.

Other researchers use statistical measures of association, such as how Lorenz (1999) utilized t-scores and mutual information data (MI) to identify high exclusive co-occurrence. MI measures the strength of co-occurrence between collocates. In other words, it measures "if the relative proportion of mutual occurrences of some words is large compared with their total frequencies" (Shin, 2006, p. 31). For instance, *tit*/*tat* has a very high MI of 15.01 (Davies, 2008). However, t-scores and MI can be problematic. MI emphasizes collocations whose components are not often found apart (Stubbs, 1995); thus, word pairs that clearly collocate but also have high frequencies might be excluded. Conversely, collocations with high t-scores will tend to be high-frequency words, and the measure may fail to identify collocations that have high frequencies of co-occurrence but low frequencies as individual words. Durrant and Schmitt (2009) give the following examples to highlight the issues with these statistical measures of association: "pairs like *good*

example, long way, and *hard work* attain high t-scores but low MI scores, while pairs like *tectonic plates* attain the reverse'' (p. 167).

Previous research on identifying useful collocations has led to various other criteria and sub-categories. For example, some researchers subdivided collocations into literals, figuratives, and core idioms (Grant & Bauer, 2004). They explain that if each word in an MWU is replaced with its definition, and the meaning of the word does not change, then it is a ''literal''. If it is possible to understand the meaning of an MWU by recognizing an untruth and pragmatically reinterpreting it in a way that correctly explains the MWU, then it is a ''figurative''. If only one word in an MWU is either literal, then such an MWU would be a ''ONCE''. Finally, they explain that if the MWU did not fall into any of these above categories, then it should be considered a ''core idiom''. However, while these semantic sub-categories of collocations do exist, researchers such as Wray (2000) insist that we deal with semantically transparent items in addition to those that are opaque. Nesselhauf (2005) agrees, finding that students tend to assign literal meaning to collocations with a figurative meaning, and vice versa.

L1-L2 collocation congruency (i.e., how similar/dissimilar a collocation's translation is in the learner's native language) is another criteria considered by many researchers. Feyez-Hussein (1990) found that approximately 50% of collocation errors were due to L1 influence. Thus, whether or not a collocation is semantically transparent or is a free combination becomes moot when the collocation differs greatly in comparison to how it is said in the learner's L1.

Notably, Nation (2001a) states that a collocation's balanced dispersion in many different categories of text is a necessary criterion for identifying useful collocations. Such collocations can easily be identified when corpora provide *dispersion* data, or the distribution of frequency among genres within the corpus. Gries (2008) believes that dispersion data analysis is essential as well, stating that raw frequency data can be misleading in regard to a word's general importance when the dispersion of its frequency data is unbalanced. However, only a few small-scale studies on identifying useful collocations have utilized dispersion data from corpora to delimit their selections of useful collocations. One such study is Cortes (2002). However, its corpus consisted of only approximately 360,000 tokens. Biber et al. (2004) also employed dispersion criteria, but their corpus consisted of two million tokens. Furthermore, not only has dispersion not been adequately applied to identify useful collocations, neither has chronological stability. Furthermore, any cut-off set for dispersion or chronological data will also be unavoidably arbitrary. For instance, Nation and Hwang (1995) specifically state that their choice of vocabulary occurring in 10 out of 15 sections of the corpora in their study for balanced dispersion was arbitrary.

In regard to frequency, Cortes (2002) set a frequency cut-off of 20 occurrences per million, and Biber et al. (2004) set theirs at 40 occurrences per million. Other studies were more inclusive. Shin (2006) set a cut-off of three occurrences per million, and Liu (2003) at two occurrences per million. However, the massive amount of data to be examined remained an issue. For example, despite examining significantly more items than previous studies (this study examined one occurrence per million tokens), items occurring approximately half as often could still be

considered to have value to language learners. For example, the lemma pairs *nice*/*vacation*, *finish*/*workout*, and *tend*/*exaggerate*, all occur approximately only once per two million tokens (Davies, 2008). Practically speaking, a native speaker would not consider these as low-frequency language not worthy of direct learning despite their low frequencies of co-occurrence. The frequency cut-offs used in the above previous studies could thus be considered conservative.

Thus, a significant gap exists in the research as to the extent that frequency, dispersion, and chronological data from corpora can help identify the most useful collocations. This brings us to this study's research questions.

3 Research Questions

1. To what extent can utilizing corpus frequency data help identify useful collocations?
2. To what extent can utilizing corpus dispersion data help identify useful collocations?
3. To what extent can utilizing corpus chronological data help identify useful collocations?

4 Methodology

4.1 Materials

This study utilized data from the *Corpus of Contemporary American English* (*COCA*) (Davies, 2008). The COCA provides collocation lists that have been compiled with consideration for constituency and positional variation, and this was one of the reasons why it was chosen as a data source. This study thus utilized Davies' (2010) *Word List Plus Collocates*, a lemmatized concgram list. It consists of the most frequent 739,255 collocates that co-occur with the most frequent 5,000 lemmas in the corpus.

In addition, the COCA divides itself into five separate genres: spoken, fiction, magazine, newspaper, and academic, and these sections all have nearly as many tokens in total and per year. The division of data into these sections made the dispersion analysis in this study possible. Also, the COCA also divides itself into chronological sections of four years per section. This study utilized the completed chronological sections from 1990 to 2009.

4.2 Procedure

This study began by piloting various frequency cut-offs on Davies' (2010) collocation list. The aim of this study was to find a frequency cut-off which resulted in the vast majority of collocates identified being judged as useful and worthy of direct instruction, and also consisting of between 2,000 and 3,000 word families.

Frequency cut-offs were piloted to determine how many useful collocations were at each level. The study took a cue from previous research and started at Biber et al. (2004) cut-off of 40 occurrences per million tokens, and continued to

Kjellmer's (1987) two occurrences per million. Cobb's (2015) *Vocabprofile* programme, which consists of the most frequent 25,000 word families in the *BNC* and *COCA*, was utilized to determine the total word families the collocations consisted of to avoid exceeding 3,000 word families while not falling below 2,000 word families. After identifying a cut-off that resulted in between 2,000 and 3,000 word families, the list was then examined by a native speaker for general item usefulness to ensure that the list was not overly inclusive. For example, if a frequency cut was too inclusive, it could include very low-frequency collocates of high-frequency vocabulary that would be of little value to learners (see Table 1).

This study found that utilizing a frequency cut-off of one occurrence per million tokens was ideal. How this cut-off was decided upon is explained further in Section 4.1 of this paper.

Often, the collocation that occurred was a node word itself within the most frequent 5,000 lemma of Davies' (2010). Therefore, the list also includes many duplicate entries such as *take*/*walk* and also *walk*/*take*. Such duplicates were first removed.

Then, dispersion and chronological data for identified collocates were collected from the COCA. Its interface allows users to extract dispersion data for five genres: spoken, fiction, magazine, newspaper, and academic. The interface also allows for the extraction of chronological data in four-year increments: 1990–1994, 1995–1999, 2000–2004, 2005–2009, and 2010–2012. Since the four-year section 2010–2013 was yet to be completed, its data were not included in this study.

Various parameters were then piloted to determine the cut-off point for balanced dispersion and chronological data distribution; these ratios were trialed, and the items flagged at each ratio were examined using native speaker intuition to judge whether it was overly inclusive or exclusive.

As for dispersion and chronological data, a range of parameters were tested due to the gap in research with the corpus used in this study. As with frequency cut-offs, any cut-off set for dispersion or chronological data will also be unavoidably arbitrary. This study experimented with parameters that best approximate balanced distribution.

For dispersion data, the parameters required that a specific percentage of the total occurrences had to occur in a majority of the *COCA*'s genres: three or more out of the five genres. Native speaker intuition was used to determine the best percentage cut-off. The lemma list was examined for items specialized in nature, and a number of these items were found to have approximately 5% or less of their occurrences in three or more of the genres. Thus, dispersion data were analyzed at three separate percentages to determine the most useful parameter: less than 10%, 5%, and 2.5% of total occurrences in three or more genres. Then pairs flagged at these parameters were examined to determine if they truly were specialized by a native speaker, and thus not worthy of direct instruction for a general English course. Next, all remaining items in the list were also scanned by a native speaker to determine if the parameters were not able to identify items that were actually specialized. Finally, all items identified as being unbalanced by a native speaker were examined to determine if they fell into a common genre (e.g., academic language).

Table 1. Examples of High- and Low-Frequency Collocates of the Lemma *Play* in the COCA (Davies, 2008)

Rank	Collocate	Frequency
1	*role*	20,747
2	*game*	8,536
99	*gin*	82
100	*hoops*	80

Table 2. System for Rating the Value of Collocates for Learners of General English

Rating value in regard to direct teaching
1. Provides no value whatsoever if directly taught. None of the examined items fell into this category so an example cannot be provided. However, this rating was included to possibly deal with any items in the list that represented corpus ''noise''. In other words, these would include mistakes in the data or in how the data was compiled, which would result in the inclusion of items that are clearly not part of natural language but are the result of the fact that the compilation of a corpus is a mere attempt at emulating balanced natural language use.
2. Provides little value if directly taught. For example, *note/supra* was found to have extremely unbalanced dispersion data, occurring mostly in the academic section in the COCA. Because of its specialized usage in an exclusive genre, native speakers may not even be aware of its meaning. Thus, clearly such an item will be of little value to a learner of general English.
3. Provides questionable value if directly taught. For example, *lemon/zest*. This item occurs often but with unbalanced distribution data in the COCA because of inclusion of many magazines with recipes in the COCA's ''magazine'' section, and is of clear questionable value for learners trying to master general English.
4. Provides value, but with limitations if directly taught. For example, *championship/year*, despite having unbalanced dispersion data distribution, would be of value to teach because of the generalness of the term and ubiquitous nature of sports in society. The term is general in that it can be used to describe all sports despite the fact that sports itself is somewhat specialized.
5. Provides clear value if directly taught. For example, how *email/address*, despite having unbalanced chronological data distribution, is clearly a valuable and stable item to learn.

A similar methodology was employed for chronological data analysis. Again, native speaker intuition was used to determine the best percentage cut-off. First, the lemma list was examined using native speaker intuition for pairs which were either dated, too modern or only occurred during a specific time period. Very few such items existed, but the items that were found had approximately 5% or less occurrences in one or more of the four chronological sections. Just as dispersion data were analyzed, chronological data were also analyzed to find items having less than 10%, 5%, and 2.5% of total occurrences in one or more sections. Then pairs flagged at these parameters were examined to determine if they truly were dated, too modern, or not useful because they only occurred during a specific time period by a native speaker, and thus not worthy of direct instruction for a general English

course. Next, all remaining items in the list were also examined by a native speaker to determine if the parameters were unable to identify items that were dated, too modern, or had little value because they only occurred during a specific time period. Finally, all items identified as being unbalanced by a native speaker were examined to determine whether they were either dated, only occurred during a specific time period, or were too modern. How these parameters were decided upon is explained further in Sections 4.2 and 4.3 of this paper.

Last, to determine the extent to which the dispersion and chronological data distribution cut-offs truly identified items that were not worthy of direct instruction, the collocates were then judged by a native speaker in regard to their usefulness. Each item was given a rating (see Table 2) in regard to its value for learners of general English.

After being rated, any items flagged by each of the cut-off parameters that were rated 1 or 2 were tallied. Furthermore, any items not flagged by the cut-off parameters that received ratings of 1 or 2 were also tallied. These two steps would then be used to judge the cut-off parameter's ability to identify collocations that truly are of little or no use for general learners of English in regard to balanced dispersion and chronological data.

5 Results

5.1 Frequency Cut-off Results

After trialing the various frequency cut-offs used by previous researchers, it was found that the cut-off of two occurrences per million tokens resulted in a list of lemma pairs consisting of only 1,671 families. Taking a cue from the recommendation of directly teaching between 2,000 (Nation, 2001a) and 3,000 (Nation, 2001b) word families, it was therefore determined that a more inclusive cut-off could be used. A cut-off of once per million tokens and once per 500,000 tokens was then piloted, which resulted in 2,540 families and 4,122 families, respectively. The cut-off of one occurrence per million tokens was thus determined to be ideal.

