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).

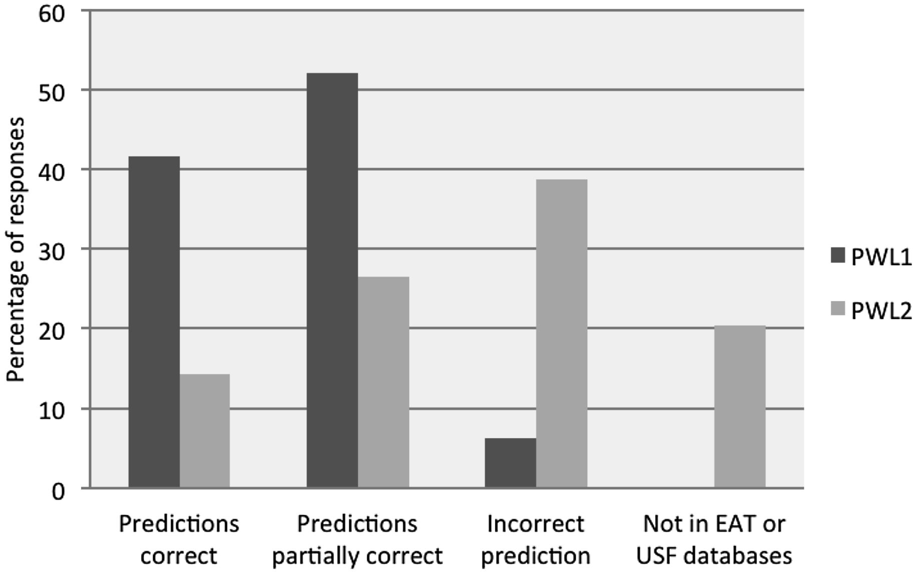

Figure 1. Prediction scores for two groups of prompt words.

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.