Cobb's (2015) *Vocabprofile* programme showed that these pairs covered 75.6% of the top 3,000 word families. It should also be noted that 97.8% of the tokens in the lemma pair list occur within the top 3,000 word families. An analysis of the data is presented in Table 3.

After duplicate entries and proper nouns were removed, the cut-off resulted in 14,035 pairs being included. Due to the large number of items, this list was checked by an experienced, native-speaking teacher of English for usefulness, and the vast majority were found useful and worthy of direct teaching. Therefore, it was confirmed that the frequency cut-off was not too inclusive.

5.2 Dispersion Data Analysis Results

Out of all three parameters tested, the 5% or more cut-off in three or more genres was shown to be the most reliable in regard to both properly flagging items of little use for learners of general English, and not flagging items the native

Table 3. Word Frequency Breakdown of Lemma Pairs Occurring Once Per Million Tokens According to *Vocabprofile's* 25,000 Word Families of the BNC and COCA

Frequency level	Families (%)	Types (%)	Tokens (%)	Cumul. token%
K-1 Words:	806 (32.59)	1,095 (38.17)	17,461 (69.15)	69.15
K-2 Words:	704 (28.47)	847 (29.52)	4,945 (19.58)	88.73
K-3 Words:	595 (24.06)	660 (23.00)	2,280 (9.03)	97.76
K-4 Words:	207 (8.37)	211 (7.35)	302 (1.20)	98.96
K-5 Words:	91 (3.68)	91 (3.17)	104 (0.41)	99.37
K-6 Words:	38 (1.54)	40 (1.39)	46 (0.18)	99.55
K-7 Words:	13 (0.53)	13 (0.45)	13 (0.05)	99.60
K-8 Words:	9 (0.36)	9 (0.31)	10 (0.04)	99.64
K-9 Words:	4 (0.16)	4 (0.14)	4 (0.02)	99.66
K-10 Words:				
K-11 Words:	2 (0.08)	2 (0.07)	2 (0.01)	99.67
K-12 Words:	2 (0.08)	2 (0.07)	2 (0.01)	99.68
K-13 Words:	1 (0.04)	1 (0.03)	1 (0.00)	
K-14 Words:	1 (0.04)	1 (0.03)	1 (0.00)	
K-15 Words:				
K-16 Words:				
K-17 Words:				
K-18 Words:				
K-19 Words:				
K-20 Words:				
K-21 Words:				
K-22 Words:				
K-23 Words:				
K-24 Words:				
K-25 Words:				
Off-List:		44 (1.53)	80 (0.32)	100.00
Total (unrounded)	2,473	2,869 (100)	25,251 (100)	100.00

speaker judged to be useful (see Table 4). At 5%, 845 items were considered to be either erroneously flagged or left unflagged by the parameters after native speaker analysis. The next most reliable parameter was at 2.5%, where a total of 1,283 items were considered either erroneously flagged or unflagged. At this parameter, the vast majority of the 1,283 items fell beyond the parameter (1,211) and thus it was not inclusive enough. The most unreliable parameter was 10%, where a total of 1,487

Table 4. Dispersion Data Analysis Results

Parameter	Accurately flagged	Items judged unbalanced by a native which were not flagged	Erroneously flagged	Total items parameters either did not flag or erroneously flagged
2.5%	616	1,211	72	1,283
5%	1,171	656	189	845
10%	1,618	209	1,278	1,487

Table 5. Most Common Types of Language from which Flagged Items are Derived

Parameter	Academic	Fiction	Food	Television
2.5%	238	26	249	56
5%	215	106	143	21
10%	300	17	54	4

items were either considered erroneously flagged or the parameters did not flag. Conversely, in comparison with the 2.5% parameter, 10% proved to be too inclusive in that the vast majority of the 1,487 items were erroneously flagged (1,278). In total, native speaker judgment identified 1,827 of the 14,035 pairs (13%) as having limited value for learners of general English.

When items were judged by a native speaker to determine their type of specialized language, four specific types accounted for the vast majority of items: academic language, descriptive language primarily used in fiction, language related to food, and language used primarily on television. Table 5 shows the number of items in each of these four types at all three parameters (non-combined).

A total of 1,539 flagged pairs at all three parameters were judged erroneously flagged by a native speaker. That is, the native speaker felt these items did have value for learners of general English. In addition, there were 209 pairs judged by a native speaker to be specialized and of little use to general learners that were not flagged at any of the three parameters.

5.3 Chronological Data Analysis Results

Out of all three parameters tested, the 2.5% and 5% or more cut-off in one or more chronological sections were shown to be the most reliable in regard to both flagging items of little use for learners of general English because of chronological issues, and not flagging items the native speaker judged to be useful for learners of general English (see Table 6). At both 2.5% and 5%, 100 items were either erroneously flagged or the parameters did not flag. At 10%, 145 items were either erroneously flagged or left unflagged. Only five items beyond the parameters tested were judged by a native speaker to be of little use for leaners because of chronological issues.

Despite the 2.5% and 5% parameters being the most reliable, the vast majority of the items flagged by all three parameters were found erroneously flagged by a

Table 6. Chronological Data Analysis Results

Parameter	Accurately flagged	Items judged unbalanced by a native which were not flagged	Erroneously flagged	Total items parameters either did not flag or erroneously flagged
2.5%	14	40	59	99
5%	13	22	77	99
10%	23	4	140	144

native speaker. These items were either useful to learners or did not actually exhibit chronological issues. Furthermore, the entire analysis only resulted in a total of 55 items being flagged for chronological issues, which amounts to only 0.39% of the total items examined.

6 Discussion

6.1 Frequency Data Analysis

Determining the extent to which frequency data can help inform useful collocation selection revealed both potential and limitations. First, it was shown that it is possible to set a frequency cut-off that results in a list of collocations that can be practically taught. What at first seemed an impractical amount of items to teach was in reality only 2,473 word families combining with each other in 14,035 different ways, which is within the 2,000–3,000 word family estimate of what can be taught directly. And while many useful collocations do occur beyond the frequency cut-off of this study, a list of collocations resulted that showed very good coverage of high-frequency vocabulary (75.6%) in addition to having 97.8% of the word families within the pairs being within the most frequent 3,000 word families. However, a number of other steps must still be taken to make the data practically usable, despite these positive results.

Shin and Nation (2008) refer to collocations as having two parts: a pivot word and its collocate. In this study, pivot words were the top 5,000 most frequent lemma in the COCA and the collocates, the words that co-occur with these frequently. However, there was the issue of removing duplicates, or instances when a collocate of one pivot word is also a pivot word itself.

This is a time-consuming, manual process that is essential. Moreover, proper nouns also need to be removed. This step is also time-consuming because it must be done manually. It was also difficult to judge whether a lemmatized collocational pair is part of a larger proper noun without examining concordance data.

Thus, the answer to Research Question 1 is that frequency data can, to a large extent, help identify useful collocations. The limitation is that many of the items identified may have duplicate entries and proper nouns would also need to be removed. Finally, such a list may contain a significant amount of items that are of little value to learners of general English due to their specialized nature.

6.2 Dispersion Data Analysis

Considering a collocational pair's general value in regard to its usefulness across multiple genres proved to be an important criterion; it identified 13% of the 14,035 pairs as not being of significant value to general learners of English. However, dispersion data alone was not sufficient in identifying unbalanced items. Often the parameter set was either too inclusive or not inclusive enough, and thus items would be included that were of little value or items of little value were not identified for removal. The most reliable parameter was shown to be a cut-off of 5%

of occurrences across three or more genres. While the parameter was useful in helping to flag items to reconsider, native speaker judgments were unavoidable. The parameter could only flag 64.1% of the items that were truly of little value, while 10.3% of the items flagged were later judged to be valuable.

The largest group that had unbalanced dispersion data was pairs occurring mostly in the academic section. While these pairs would be highly useful for students who plan to do scientific research or read academic journals, such items may not be useful for more general language needs. Thus, identifying such genre-specific, unbalanced items can be extremely valuable, either to exclude them or even focus on them if appropriate.

The same can also be said for the large number of pairs that occurred mostly in the fiction section. They consisted of language employed by fiction writers to describe what the reader cannot see. Thus, these items do not occur often in any other genres. Again, their inclusion or exclusion depends on the course of study.

Biber, Conrad, and Reppen (1998) reminded us that large corpora can skew the type of data we are looking for. This was evident in the disproportionate amount of collocations related to cooking found in the magazine and newspaper sections. Since the magazines and newspapers sourced by the COCA regularly featured recipe articles, such items had disproportionate frequency totals. The pedagogical value of directly teaching such items to general learners is questionable except for those who plan to work in the food industry. Thus despite their high frequency, their pedagogical value is in doubt.

Items mostly occurring in the spoken section were also apparently influenced by the data source. The COCA sourced much of its spoken section data from television, and in particular, news or talk shows. Thus, the vast majority of the items with unbalanced dispersion in the spoken section consisted of the language newscasters or talk show hosts use, such as commercial break transitions, etc. The value of such items for learners of general English is also arguably low for second language learners, and their discovery shows the importance of dispersion data.

Also of note is how the COCA divides its genres, and the effects that it has on dispersion data. While much academic and fiction-related language was easily identified, the same cannot be said for other specialized genres, such as business-related collocations, despite it being a clearly specialized genre. Business-related terms were distributed throughout the spoken, magazine, and newspaper genres of the *COCA*, but not in particularly high-frequency counts in comparison with academic language, which had its own dedicated genre. Only a small portion of the spoken, magazine, and newspaper genres took its data from business-related sources, such as financial magazines. If the COCA were designed with this in mind, such language could have also been easily identified. Such data would be of clear value to the many learners of business English.

In summary, the data analysis showed that the most reliable parameter was able to identify 64% of the items deemed to be of little value for learners of general English by a native speaker. Thus, in regard to answering Research Question 2, the extent to which dispersion data can identify useful collocations is limited in that the parameter was only able to identify 64% of the items that needed to be excluded.

As with frequency data, a native speaker manual analysis of the items in regard to dispersion was deemed essential as well.

6.3 Chronological Data Analysis

Considering a collocational pair's balanced chronological data distribution, when determining its value for learners, proved to be much less effective than the dispersion data analysis, since only 0.39% of the 14,035 pairs were found to be either dated, too modern, or only occurred in a limited time span in the past. Furthermore, each parameter was shown to be quite unreliable in that the vast majority of the items it flagged as having unbalanced distribution was deemed valuable for learners of general English.

Often items erroneously flagged by the parameters were new collocations deemed by a native speaker to have high potential to be used regularly in the future, such as *internet/access*. The types of items that were accurately flagged or deemed by a native speaker to have chronological issues were mostly related to temporal events, such as with *new/millennium*. Items with sudden surges in frequency counts were mostly connected to political events, wars, or other time-sensitive events.

Some items were also deemed too modern, so their future value was unclear. For instance, *cell/embryonic* was flagged by one of the parameters and considered by a native speaker to be of questionable value. It may have high-frequency counts simply because it is a new technology and being discussed often, and it is unclear how whether the collocation will continue to be used. The science may become commonplace or outdated, and thus the term may not be discussed as often in the future.

Only a few items were considered as dated, such as *word/processor*. Notably, the corpus only provides data back to 1990. If older data were available, then there would be more dated collocations identified. However, within the data's 19-year span, very few dated collocations were found. In addition, if a more detailed chronological breakdown of data were available (i.e., a breakdown by year instead of four-year sections), a more in-depth analysis would have been possible.

As for Research Question 3, the data clearly demonstrated the limited efficacy of chronological data analysis. Not only was there a very small number of items that actually had chronological issues, all of the parameters tested were highly unreliable, thus again requiring native speaker judgment. Thus, this criterion was shown to be of limited value for useful collocation identification.

7 Conclusion

7.1 Summary of Results

This paper has described the extent to which frequency, dispersion, and chronological data can help identify useful collocations. The frequency cut-offs that were tested identified a particular cut-off that resulted in a practical number of useful items to be taught. It resulted in a list of collocations with very high coverage of high-frequency vocabulary. However, many useful collocations were also shown

to remain beyond the frequency cut-off. It furthermore showed that some highly time-consuming steps were still required to make the results usable, such as removing duplicate entries and proper nouns.

Frequency data analysis alone proved insufficient in producing a list of collocations that all have value to learners of general English. The dispersion data analysis conducted in this study identified 13% of the items as not being of value for learners. Although one particular dispersion data cut-off was more reliable in identifying items deemed by a native speaker to be unbalanced in their usage across genres of English, this parameter could not identify all of the items a native speaker deemed as being of little value due to unbalanced dispersion. This parameter failed to identify 35.9% of the total items deemed to be of little value due to unbalanced dispersion. Furthermore, 13.9% of the items this parameter did identify as having unbalanced dispersion data were actually judged to be of value, and thus had been erroneously flagged. So despite being useful in flagging many of the items that truly had little value for learners, such data cannot be considered reliable. Native speaker analysis, a time-consuming manual process, was thus shown to be necessary.

This study also revealed that dispersion data analysis can not only identify useful items for learners of general English, but can also identify specialized vocabulary, which is prominent in academic language and fiction writing, etc. Identifying such specialized vocabulary is not only useful for teachers who need to exclude it from a more general language course, but it is conversely useful for specialized classes needing to focus on highly salient collocations. The analysis also revealed that even data in large corpora can exhibit skewed frequencies. Any corpus is only as good as its source, and dispersion data analysis can identify deficiencies in corpora, such as the COCA's heavy inclusion of food/recipe-related language.

This study's chronological data dispersion analysis was shown to be of far lesser value in comparison with the genre dispersion data analysis. In total, only 0.39% of the items examined were found to have chronological issues. The parameters tested were also shown to be quite unreliable, in that the vast majority of items they flagged as having unbalanced chronological data distribution were either deemed of value to learners or to not have chronological issues in the first place. Therefore, it is less clear whether examining collocations for chronological balance is warranted or productive.

As mentioned earlier, the number of items to be examined when determining useful collocations is staggering. The goal of this study was to identify collocations that could be practically taught. So while this study examined significantly more items in comparison with previous research, the resulting list cannot be considered completely comprehensive. Furthermore, there are also limitations in how the results of this study can be interpreted due to the fact that only one native speaker gave judgments in regard to which items seemed worthy of direct instruction because of practical time limitations for judging such a large amount of data. If more than one speaker examined the data, an inter-rater reliability analysis could have been conducted to provide more solid data.

This study also acknowledges the limitations of its parameters in regard to frequency, dispersion, and chronological data analysis; no relevant precedents for the corpus used in this study had existed, and thus to an extent its parameter

cut-offs were subjective. The present study simply aimed to validate the usefulness of some specific parameters to help identify useful collocations. It acknowledges that results will never be indisputable, but rather offers the best approximation possible within unavoidable constraints. To our knowledge, this was the first study using dispersion and chronological data from the COCA to determine useful collocations; thus, parameters were experimented with that best approximate balanced distribution. Regardless, there are clearly limitations to interpreting the results of this study due to these issues.

Similarly, native speaker intuition judgments on the value of items for learners of general English are subjective. However, the data revealed that such judgments were essential. In addition, this paper acknowledges the limitation of having only one native speaker make judgments on the value of items in regard to the frequency cut-off and parameters. While employing native speakers is ideal, due to time constraints and the large amount of items examined, relying upon a single native speaker was an acceptable expedient.

Despite the above limitations, we believe this paper contributes to collocation research and can inform future works. These limitations should be considered as opportunities for future researchers to improve methodology and resource design. These improvements will hopefully lead to further insights in regard to the identification of useful collocations.

References

Almela, M., & Sanchez, A. (2007). Words as "lexical units" in learning/teaching vocabulary. *International Journal of English Studies*, *7*(2), 21–40. Retrieved from http://revistas.um.es/ijes/issue/view/4811.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*, 371–405. doi:10.1093/applin/25.3.371

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London, UK: Pearson Education.

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, *11*, 411–433. doi:10.1075/ijcl.11.4.04che

Cobb, T. (2015). *Vocabprofile*. Retrieved from http://www.lextutor.ca/vp/bnc/

Cortes, V. (2002). Lexical bundles in freshman composition. In R. Reppen, & S. M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131–145). Amsterdam, the Netherlands: John Benjamins Publishing Company.

Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990–present. Retrieved from http://corpus.byu.edu/coca/

Davies, M. (2010). *Word list plus collocates*. Retrieved from http://www.wordfrequency.info/purchase1.asp?i=c5a

DeCock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67–79). London, UK: Longman.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, *47*, 157–177. doi:10.1515/iral.2009.007

Feyez-Hussein, R. (1990). Collocations: The missing link in vocabulary acquisition amongst EFL learners. In J. Fisiak (Ed.), *Papers and studies in contrastive linguistics: The Polish English contrastive project* (Vol. 26, pp. 123–136). Poznan, Poland: Adam Mickiewicz University.

Gitsaki, C. (1996). *The development of ESL collocation knowledge* (Unpublished doctoral dissertation). University of Queensland, Queensland, Australia.

Grant, L., & Bauer, L. (2004). Criteria for re-defining idioms. Are we barking up the wrong tree? *Applied Linguistics*, *25*(1), 38–61. doi:10.1093/applin/25.1.38

Gries, S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*, 403–437. doi:10.1075/ijcl.13.4.02gri

Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 47–67). Hove, UK: Language Teaching.

Hill, J., Lewis, M., & Lewis, M. (2000). Classroom strategies, activities and exercises. In M. Lewis (Ed.), *Teaching Collocation: Further developments in the lexical approach* (pp. 88–117). Hove, UK: Language Teaching.

Kallkvist, M. (1998). Lexical infelicity in English: The case of nouns and verbs. In K. Haastrup, & A. Viberg (Eds.), *Perspectives on lexical acquisition in a second language* (pp. 149–174). Lund, UK: Lund University Press.

Kjellmer, G. (1987). Aspects of English collocations. In W. Meijs (Ed.), *Corpus linguistics and beyond* (pp. 133–140). Amsterdam, the Netherlands: Rodopi.

Liu, D. (2003). The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly*, *37*, 671–700. doi:10.2307/3588217

Lorenz, G. (1999). *Adjective intensification – Learners versus native speakers: A corpus study of argumentative writing*. Amsterdam, the Netherlands: Rodopi.

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

Nation, I.S.P. (2001b). How many high frequency words are there in English? In M. Gill, A. W. Johnson, L. M. Koski, R. D. Sell, & B. Warvik (Eds.), *Language, learning, literature: Studies presented to Hakan Ringbom* (pp. 167–181). Turku: Abo Akademi University.

Nation, I.S.P., & Hwang, K. (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

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam, the Netherlands: John Benjamins.

Rogers, J., Brizzard, C., Daulton, F., Florescu, C., MacLean, I., Mimura, K., ... Shimada, Y. (2014). A methodology for identification of the formulaic language most representative of high-frequency collocations. *Vocabulary Learning and Instruction*, *3*(1), 51–65. doi:10.7820/vli.v03.1.2187-2759

Shin, D. (2006). *A collocation inventory for beginners* (Unpublished doctoral dissertation). Victoria University of Wellington, Wellington, New Zealand.

Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*, *62*, 339–348. doi:10.1093/elt/ccm091

Snellings, P., van Gelderen, A., & de Glopper, K. (2002). Lexical retrieval: An aspect of fluent second–language production that can be enhanced. *Language Learning*, *52*, 723–754. doi:10.1111/1467-9922.00202

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative methods. *Functions of Language*, *2*(1), 23–55. doi:10.1075/fol.2.1.03stu

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, *21*, 463–489. doi:10.1093/applin/21.4.463

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

Vocabulary Learning and Instruction
Volume 4, Issue 2, December 2015
http://vli-journal.org

Article

Replacing Translation Tests With Yes/No Tests

Raymond Stubbe
Kyushu Sangyo University
doi: http://dx.doi.org/10.7820/vli.v04.2.stubbe

Abstract

Along with personal interviews, individual word translation tests from the target language to the mother tongue are recognized as a reliable method of determining students' actual lexical knowledge. However, as most English as a foreign language teachers are aware, the marking of these tests can be a laborious task. A far easier vocabulary testing format is the Yes/No (YN) checklist test, which can examine a large number of words while not over-burdening the marker. Pseudowords, which look like real words but do not bear meaning, have been added to the YN format to check for evidence of overestimation of lexical knowledge by test-takers. Four scoring formulae, which adjust YN results according to the number of pseudoword reports, have become established in the literature. Of these, the *h-f* formula has become recognized as the simplest to use for adjusting YN scores. This study presents a regression-based prediction formula derived from the *h-f* results in a pilot study, which was then applied to the YN *h-f* adjustments in a second study (the main study) to predict actual vocabulary knowledge as demonstrated by a meaning recall translation test of the same items. This prediction formula, labeled *h-fRF*, was compared with another regression-based formula as well as the original *h-f* formula. Results showed that 54% of the 455 individual *h-fRF* predictions were within 5% (4.8 of 96 words) of matching translation test scores, and 82% were within 10%, which were better than the other formula predictions. These results may be of interest to classroom teachers as they suggest that by using the *h-fRF*, the burden of marking translation tests can be reduced by the far easier YN test format.

1 Background

The ability to recall the meaning of individual words while reading has long been recognized as a pre-requisite for successful reading comprehension (Beglar & Hunt, 1999; Qian, 1999, Stæhr, 2008 and others). For second/foreign language (L2) learners, the easiest way to demonstrate this recall ability is by translating the encountered words into their native language (L1). Consequently, when judging the lexical suitability of texts for use in the English as a foreign language (EFL) classroom, teachers often utilize L1–L2 passive recall translation tests. This study examined whether or not the easier to administer Yes/No (YN) test can be used to predict scores on a meaning recall translation test and possibly replace this latter, more cumbersome format.

Please cite this article as: Stubbe, R. (2015). Replacing translation tests with Yes/No tests. *Vocabulary Learning and Instruction*, *4* (2), 38–48. doi: 10.7820/vli.v04.2.stubbe

1.1 YN Vocabulary Tests

YN vocabulary tests present learners with a list of words, usually selected from word frequency lists, and ask them to signify their knowledge of each item by either checking that word or by selecting either "yes" or "no". Read (2007, pp. 112–113) notes: "Despite its simplicity, the Yes/No format has proved to be an informative and cost-effective means of assessing the state of learners' vocabulary knowledge, particularly for placement and diagnostic purposes." As YN tests rely solely on self-reporting, the actual lexical knowledge of the students cannot be verified. One concern with the YN format is whether test results accurately reflect the test takers' knowledge of the selected items, or if the results overestimate the number of words actually known (Read, 1993, 2000). To compensate for the potential of students claiming knowledge of words they actually do not know the meaning of (overestimation), pseudowords, or non-real words, were introduced to the vocabulary checklist test by Anderson and Freebody (1983). Pseudowords were introduced to the field of L2 acquisition by Meara and Buxton (1987).

The use of pseudowords in YN tests has remained widespread through present-day versions. In these tests, knowledge of a real word is known as a hit, while claiming knowledge of a pseudoword is a false alarm (FA). Not claiming knowledge of a real word is labeled a miss and not claiming knowledge of a pseudoword is a correct rejection. Claiming knowledge of words that do not exist is seen as an indication of falsely claiming knowledge of real words (overestimation).

There presently exists three approaches to utilizing YN test pseudoword data. One use is to set a maximum acceptable number of pseudowords beyond which "the data are discarded as unreliable" (Schmitt, 2010, p. 201). Schmitt, Jiang, and Grabe (2011) set their acceptance limit at three (10% of their 30 pseudowords). Barrow, Nakanishi, and Ishino (1999) set the same cut-off point for the 30 pseudowords used in that study. Stubbe (2012b) demonstrated that a cut-off point of four (12.5% of the 32 pseudowords) better suited those YN test results.

Another use of YN pseudowords is to adjust the YN scores using a correction for guessing formula. The test results from learners checking pseudowords are adjusted using a variety of formulae, to better reflect their actual vocabulary knowledge. Four such established formulae were compared in Huibregtse, Admiraal, and Meara (2002): *h-f, cfg, △m, and Isdt.* With the first formula, *h-f,* (Anderson & Freebody, 1983), the proportion of FAs relative to the total number of pseudowords, the FA rate (*f*), is subtracted from the proportion of hits relative to the total number of real-word items, the hit rate (*h*), to create the formula: true hit rate = *h-f.* The remaining correction for guessing formulas are slightly more complicated and are presented below:

cfg (correction for guessing: Meara & Buxton, 1987):

$$cfg = \frac{h\text{-}f}{1 - f}$$

$\triangle m$ (Meara, 1997):

$$\Delta m = \left(\frac{h\text{-}f}{1-f}\right) - \left(\frac{f}{h}\right)$$

Isdt (Huibregtse et al., 2002):

$$Isdt = 1 - \frac{(4 * h * (1-f)) - (2 * (h\text{-}f)) * (1 + h\text{-}f)}{(4 * h * (1-f)) - (h\text{-}f) * (1 + h\text{-}f)}$$

Huibregtse et al. (2002) found that their *Isdt* formula had the best prediction ability of the four correction formulae, but that the simpler *h-f* formula (Anderson & Freebody, 1983) worked just as well under most conditions. Mochida and Harrington (2006) and Stubbe (2012a) similarly report that *Isdt* had the highest correlation of the four correction formulae with a second multiple-choice test of the same items, while YN raw hits had the lowest correlation. Eyckmans (2004) however, comparing YN test results with a meaning recall (L2 to L1 translation) test reported that the *cfg* formula had higher correlations than *Isdt*. Eight years following Huibregtse et al.'s (2002) study, Schmitt (2010, p. 201) noted that "it is still unclear how well the various adjustment formulae work."

Pellicer-Sánchez and Schmitt (2012) compared the same four scoring formulae, using subsequent meaning recall student interviews as the criterion measure. They found that each formula provide a mean score that was higher than the matching interview score, with $\triangle m$ providing the closest mean score. In other words, all four of the established YN scoring formulae overestimated the testees' demonstrable lexical knowledge. A correlational analysis found that all formulae provided high correlations with the interview results for the non-native speakers ($r > .796$), with *h-f* proving superior. Despite the high correlations, each formulae provided adjusted YN scores that overestimated the testees' actual vocabulary knowledge.

In a more recent study, Stubbe and Hoke (2014; using pre-existing data from Stubbe, 2013) compared the same four scoring formulae evaluated in Huibregtse et al.'s (2002) study. It was also found that *h-f* had the highest correlations with translation scores (see Table 1). Results also suggested that residuals, which are the differences between the correction formula predicted score and the actual translation score for each participant, were lowest for *h-f.* Residuals were calculated using the root mean square error (RMSE) method described in De Veaux, Velleman, and Bock (2008). As Stubbe and Hoke (2014) demonstrated that *h-f* was superior to *cfg,* $\triangle m$ and *Isdt* in terms of predicting translation scores, only *h-f* was chosen for inclusion in this present study.

1.2 Improving YN Scores using Regression Analysis

A third usage for pseudowords (or false alarm data) was introduced by Stubbe and Stewart (2012): the creation of a standard least squares (multiple regression) model and formula which can be used to predict translation test scores using self-reports of lexical knowledge (real-word and pseudoword) on a YN test.

Table 1. Means, SDs, Range, Correlations, and Residuals of Applying the Four Correction Formulae ($n = 455$; from Stubbe & Hoke, 2014, p. 74)

Test/Formula	Mean	SD	*r*	residual
Tr Scores	27.05	12.16	1	–
YN hits	48.82	17.23	0.721	24.82
FA Counts	2.17	3.16	-0.142	–
h-f	42.29	16.53	0.833	17.84
cfg	45.46	17.50	0.807	21.19
△ m	33.11	26.57	0.739	20.31
Isdt	50.02	13.55	0.775	24.56

Note. Tr = translation test; FA = false alarms (pseudoword reports); SD = standard deviation; *r* = correlation (Pearson Product-Moment) with translation test scores; Residual was calculated by squaring the differences between each translation test score and each of the five predictions, summing those squares, calculating the mean (df = 453) and finally acquiring the square root.
Source. Stubbe and Hoke (2014) "Comparing Yes/No Test Correction Formula Predictions of Passive Recall Test Results" which first appeared in The 2013 Pan-SIG Conference Proceedings published by the Japan Association of Language Teaching (pp. 72–78).

The use of regression analysis with YN test results, though not that common, is not unprecedented (Mochida & Harrington, 2006).

Stubbe and Stewart (2012) presented two scoring formulae derived using multiple regression analysis, with YN test real-word scores and pseudoword scores as two independent (predictor) variables and translation test scores as the dependent variable. The first formula was based on the full 120 real-word and 32 pseudowords YN item list, and had an r^2 of 45.2%. This formula was reported as "True number of words known $= 8.14 + (0.41 \times \text{YN Score}) - (1.94 \times \text{FAs})$" (Stubbe & Stewart, 2012, p. 5), where 8.14 words represents the intercept on the *y*-axis. For every word reported known on the YN test, add 0.41 words truly known. For every FA, subtract 1.94 words. To illustrate, one student reported 78 words as known on the YN test, and checked two pseudowords, so her true score would be calculated as 35.34, $\{8.14 + (.41 \times 78) - (1.94 \times 2)\}$.

This original prediction formula was improved by utilizing item analysis to select 40 of the 120 real words on the YN test which had "the highest phi correlations to translation test results, and the 9 pseudowords with the highest negative point biserial correlations to overall translation test scores" (Stubbe & Stewart, 2012, p. 6). The resulting prediction formula was reported as "True number of words known $= 3.26 + (.51 \times \text{YN Score}) - (2.39 \times \text{FAs})$" and had an r^2 of 59.1% (Stubbe & Stewart, 2012, p. 6; hereinafter referred to as *S&SRF*, for Stubbe & Stewart Regression-based Formula). Using this prediction formula, the same student as above would receive a true score of 38.77, which is considerably closer to her actual translation score of 39 than the original prediction of 35.34.

2 Aim

Three of the established YN scoring formula, *cfg, Δm,* and *Isdt*, may be too cumbersome to be of much use to regular classroom teachers. Though easier to use, the *h-f* adjusted YN scores were not very close to the actual translation test scores (see Table 1). The aim of this study is to introduce and assess a simple regression-based approach to scoring YN tests and to compare that approach to *S&SRF* and *h-f.* As *h-f* had a stronger correlation and a smaller residual than *cfg, Δm,* and *Isdt*, these latter three formulae will not be included.

3 Method

For clarity, the methodologies employed in the pilot study (Stubbe & Yokomitsu, 2012) and the main study (Stubbe, 2013) are reviewed in this section.

3.1 Pilot Study

For the pilot study, four English loanwords (LWs) and four non-loanwords (NLWs) were randomly selected from the top half and the bottom half of each of the eight word frequency levels in the *JACET List of 8000 Basic Words* (JACET Basic Word Revision Committee, 2003; hereinafter the *JACET8000*); for a total of 64 items for each group. Regrettably, three words were found to be in the wrong frequency level and one NLW turned out to be a LW. These four items along with their corresponding member from the opposite group (LW or NLW) had to be deleted from the item pool, leaving 120 words to be tested (Stubbe & Yokomitsu, 2012). A YN test was created containing these 120 words plus 32 pseudowords, all randomly ordered. All pseudowords were randomly selected from *Tests 101–106* of the *EFL Vocabulary Tests* (Meara, 2010). A translation test (English to Japanese, L2–L1) was also created which contained the same 120 words, also randomly ordered. The L2–L1 format test was chosen because translation ability is a strong indicator of which words students can actually understand while reading (Waring & Takaki, 2003) and "asking participants to provide mother-tongue equivalents of the target language words was the most univocal way of verifying recognition" (Eyckmans, 2004, p. 77).

Both the YN and translation tests used in the pilot study were given to Japanese university students enrolled in mandatory English classes ($n = 71$). TOEIC Bridge scores for the participants ranged from 90 though 140, roughly equivalent to 200 through 240 on the TOEIC. The YN test was given at the beginning of class and the translation test was given towards the end of that same class. This was done to ensure each YN test was paired with a translation test.

3.2 Main Study

To improve the separation between adjacent *JACET8000* levels in the main study (Stubbe, 2013), words were sampled only from the bottom half of each level. It was also decided to reduce the total number of tested items to 96; six LWs and six NLWs from each of the eight levels of the *JACET8000*. Forty-four of the 120 words

in the pilot study were included in the main study's item pool. The other 52 words were randomly selected from the various levels of the *JACET8000* as required to complete the desired six LWs and six NLWs per frequency level. Also, only 16 of the best predicting 40 words used to create the prediction formulae (*S&SRF* and the *h-fRF*) were included in the main study's item pool. Again two tests were created: a YN test with 96 words, plus 32 pseudowords including the nine best predicting pseudowords identified in Stubbe and Stewart (2012), with 23 more randomly selected from Tests 101–106 of the *EFL Vocabulary Tests* (Meara, 2010). An L2–L1 translation test, which contained the same 96 words, also randomly ordered, was also created.

Participants in the main study (Stubbe, 2013) took the YN test at the beginning of a class. As in the pilot, this was a paper test in which the students signaled whether they knew a word by filling in either a "Yes" bubble or a "No" bubble beside each item. The same students ($n = 455$) took the paper translation test towards the end of that same class in order. The YN test was scored by means of an optical scanner; the translation test was hand-marked by three native Japanese raters. Interrater reliability was 92%, and Facets analysis (Linacre, 2012) indicated that the raters were basically equal with overall measures of 0.02, 0.02, and –0.04 logits. Participants were all EFL students enrolled in one of four Japanese universities. About 40% of these participants had TOEIC scores in the 350–450 range, considerably higher than the pilot study range of 200–240.

3.3 Regression-Based Prediction Formulae

As discussed above, the improved prediction formula *S&SRF* was generated using the reduced 40-word and 9-pseudoword item set in the pilot study (Stubbe and Stewart, 2012). Two outliers with six FAs each out of a possible nine were deleted ($n = 69$ of 71) from the data. Again, the *S&SRF* formula (p. 6) is:

"True knowledge of tested words $= 3.26 + (0.51 \times \text{YN Score}) - (2.39 \times \text{False Alarms})$".

The formula for *h-fRF,* generated by running a simple regression analysis with the pilot study *h-f* adjusted YN scores as the independent variable and matching translation scores as the dependent variable ($n = 71$; using the reduced item set), was calculated to be:

$$\text{True knowledge of tested words} = 3.28 + (0.51 \times \textit{h-f}).$$

Both of these formulae were created using pilot study data only, to be applied to the main study YN test results. Also it is notable that only 16 of the 40 words as well as the 9 pseudowords used to create the two formulae above were also included in the main study's item set of 96 words and 32 pseudowords. Further, the ability levels of 40% of the participants in the main study were considerably higher than the pilot study participants.

4 Results and Discussion

Means and standard deviations (SDs) for the main study YN test as well as the translation test are presented in Table 2 (Stubbe, 2013). Similar to the pilot

Table 2. Summary of YN and Translation Test Results

Test	Mean	SD	Reliability
YN hits	48.82	17.23	0.96
YN FAs	2.17	3.16	n/a
Tr score	27.06	12.16	0.92

Note. Tr = translation test; SD = standard deviation; Reliability = Cronbach's alpha; $n = 455$; $k = 96$ real-words and 32 pseudowords on the YN test and 96 real-words on the translation test. The mean and SD figures reported in Stubbe (2013) were percentages, and are thus slightly higher.

study (Stubbe & Yokomitsu, 2012), YN test means were considerably higher than the translation mean (48.82 versus 27.06, respectively). This 44.6% decrease between YN and translation test scores is less than the nearly 50% decrease found in the pilot study, likely because the larger population in this main study (Stubbe, 2013) included English learners of higher proficiency. The reliability (Cronbach alpha) for these two tests was high at .96 and .92, respectively.

Means, SDs, correlations with the translation test scores (r), and residuals for the two *regression formulae* (RF) as well as *h-f* are presented in Table 3. Of the three formulae, *h-fRF* had the closest mean to the translation mean (24.85 and 27.06, respectively). Both *h-fRF* and *h-f* shared the highest correlation with translation scores (.833). Using Chen and Popovich's (2002) *t (difference)* formula for paired t-tests as adapted by Field (2009), the difference between the *h-fRF* correlation and *S&SRF* (.833 and .789, respectively) was statistically significant ($t = 6.14$, $df = 452$, $p < .0001$). The *h-fRF* formula also had the lowest of all residuals (RMSE) at 7.30. With the closest mean to translation scores, the highest correlation and smallest residual, *h-fRF* is clearly the best prediction formula.

A one-way analysis of variance revealed that the differences between the means of the translation scores and the two regression-based scoring formulae were statistically significant ($F(2, 1362) = 18.68$, $p < .0001$). Post hoc analysis revealed that the difference between the means of all three pairings was also statistically significant: (a) translation score with *h-fRF*; (b) translation scorewith *S&SRF*; and (c) *S&SRF* with *h-fRF* ($t = 6.69$, 11.70, and 14.83, respectively; df = 454; $p < .0001$, Bonferroni adjustment: .05/3 = .017). Effect sizes (Cohen's d) between translation scores and the two formulae were: *S&SRF* = .377; *h-fRF* = .209. As Cohen (1988) considered an effect size of .2 to be small and .5 to be medium, the difference between the translation scores and the *h-fRF* predicted scores was small. Although the difference between the means of the translation test and the *h-fRF* predictions was significant, the small effect size (.209) suggests that this formula may be able to predict recall knowledge reasonably well.

4.1 Proximity of Individual Predicted Scores

In a subsequent analysis, the score predicted by the two regression-based formulae, *S&SRF* and *h-fRF,* as well as *h-f*, were subtracted from the translation

Table 3. Means, SD, Correlations, and Residuals of Applying the Formulae: *S&SRF*, *h-fRF*, and *h-f*

Test/Formula	Mean	SD	*r*	Residual
Tr Score	27.06	12.16	1	n/a
YN hits	48.82	17.23	0.721	24.90
YN FAs	2.17	3.16	-0.142	n/a
S&SRF	22.96	9.39	0.789	8.54
h-fRF	24.85	8.43	0.833	7.29
h-f	42.29	16.53	0.833	17.90

Note. Tr = translation test; SD = standard deviation; *r* = correlation (Pearson Product-Moment) with Tr (translation test) scores. Residuals were calculated as per Table 1 Note, above ($N = 455$).

score for each of the 455 individual participants to evaluate the usefulness of the individual predictions. Table 4 displays the number of individual participants with predicted scores within 1 percentage point (.96 of 96 words), 5 percentage points (4.8 of 96 words), and 10 percentage points (9.6 of 96 words) of his/her actual translation score. With 54% of predicted scores within 5 percentage points of translation results and 81.5% within 10 percentage points, the *h-fRF* again appears to predict translation scores reasonably accurately. The efficacy of applying regression analysis to YN test results is clearly demonstrated by the *h-fRF* residuals, which are substantially lower than *h-f* (7.3 versus 17.9, see Table 3). Remembering that *h-fRF* and *h-f* share the same *r* value (.833), these results also demonstrate the necessity of calculating residuals, and not just correlations, when comparing YN scoring formulae.

A couple of important differences exist between the pilot study (Stubbe & Yokomitsu, 2012; from which *h-fRF* and *S&SRF* were developed) and the main study (Stubbe, 2013; upon which the prediction formulae were tested). Only Stubbe and Stewart's (2012) *reduced item set* of the best 40 words and 9 pseudowords was used to create the regression-based formulae. This *reduced item set* shared only 16 words and the nine pseudowords with the main study item set of 96 words and 32 pseudowords. As 103 of the total 128 items, in the main study YN item set (82.4%) were used only in the main study, it appears as if *h-fRF* is not overly item dependent. In other words, this formula seems to work well even when tests contain different vocabulary items. Another important difference is between the participants in the two studies in terms of sample size and English proficiency levels.

Table 4. Proximity of Predicted Scores to Actual Translation Scores

Formula	within 1%	within 5%	within 10%	outside 10%
S&SRF	44 (9.7%)	224 (49.2%)	353 (77.6%)	102 (22.4%)
h-fRF	58 (12.7%)	247 (54.3%)	373 (82.0%)	82 (18.0%)
h-f	5 (1.1%)	40 (8.8%)	122 (26.8%)	333 (73.2%)

Note. $n = 455$; 'within x%' denotes percentage points, i.e., 'within 5%' means within five percentage points of the total 96 items, i.e. 4.8 words.

Whereas *h-fRF* was based on 71 low level learners with TOEIC scores of about 200–240, the participants in the main study numbered 455 with TOEIC scores ranging from 200 through 450. Hence, *h-fRF* appears to work well with a wider range of proficiency levels.

5 Conclusion

This study was an investigation into predicting meaning recall (L2–L1) translation test scores from YN test real-word and pseudoword results. Two regression-based prediction formulae, developed using test results from the pilot study following the method described in Stubbe and Stewart's (2012) study, were compared with the established *correction for guessing* scoring formula *h-f*. Results suggest that the two regression-based formulae delivered better predictions. The *h-fRF* formula proved to be the overall best predictor. It was also found that *h-fRF* was not overly item dependent, nor ability level dependent.

This study has demonstrated the usefulness of a new YN scoring formula derived from a simple regression analysis of the *h-f* adjusted YN scores in one study and applied to the *h-f* adjusted YN scores in a different study. By first calculating *h-f* adjustments to YN test results and modifying those adjusted scores using a regression-based prediction formula, such as *h-fRF*, the prediction ability of the YN test can be substantially improved. These results have certain implications for EFL teachers. One trusted means of checking vocabulary knowledge is to test the students on a selection of words they will encounter in a language activity, or text, using a meaning recall translation test. However, the marking of this testing format can be quite cumbersome, especially with large numbers of students and/or items. The present study has found that by adjusting the results of a YN test of the same words, using a simple regression-based scoring formula, *h-fRF*, produced predicted scores that are within 10 percentage points of the actual translation test scores for 82% of participants. For example, an individual *h-fRF* predicted score of 60% of the total number of tested words has an 82% chance of falling within an actual knowledge range between 50% and 70% as demonstrated by a translation test, and a 54% chance of falling between 55% and 65%. Thus, teachers can have reasonable confidence in YN scores adjusted by *h-fRF*, while avoiding the drudgery of marking translation tests.

The *h-fRF* formula, "True knowledge of tested words $= 3.28 + (0.51 \times h\text{-}f)$", was derived from and consequently appears to work well with low-level Japanese EFL students. This formula may require re-calibration for learners of different ability levels, and/or cultures. By first giving students a YN test followed by a translation test of the same items, calculating the *h-f* adjusted YN scores, and then performing a simple regression analysis (with *h-f* adjusted YN scores as the independent variable and translation scores as the dependent variable), a revised *h-fRF* can be calculated. From the regression table, coefficients similar to "3.28" for the intercept and "0.51" for *h-f* will be provided. Using the revised formula, teachers should be able to replace translation tests with the easier YN test format.

References

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. A. Hutson (Ed.), *Advances in reading/language research* (Vol. 2, pp. 231–256). Greenwich, CT: JAI Press.

Barrow, J., Nakanishi, Y., & Ishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System*, *27*(2), 223–247. doi:10.1016/S0346-251X(99)00018-4

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, *16*, 131–162. doi: 10.1177/026553229901600202

Chen, P., & Popovich, P. (2002). *Correlation: Parametric and non-parametric measures*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-139. Thousand Oaks, CA: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

De Veaux, R., Velleman, P., & Bock, D. (2008). *Stats: Data and models*. Essex, UK: Pearson Education Ltd.

Eyckmans, J. (2004). *Measuring receptive vocabulary size*. Utrecht, the Netherlands: LOT (LandelijkeOnderzoekschoolTaalwetenschap).

Field, A. (2009). *Discovering statistics using SPSS (3rd ed.)*. London, UK: Sage Publications.

Gyllstad, H., Vilkaite, L., Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL International Journal of Applied Linguistics 166*(2), 276–303.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes–no vocabulary test: correction for guessing and response style. *Language Testing*, *19*(3), 227–245. doi:10.1191/0265532202lt229oa

JACET Basic Word Revision Committee. (2003). *JACET list of 8000 basic words*. Tokyo: Japan Association of College English Teachers.

Linacre, J. M. (2012). *Facets computer program for many-facet Rasch measurement, version 3.70.0*. Beaverton, Oregon: Winsteps.com. Retrieved from: http:// www.winsteps.com/index.htm

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

Meara, P. (2010). *EFL vocabulary tests*. Swansea: Lognostics second edition. Retrieved from: http://www.lognostics.co.uk/vlibrary/meara1992z.pdf

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142–154.

Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*(1), 73–98. doi:10.1191/ 0265532206lt321oa

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*(4), 489–509. doi:10.1177/0265532212438053

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*(2), 282–308. doi:10.3138/cmlr.56.2.282

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355–371. doi:10.1177/026553229301000308

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

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies*, *7*(2), 105–125.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York: Palgrave Macmillan. doi: 10.1057/9780230293977

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

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

Stubbe, R. (2012a). Searching for an acceptable false alarm maximum. *Vocabulary Education & Research Bulletin*, *1*(2), 7–9.

Stubbe, R. (2012b). Do pseudoword false alarm rates and overestimation rates in YN vocabulary tests change with Japanese university students' English ability levels? *Language Testing*, *29*(4), 471–488. doi: 10.1177/0265532211433033

Stubbe, R. (2013). Comparing regression versus correction formula predictions of passive recall test scores from yes-no test results. *Vocabulary Learning and Instruction*, *2*(1), 39–46.

Stubbe, R., & Hoke, S. (2014). Comparing YN test correction formula predictions of passive recall test results. In R. Chartrand, G. Brooks, M. Porter, & M. Grogan (Eds.), *The 2013 PanSIG conference proceedings* (pp. 72–78). Nagoya, Japan: JALT.

Stubbe, R., & Stewart, J. (2012). Optimizing scoring formulae for YN vocabulary checklists using linear models. *Shiken Research Bulletin*, *16*(2), 2–7.

Stubbe, R., & Yokomitsu, H. (2012). English loanwords in Japanese and the JACET 8000. *Vocabulary Education & Research Bulletin*, *1*(1), 10–11.

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. Retrieved from: http://nflrc.hawaii.edu/rfl/October2003/waring/waring.pdf.

Vocabulary Learning and Instruction
Volume 4, Issue 2, December 2015
http://vli-journal.org

Commentary

Low-Confidence Responses on the Vocabulary Size Test

T. P. Hutchinson
Centre for Automotive Safety Research, University of Adelaide
doi: http://dx.doi.org/10.7820/vli.v04.2.hutchinson

Abstract

McDonald and Asaba (*Vocabulary Learning and Instruction*, 2015) reported an administration of the *Vocabulary Size Test* that was modified to include "I don't know" as a fifth response option on all items, and in which participants later responded to the items originally marked as "I don't know". McDonald and Asaba were inclined to favour the score calculated without the later (reluctant or low-confidence) responses. It is argued here that this goes beyond the data. In many educational contexts, strong encouragement is given to respond when unsure, as examinees to have a better-than-chance probability of being correct, and will be disadvantaged if they do not respond.

In a vocabulary or other educational test, how important are correct responses given with low confidence? I would like to add a few comments to those of McDonald and Asaba (2015) in contrasting the performance of two students who they referred to as Rena and Risako. The test was the *Vocabulary Size Test* (Nation and Beglar, 2007) with responses in Japanese, modified to include "I don't know" as a fifth response option on all items. The four participants were instructed not to guess if unsure, but to select "I don't know" instead. On a second pass through the test, participants answered the items originally marked as "I don't know". It seems reasonable to refer to this second set of attempts as low-confidence or reluctant responses. McDonald and Asaba also interviewed the four participants about their reasoning in selecting responses in the second pass. I regard empirical studies using unusual response formats as very valuable. The research by McDonald and Asaba qualifies for two reasons—distinguishing responses according to level of confidence, and introspection by the participants.

1 Reconsideration of Results

Rena scored 84 without guesses, and 91 with all responses included. Risako scored 54 without guesses, and 89 with all responses included. Two further participants, Rika and Mari, respectively scored 51 and 50 without guesses, and 83 and 81 with all responses included. McDonald and Asaba are rightly cautious about giving too much weight to one type of score (e.g., without guesses) rather than another (e.g., all responses), but nevertheless they rather favour the score

Please cite this article as: Hutchinson, T. (2015). Low-Confidence Responses on the Vocabulary Size Test. *Vocabulary Learning and Instruction*, *4* (2), 49–51. doi: 10.7820/vli.v04.2.hutchinson

without guesses. In both the Abstract and the Conclusion of the paper, they say there are much greater differences between types of score for the lower proficiency learners, and so they must be accepting the scores (without guesses) of 54 for Risako, 51 for Rika, and 50 for Mari as valid indicators of lower proficiency.

My opinion is that in giving precedence to the score without guesses, McDonald and Asaba are going beyond their data. In many educational contexts, there is strong encouragement of examinees to respond when unsure, because examinees usually have a better-than-chance probability of being correct when unsure, and they will be disadvantaged if they do not respond. This disadvantage may be greater for some personality types than others. It is widely thought, I believe, that strong encouragement to examinees to respond to all items is a simple way of minimising the penalty for not guessing that is otherwise imposed on some people. The arguments otherwise of McDonald and Asaba are not strong.

Table 1 is a reorganisation of results in Tables 2 and 3 of McDonald and Asaba (2015). It shows, for each of the participants, the numbers of responses of each type. In particular, Rena confidently answered 33 items wrongly. (This is calculated as 140 items, minus 23 "I don't know" responses, minus 84 confident correct responses = 33 confident wrong responses.) With the data arranged like this, a striking story can be seen.

(1) Confident responses. Rena gets a lower proportion correct than the other three participants.
(2) Confident responses. For all participants, some were wrong, despite the instruction to select "I don't know" if unsure.
(3) Guess responses. Rena gets a lower proportion correct than the other three participants. Rena's responses were correct barely above the chance level (7 out of 23 is 30%, compared with a chance level of 25%). Other participants showed partial knowledge, in the sense of a better-than-chance probability of being correct (47%, 38%, and 38%).
(4) Rena responded at first pass (i.e., confidently) more frequently than the other participants.

Table 1. Confident Responses and Guesses: Correct and Wrong Answers

	Confident responses		Guess responses	
	Correct	Wrong	Correct	Wrong
Rena	84	33	7	16
Risako	54	11	35	40
Rika	51	4	32	53
Mari	50	9	31	50

Note: Confident refers to responses given on the first pass through the test; guess refers to responses given on the second pass to items initially marked as "I don't know". A guess in this sense may be based on some knowledge, it is not necessarily a random guess.

Above-chance performance in items answered with low confidence is a common finding in the literature (Hutchinson, 1982).

Rena was the participant who scored highest, yet at each level of confidence, her probability of being correct was the lowest of all four participants. It may be that Rena differed from the other participants in how she used the term "unsure"—McDonald and Asaba say that "Participants were explicitly instructed not to guess on items they were unsure about, but to select 'I don't know' in these instances instead". That is, she may not be more proficient than Risako, but instead be more confident (or more willing to claim confidence) in her responses. Educational testing is usually intended to measure proficiency (or achievement, ability, aptitude, etc.), and not personality traits and states. I expect that there are some occupations for which Rena's personality and behaviour are advantageous, but I also expect there are others for which people like Risako are better suited.

2 Discussion

McDonald and Asaba had participants express a level of confidence in their responses, and also report on their thought processes. Various other formats have been used from time to time to try to estimate either partial knowledge or willingness to respond (e.g., answer-until-correct, "None of the above" as a response option, and nonsense items). Data from such formats are used best if there is some theory available—a theory with more psychological content than Item Response Theory has, and yet is broad-brush rather than requiring detailed analysis of each item individually (Hutchinson, 1982). It may be over-ambitious to seek quantitative comparison of data with theory, but possible to look for qualitative features in the data, such as above-chance probability of correctness with reluctant responses (as in McDonald and Asaba) and with second attempts after a first response is wrong.

Acknowledgements

The Centre for Automotive Safety Research (CASR) receives core funding from both the South Australian Department for Planning, Transport and Infrastructure and the South Australian Motor Accident Commission. The views expressed are those of the authors and do not necessarily represent those of the University of Adelaide.

References

Hutchinson, T. P. (1982). Some theories of performance in multiple choice tests, and their implications for variants of the task. *British Journal of Mathematical and Statistical Psychology*, *35*(1), 71–89. doi:10.1111/j.2044-8317.1982.tb00642.x

McDonald, K., & Asaba, M. (2015). "I don't know" use and guessing on the bilingual Japanese Vocabulary Size Test: A preliminary report. *Vocabulary Learning and Instruction*, *4*(1), 16–25. doi: 10.7820/vli.v04.1.mcdonald.asaba

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13. Retrieved from http://jalt-publications.org/tlt/issues/2007-07_31.7

Vocabulary Learning and Instruction
Volume 4, Issue 2, December 2015
http://vli-journal.org

Cardiff University PhD Program in Applied Linguistics (Lexical Studies)

Introduction

This program is designed specifically to meet the needs of part-time, distance-learning students.

All students on the program work in the broad areas of lexical research and lexical perspectives on language issues. Specific topics include, for example, vocabulary processing, assessment of vocabulary knowledge, second-language acquisition, lexical attrition in L1 or L2, lexical loss in acquired communication disorders, formulaic language, corpus approaches to lexical investigation, and the interface of lexis and syntax.

Main Supervisors on the program include Professor *Tess Fitzpatrick*, Professor *Alison Wray*, and Dr *Dawn Knight*. In addition, Professor *Paul Meara*, Honorary Professor at Cardiff University and international expert on vocabulary acquisition, advises, and contributes to the Lexical Studies research group. Other staff who contribute to supervision, bringing a wealth of additional knowledge and perspectives, include Drs Michelle Aldridge, Mercedes Durham, Lisa El Refaie, Lise Fontaine, Chris Heffer, Nick Wilson, and Tereza Spilioti.

Program Description

The program takes an empirical approach to the investigation of lexical acquisition, processing, production, and attrition. Many research students on the program are also language teaching practitioners, and choose to focus on the role of vocabulary in pedagogical and assessment processes. Others investigate lexical phenomena such as polysemy, collocation, and associative links, or use lexical profiling techniques to examine language use in specific contexts. Methods typically used include corpus-based and psycholinguistic tools, and studies might investigate the effects of specific interventions on lexical performance, or make predictions based on theoretical models. Research topics are not restricted to the L2 context; the supervisory team's expertise in forensic linguistics, healthcare communication, non-verbal communication, language variation, and communication disorders enables us to conduct lexical investigations in these areas too.

At the time of writing (summer 2015) there are 10 students on the program, and from this year on we will be recruiting up to 5 students per year, with an anticipated broadening of the range of perspectives on lexis, as made possible by the specialisms of the team. Current students' research covers the following areas: measuring collocation knowledge; polysemy and the acquisition of new word

Please cite this article as: Cardiff University. *Vocabulary Learning and Instruction*, *4* (2), 52–55.
doi: 10.7820/vli.v04.2.cardiff

meanings; word association profiles; autonomous vocabulary learning; item difficulty in vocabulary tests; vocabulary load of proper nouns; individual differences in lexical storage; phonological patterns in formulaic sequences; formulaic frames and oral fluency; use of circumlocution strategies by people with dementia. We encourage students to publish their work, and most produce journal articles, book reviews, or book chapters during their PhD candidacy (see the examples at the end of this article).

Most students on the program live and work overseas, and the program addresses the challenges of this study context through a structured framework, based on the one Paul Meara established when he set up the program 20 years ago (then at Swansea). There are regular *newsletters* and monthly research training tasks, and during each year of the program students are expected to complete an empirical study and present their year's research in an appropriate conference and/or publication. In the first year the empirical work is a replication of a published study; the benefits of this approach, including student perspectives, are discussed in Fitzpatrick (2012). The thesis is built incrementally with each study representing a chapter of the final thesis. Students receive support from the supervisory team throughout the preparation of these outputs, and are given detailed feedback on submitted work. At 6-monthly intervals, students' progress is formally reviewed, and each year the supervisory panel meets (students attend via Skype) to confirm that threshold criteria have been met in order to allow the student to progress to the next year of the program.

Students attend an *annual Lexical Studies Conference*, held in the UK, and most consider this to be the highlight of their study year. The conference provides an opportunity for 3 days of face-to-face contact between students, supervisors, and established researchers; invited speakers at our last three conferences include *Laurence Anthony* (*AntConc*), *Averil Coxhead*, *Chris Hall*, *Patrick Hanks*, *Birgit Henriksen*, *Mike McCarthy*, and *Leah Roberts*. The sense of belonging to a wider research community is particularly important for students working in distance mode, and the conference and newsletter help to maintain the sense of a virtual research network, consisting not only of current students, but also of post-doctoral researchers and scholars with an established reputation in vocabulary studies. The conference and selective sharing of the monthly tasks enable the students to learn from, and provide practical support and encouragement to, each other. Aside from the annual conference, the only required visit to the UK is for the viva examination at the end of the 6-year program. However, we can arrange for students to spend study time in Cardiff at any point during their studies. Most of the direct communication between students and supervisors is via email, supplemented with occasional Skype meetings. The program has a dedicated module on the University's virtual learning environment, and is very well supported by the University library service; this support includes one-to-one assistance from the subject librarian and an excellent electronic collection of journals and books.

The structure of the program comprises tasks and targets in four research strands: an empirical strand, a critical strand, a presentational strand, and a

general skills strand. A typical working schedule for part-time students can be summarized as:

	Year	Empirical strand	Critical strand	Presentational strand	General skills strand
Phase 1	1	1 replication	8 critical tasks; 1 book review	Poster presentation	As determined by regular reviews of training needs; may include data management, research methodology, statistical analysis, computing, other research skills
	2	1 experiment	8 critical tasks; 1 book review	Minor conference	
Phase 2	3	2 experiments	8 critical tasks; literature review	Conference	
	4	2 experiments	4 critical tasks; a literature review	Journal paper	
Phase 3	5	1 experiment	Pre-final draft of the thesis	Major conference	
	6	Final draft of the thesis		Formal defence of thesis	

Graduates of the program, many of whom worked under Paul Meara's supervision before the program moved to Cardiff, include Andy Barfield, Huw Bell, Jon Clenton, David Coulson, Tess Fitzpatrick, Simon Fraser, George Higginbotham, Marlise Horst, Tadsmitsu Kamimoto, Masamichi Mochizuki, Ian Munby, Hilary Nesi, Mitsuru Orita, Richard Pemberton, Jim Ronald, Rob Waring, Clarissa Wilks, Brent Wolter. Most graduates of the program work in academic posts and are active researchers, and remain in contact with the program via the monthly newsletter and annual conference. *Lexical Processing in Second Language Learners* (Fitzpatrick & Barfield, 2009) is a collection of work from this group.

Entry to the program is competitive, and successful applicants will produce a convincing, well-researched proposal in a relevant field. In addition, they will typically have demonstrated a strong performance at Master's level, will have completed a course in research methods and will have some experience of conducting empirical research. An IELTS score of minimum 7.5 is required for applicants who are not first-language English speakers. Decisions on the October 1 intake to the program are made in June each year, so applications should be made in advance of that time, for consideration in the coming round. However, enquiries can be made at any time. Often, applicants are invited to "lurk" on the program as observers for a while, so as to better understand the expectations and opportunities of the program. Some applicants also attend the annual conference where they can meet current students and gain a direct insight into the type of research that is done.

More information about the program, including fee bursary information, can be found on the Cardiff University website (from the home page, search for "PhD Applied Linguistics") and the program blog is at *http://blogs.cardiff.ac.uk/lexicalstudies/*.

Recent Publications by Current PhD Students

Brown, Dale (2014). The power and authority of materials in the classroom ecology. *The Modern Language Journal*, *98*(2), 658–661. doi:10.1111/modl.12095

Brown, Dale. (2013). Types of words identified as unknown by L2 learners when reading. *System*, *41*(4), 1043–1055. doi:10.1016/j.system.2013.10.013

Brown, Dale. (2013). Knowledge of collocations. In J. Milton & T. Fitzpatrick (Eds.) *Dimensions of vocabulary knowledge*. Basingstoke: Palgrave Macmillan.

Brown, Dale. (2012). The frequency model of vocabulary learning and Japanese learners. *Vocabulary Learning and Instruction*, *1*(1), 20–28. doi:10.7820/vli.v01.1.brown

Klassen, Kimberly. (2014). Review of: Stanislas Dehaene (2009). Reading in the Brain: The New Science of How We Read. New York: Penguin. *International Journal of Applied Linguistics*, *24*(1), 128–130.

Maby, Mark. (2013). Review of: Moreno Jaén, Serrano Valverde & Calzada Pérez (Eds., 2010). 'Exploring New Paths in Language Pedagogy: Lexis and Corpus-based Language Teaching'. London: Equinox. *BAAL News, 103*, 7–9.

Racine, John, Higginbotham, George, & Munby, Ian. (2014). Exploring non-native norms: A new direction in word association research. *VERB*, *3*(2), 13–15.

Racine, John. (2013). Reaction time methodologies and lexical access in applied linguistics. *Vocabulary Learning and Instruction*, 1–5.

Rooks, Matthew. (2014). The effects of motivational factors on Japanese learners of English. *International Journal of Educational Research and Development*, *3*(1), 6–22.

Stewart, Jeff. (2012). A multiple-choice test of active vocabulary knowledge. *Vocabulary Learning and Instruction*, *1*(1), 53–59. doi:10.7820/vli.v01.1.stewart

Stewart, Jeff, Batty, Aaron Olaf, & Bovee, Nicholas. (2012). Comparing multidimensional and continuum models of vocabulary acquisition: An empirical examination of the Vocabulary Knowledge Scale. *TESOL Quarterly*, *46*(4), 695–721. doi:10.1002/tesq.35

Thwaites, Peter. (2014). Maximizing learning from written output. *ELT Journal*, *68*(2), 135–144. doi:10.1093/elt/cct098

References

Fitzpatrick, T. (2012). Conducting replication studies: Lessons from a graduate programme. In G. Porte (Ed.), *Replication research in applied linguistics and second language acquisition: A practical guide* (pp. 151–170). Cambridge, UK: Cambridge University Press.

Fitzpatrick, T., & Barfield, A. (Eds.). (2009). *Lexical processing in second language learners*. Bristol, UK: Multilingual Matters.

Vocabulary Learning and Instruction
Volume 4, Issue 2, December 2015
http://vli-journal.org

The University of Nottingham Vocabulary Research Group

Introduction

The University of Nottingham offers a highly supportive research environment for PhD students interested in vocabulary issues. Students join the Vocabulary Research Group (VRG), which is part of the well-known Centre for Research in Applied Linguistics (CRAL).

The VRG is led by a staff of three *vocabulary experts*, all with international profiles:

Professor Norbert Schmitt

Author of eight books on vocabulary and applied linguistics, and over 50 articles in international peer-reviewed journals. He is interested in all aspects of second-language vocabulary studies, and lectures globally on these topics. For more information, see: http://www.norbertschmitt.co.uk/

Dr. Ana Pellicer-Sánchez

Her research sits at the intersection between applied linguistics and psycholinguistics and focuses on the use of psycholinguistic measures, like eye-tracking and reaction times (RTs), to explore the acquisition and assessment of vocabulary. Her research interests include approaches to vocabulary teaching and learning, and the relationship between vocabulary and reading. For more information, see: http://www.nottingham.ac.uk/english/people/ana.pellicer-sanchez

Dr. Michael Rodgers

His research interests include vocabulary acquisition, lexical coverage, listening comprehension, and language learning from video. He has published in international peer-reviewed journals such as *Applied Linguistics, Language Learning* and *TESOL Quarterly*. For more information, see: http://www.nottingham.ac.uk/english/people/michael.rodgers

Program Description

The VRG currently consists of nine students, each supervised by two of the staff. In addition, the students and staff all meet weekly in a group to discuss lexical issues, practice conference presentations, or practice the development of research methodology. Staff also mentor their students to publish in the best international journals, and would expect each one to publish at least two major articles from their PhD research. Likewise, students are expected to present at major applied linguistics conferences. To date, our students have been very successful in this, and since 2005, have published or co-published over 30 articles in major international

Please cite this article as: The University of Nottingham. *Vocabulary Learning and Instruction*, *4* (2), 56–58. doi: 10.7820/vli.v04.2.nottingham

journals such as *Applied Linguistics*, *Language Learning*, *Language Testing*, and *System*, seven book chapters, and presented at over 30 conferences.

Current Students are Researching the Following Lexical Areas

- Developing the next generation vocabulary test
- Describing different categories of formulaic language
- Exploring the polysemy of collocations
- Creating a pedagogical list of the most useful phrasal verbs
- Creating a list of the most useful affixes for EFL learners
- Exploring vocabulary knowledge in bilinguals
- Modeling the multi-dimensional nature of vocabulary knowledge

The main goal of the VRG is to mentor the next generation of vocabulary researchers. This requires students to come to our program well-prepared. The standard time for British full-time PhDs is three years (although often with a fourth to write up), and most students find this short. Thus, it is important to have a sound background in applied linguistics in general and in the area of vocabulary issues before starting the PhD research. It is also important to be comfortable with the use of spreadsheets and be familiar with statistical techniques (SPSS). Prospective students do not need to be experts, but they must understand statistics and be able to run basic analyses like *t*-tests, ANOVA, correlations, etc. Students are also expected to have strong language skills, both spoken and written. Apart from completing the official university application form, applicants will have to provide solid evidence that they would meet the requirements outlined above.

Each staff member usually takes one or two students each year to begin in the autumn semester (around October 1). Although we take enquiries throughout the year, we make our selection decisions sometime in the spring semester, which means that potential students need to make their formal PhD application to our school by the end of February. <http://www.nottingham.ac.uk/pgstudy/courses/english/appliedlinguisticselt-mphilphd.aspx>

The School of English at the University of Nottingham does not offer a PhD program by distance learning as such. However, students can enroll in a part-time PhD program in which most supervisory contact can take place by distance. This part-time opportunity is subject to agreement by supervisors.

Recent past students with major publications:

Ana Pellicer-Sánchez (University of Nottingham)

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring yes-no vocabulary tests: reaction time vs. nonword approaches, *Language Testing*, *29*(4), 489–509. doi:10.1177/0265532212438053

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

Phil Durrant (University of Exeter)
Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, *26*(2), 163–188. doi:10.1177/0267658309349431

Anna Siyanova-Chanturia (Victoria University of Wellington)
Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and nonnative speakers. *Second Language Research*, *27*(2), 1–22. doi:10.1177/0267658310382068

Wen-ta (Thomas) Tseng (National Taiwan Normal University)
Tseng, W-T., Dörnyei, Z., and Schmitt, N. (2006). A new approach to assessing strategic learning: The case of self-regulation in vocabulary acquisition. *Applied Linguistics*, *27*(1), 78–102. doi:10.1093/applin/ami046

Suhad Sonbul (Umm Al-Qura University)
Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning*, *63*(1), 121–159. doi:10.1111/j.1467-9922.2012.00730.x

Hilde van Zeeland (graduated in 2014)
van Zeeland, H., & Schmitt, N. (2013). 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

Vocabulary Learning and Instruction
Volume 4, Issue 2, December 2015
http://vli-journal.org

Victoria University of Wellington PhD in Applied Linguistics, Second Language Vocabulary Acquisition

Program Instructors and Areas of Specialization

Emeritus Professor *Laurie Bauer* also supervised several vocabulary-based PhD students in the past. He is a world authority on morphology. Laurie is no longer taking on PhD candidates.

Associate Professor *Frank Boers* is best known for his research in the area of collocations, idioms, and formulaic language. He is the co-editor of the journal *Language Teaching Research*.

Dr. *Averil Coxhead* specializes in the areas of vocabulary and pedagogy, multi-word units, and English for Specific/Academic Purposes.

Dr. *Irina Elgort* researches lexical development in a second and foreign language, the bilingual mental lexicon, and reading.

Dr. *Peter Gu* is interested in vocabulary testing and vocabulary learning strategies.

Dr. *Angela Joe* is interested in second-language vocabulary acquisition, English for Specific Purposes, and Language in the Workplace.

Associate Professor *John Macalister*'s research interests include extensive reading and vocabulary.

Dr. *Jonathan Newton*'s work has focused on vocabulary learning through spoken interaction and teacher cognition in vocabulary.

Emeritus Professor *Paul Nation* is interested in a wide range of vocabulary research and is currently working on a picture-based vocabulary size test for young learners. Paul is no longer taking on PhD candidates.

Dr. Anna Siyanova-Chanturia researches the acquisition, processing, and use of multi-word expressions, such as collocations, idioms, multi-word verbs, in a first and second language.

Program Description

The PhD in Applied Linguistics program at Victoria University of Wellington is a research degree requiring the presentation of a thesis after an extended period of research. The PhD is typically done full-time. Candidates must be registered for a minimum of two years, and are expected to complete the degree in three years. The maximum time to complete the degree as a full-time candidate is four years. Candidates typically work on-campus with a maximum of 10 months allowed over the duration of the degree for off-campus data collection. Victoria University has a limited number of full scholarships that fund tuition and living expenses. At

Please cite this article as: Victoria University of Wellington. *Vocabulary Learning and Instruction*, *4* (2), 59–63. doi: 10.7820/vli.v04.2.victoria

present, approximately 20% of our students are fully funded through these scholarships. Funding is also available through faculty research grants to support data collection.

To enter the PhD program you need to have achieved very strong grades in an MA in TESOL or Applied Linguistics, have a good background in research in vocabulary, and provide a suitable topic that can be supervised within the School. We recommend contacting staff about potential topics and being flexible about the area of vocabulary research you want to investigate.

The environment for research at Victoria University of Wellington is highly supportive, with regular meetings of the vocabulary research group. There are currently 12 PhD candidates researching vocabulary with numbers varying between years. We hosted the first ever Vocab@Vic Conference in Wellington, December 18–20, 2013.

Current PhD Students and Working Titles

Oliver Ballance: Concordances and language learners: exploring textual and cognitive dimensions of concordancing in language learning; *TJ Boutorwick*: Productive vocabulary and extensive reading; *Thi Ngoc Yen Dang*: Developing and validating academic written and spoken word lists; *Khadij Gharibi*: Lexical Attrition in Iranian bilingual children in New Zealand; *Lin He*: Effects of explicit instruction about sentence structure on L2 sentence processing; *Myq Larson*: Thresholds, text coverage, vocabulary size, and reading comprehension in Applied Linguistics; *Chi Duc Nguyen*: Vocabulary uptake from listening to TED talks; *Betsy Quero*: The vocabulary load of academic texts; *Brian Strong*: Exploring phrasal verb acquisition difficulties; *Friederike Tegge*: Investigating song-based language teaching and its effect on lexical learning; *Haidee Thomson*: Learning and acquisition of formulaic language by foreign language learners, *Mark Toomer*: Retention of collocations: the effects of incidental and deliberate learning conditions on explicit and implicit knowledge of lexical and grammatical collocations.

Recent and Notable Graduates

2013: Tatsuya Nakata, Optimising second language vocabulary learning from flashcards; *Yosuke Sasao*, Diagnostic tests of English vocabulary learning proficiency: Guessing from context and knowledge of word parts.

2012: Tatsuhiko Matsushita, In what order should Japanese learners learn vocabulary? A corpus-based approach; *Mike Rodgers*, English language learning through viewing television: An investigation of comprehension, incidental vocabulary acquisition, lexical coverage, attitudes and captions; *Joseph Sorell*, Making a high-frequency word list; *Anna Piasecki*, The effects of proficiency on sub-lexical processing in bilingual visual word recognition.

2011: Patrick Foss, Vocabulary use and development in a corpus of Japanese learner blogs; *Zheng Wei*, Word roots in English: Learning English words through form and meaning similarity.

Averil Coxhead, *Irina Elgort*, *Angela Joe*, and *Jonathan Newton* are graduates who currently hold positions at Victoria University of Wellington. *Anna C-S Chang*, Hsing-Wu University, Taiwan; *Teresa Mihwa Chung*, Korea University, Korea; *Tatsuhiko Matsushita*, The University of Tokyo, Japan; *Anna Piasecki*, University of the West of England; *Michael Rodgers*, Nottingham University; *Yosuke Sasao*, Toyohashi University of Technology, Japan; *Stuart Webb*, Western University, Canada; *Wei Zheng*, Beijing Foreign Studies University, China.

Notable Published Papers

Boers, F., & Strong, B. (in press). An evaluation of textbook exercises on collocations. In B. Tomlinson (Ed.), *Second language acquisition research and materials development for language learning*. Taylor & Francis.

Boers, F. (2015). Words in second language learning and teaching. In J. Taylor (Ed.), *The Oxford handbook of the word* (pp. 582–596). Oxford: Oxford University Press.

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effectiveness of exercises on verb-noun collocations. *Language Teaching Research*, *18*(1), 50–70.

Boers, F., Eyckmans, J., & Lindstromberg, S. (2014). The effect of a discrimination task on recall of L2 collocations and compounds. *International Journal of Applied Linguistics*, *24*, 357–369. doi:10.1111/ijal.12033

Boers, F., Lindstromberg, S., & Eyckmans, J. (2014). Is alliteration mnemonic without awareness-raising? *Language Awareness*, *23*, 291–303.

Boers, F., Lindstromberg, S., & Webb, S. (2014). Further evidence of the comparative memorability of alliterative expressions in second language learning. *RELC Journal*, *45*, 85–99. doi:10.1177/0033688214522714

Boers, F., & Webb, S. (2015). Gauging the semantic transparency of idioms: Do natives and learners see eye to eye? In R. Heredia & A. Cieslicka (Eds.), *Bilingual figurative language processing* (pp. 368–392). Cambridge University Press. doi:10.1017/cbo9781139342100.018

Chang, A. (in press). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition*.

Coxhead, A. (2013). Vocabulary and ESP. In B. Paltridge & S. Starfield (Eds.). *The Handbook of English for specific purposes* (pp. 115–132). Boston, MA: Wiley-Blackwell.

Coxhead, A. (in press). (Ed.). Special issue on vocabulary and pedagogy. *Language Teaching Research*.

Coxhead, A. (2014) (Ed.). *New ways in teaching vocabulary, revised*. Alexandria, VA: TESOL Inc.

Coxhead, A., & Bytheway, J. (2015). Learning vocabulary using two massive online resources: You will not blink. In D. Nunan & J. Richards (Eds.), *Learning beyond the classroom* (pp. 65–74). New York: Routledge.

Coxhead, A., Nation, P., & Sim, D. (2015). Vocabulary size and native speaker secondary school students. *New Zealand Journal of Educational Studies.* doi:10.1007/s40841-015-0002-3

Dang, T., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66–76.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*(2), 253–272.

Elgort, I., Perfetti, C. A., Rickles, B., & Stafura, J.Z. (2014). Contextual learning of L2 word meanings: Second language proficiency modulates behavioural and ERP indicators of learning. *Language, Cognition and Neuroscience*, *30*(5), 506–528. doi:10.1080/23273798.2014.942673

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.

Greene, J., & Coxhead, A. (2015). *Academic vocabulary for middle school students: Research-based lists and strategies for key content areas.* Baltimore, MD: Brookes Publishing.

Gu, Y. (2013). Vocabulary learning strategies. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics.* Oxford: Wiley-Blackwell.

Gu, Y. (2012). Second language vocabulary. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 307–309). London: Routledge.

Nation, I. S. P., & Anthony, L. (2013). Mid-frequency readers. *Journal of Extensive Reading*, *1*(1), 5–16.

Nation, I. S. P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language*, *26*(2), 1–16.

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

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

Nation, P., & Coxhead, A. (2014). Vocabulary size research at Victoria University of Wellington, New Zealand. *Language Teaching*, *47*(3), 398–403. doi:10.1017/S0261444814000111

Newton. J. (2013). Incidental vocabulary learning in classroom communication tasks. *Language Teaching Research*, *17*(2), 164–187. doi:10.1177/1362168812460814

Rogers, J., Webb, S., & Nakata, T. (2014). Do the cognacy characteristics of loanwords make them more easily learned than noncognates? *Language Teaching Research*, *19*(1), 9–27. doi:10.1177/1362168814541752

Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. *Corpus Linguistics and Linguistic Theory.* doi:10.1515/cllt-2014-0016

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, *27*, 251–272. doi:10.1177/0267658310382068

Siyanova-Chanturia, A., & Martinez, R. (2014). The idiom principle revisited. *Applied Linguistics*. doi:10.1093/applin/amt054

Siyanova-Chanturia, A., & Spina, S. (Forthcoming, 2015). Investigation of native speaker and second language learner intuition of collocation frequency. *Language Learning*.

Webb, S., & Boers, F. (in press). Research timeline: Teaching and learning collocation in adult second and foreign language learning. *Language Teaching: Surveys and Studies*.

Webb, S., & Macalister, J. (2013). Is text written for children appropriate for L2 extensive reading? *TESOL Quarterly*, *47*(2), 300–322. doi:10.1002/tesq.70

Webb, S., Newton, J., & Chang, A. C-S. (2013). Incidental learning of collocation. *Language Learning*, *63*(1), 91–120. doi:10.1111/j.1467-9922.2012.00729.x

Wei, Z. (2015). Does teaching mnemonics for vocabulary learning make a difference? Putting the keyword method and the word part technique to the test. *Language Teaching Research*, *19*(1), 43–69. doi:10.1177/1362168814541734

Wei, Z., & Nation, P. (2013). The word part technique: A very useful vocabulary teaching technique. *Modern English Teacher*, *22*(1), 12–16.

Vocabulary Learning and Instruction
Volume 4, Issue 2, December 2015
http://vli-journal.org

Carnegie Mellon University Doctoral Program in Second Language Acquisition, Department of Modern Languages

Core Faculty

Mariana Achugar, Discourse Analysis

Keiko Koda, Reading and Biliteracy

Brian MacWhinney, Psycholinguistics

Naoko Taguchi, Interlanguage Pragmatics

G. Richard Tucker, Language Policy and Planning

Remi A Van Compernolle, Sociocultural Theory

Program Description

The PhD in Second Language Acquisition (SLA) provides students and future professionals with a theoretical grounding in linguistics and cognitive psychology, coupled with broad training in qualitative and quantitative research methodology. The goal of the PhD program in SLA is to create independent and insightful researchers capable of using analytical and empirical methods to illuminate and understand the acquisition, use, and maintenance of second languages.

The program offers four foci: reading development in second languages, social dimensions of second language learning, instruction and second language learning, and cognitive aspects of second language learning.

The reading concentration addresses the following questions:

- Does reading development differ among first- and second-language learners?
- How do reading skills acquired in one language affect reading development in another language?
- How does oral language proficiency relate to reading development in a second language?
- In what ways is learning to read similar (or different) across languages?
- What type of intervention and scaffolding can content teachers provide to help language learners develop their academic literacy?

Four-year Residential Program

Currently, 11 students are in residence (4 focusing on second language literacy development, 2 on instruction, 2 on cognitive aspects, and 2 on social dimensions).

Please cite this article as: Carnegie Mellon University. *Vocabulary Learning and Instruction*, *4* (2), 64–65. doi: 10.7820/vli.v04.2.carnegie.mellon

Notable Graduates and Notable Published Papers

Megumi Hamada (2005). Associate Professor of English at Ball State University.

Hamada, M., & Koda, K. (2008). Influence of first language orthographic experience on second language decoding and word learning. *Language Learning*, *58*, 1–31. doi:10.1111/j.1467-9922.2007.00433.x

Hamada, M., & Koda, K. (2010). The role of phonological decoding on second language word-meaning inference. *Applied Linguistics*, *31*(4), 513–531. doi:10.1093/applin/amp061

Hamada, M., & Koda, K. (2011). The role of the phonological loop in English word learning: A comparison of Chinese ESL learners and native speakers. *Journal of Psycholinguistic Research*, *42*, 75–92. doi:10.1007/s10936-010-9156-9

Hamada, M., & Koda, K. (2011). L2 word-form learning: L1 orthographic experience and sensitivity to word-form. *System*, *39*, 500–508. doi:10.1016/j.system.2011.10.011

Chan Lu (2009). Assistant Professor of Modern Languages and Literatures at Loyola Marymount University.

Lu, C., & Koda, K. (2011). Impacts of home language and literacy support in English-Chinese biliteracy acquisition among Chinese heritage language learners. *Heritage Language Journal*, *8*, 44–80.

Lu, C., Koda, K., Zhang, D., & Zhang, Y. (in press). Effects of semantic radical properties on character meaning extraction and inference. *Writing System Research*.

Dongnbo Zhang (2010). Assistant Professor of Teacher Education, Michigan State University.

Zhang, D., & Koda, K. (2011). Home literacy environment and word knowledge development: A study of young learners of Chinese as a Heritage Language. *Bilingual Research Journal*, *34*, 4–18. doi:10.1080/15235882.2011.568591

Zhang, D., & Koda, K. (2012). Morphological awareness, lexical inferencing vocabulary knowledge and L2 reading comprehension: Testing direct and indirect effects. *Reading and Writing*, *25*, 1195–1216. doi:10.1007/s11145-011-9313-z

Zhang, D., & Koda, K. (2013). Morphological awareness and reading comprehension in a foreign language: A study of young Chinese EFL learners. *System*, *41*, 901–931. doi:10.1016/j.system.2013.09.009

Pooja Reddy Nakamura (2011). Senior Research Associate, American Institutes of Research.

Reddy, P., & Koda, K. (2012). Orthographic constraints on phonological awareness in biliteracy development. *Writing System Research*, *4*, 1–21.

Nakamura, P., Koda, K., & Joshi, M. (2014). Biliteracy acquisition in Kannada and English: A developmental study. *Writing System Research*, *6*, 132–147. doi:10.1080/17586801.2013.855620