Reading in a Foreign Language October 2019, Volume 31, No. 2
ISSN 1539-0578 pp. 150–172

Two approaches to extensive reading and their effects on L2 vocabulary development

TJ Boutorwick
Victoria University of Wellington
New Zealand

John Macalister
Victoria University of Wellington
New Zealand

Irina Elgort
Victoria University of Wellington
New Zealand

Abstract

One avenue for developing second language (L2) vocabulary knowledge is through Extensive Reading (ER). ER can provide opportunities for incidental learning to occur. Class time is often too restricted for sufficient attention to deliberate learning (Hunt & Beglar, 2005) meaning ER is important for L2 vocabulary development. This article builds on ideas in the recent two-part *Reading in a Foreign Language* ER discussion forum by investigating two implementations of ER and their effects on L2 vocabulary development: a traditional ER-only approach, and an ER-plus approach which supplements ER with post-reading discussion implemented in small groups. L2 English learners enrolled at a university in Aotearoa New Zealand read five graded readers during normal class time. Latent Semantic Analysis was used to measure the development of word association knowledge of 60 target words. The findings revealed facilitative effects of both ER approaches. Supplementing ER with discussion provided opportunities for further development.

Keywords: ER, semantic knowledge, word associations, interaction, language production, vocabulary, incidental learning

Over the years, extensive reading (ER) has been variably understood. Bruton (2002), for example, identified four different ways in which ‘extensive’ had been interpreted in the literature at the time. A major contribution to developing a common understanding of ER came with the proposal of ten principles (Day & Bamford, 2002). These principles have provided a solid basis for conceptualizing ER in the language-learning classroom for almost two decades. At times they appear to have attained the status of 'rules', or principles to adhere to at all costs (Macalister, 2015). However, in two recent issues of *Reading in a Foreign Language*, invited contributors considered the ten principles and a consensus emerged that some principles are more important than others. In their contribution to the forum, Jeon and Day (2015) identified five core principles in ER programmes, and Day (2015) ranked the principles as applied in 44 reports of ER programmes. The top four principles were that learners read as much as possible, they choose what they want to read, a variety

of reading material on a wide range of topics is available for them, and this reading material is easy. These four principles would seem to embody the elements of the folk definition of ER: lots of easy, enjoyable reading.

Closely linked to the question of what ER is, is the question of how ER should be implemented. The traditional view, and one that captures Day and Bamford's (2002) principles, has ER as a stand-alone activity. This traditional view has, however, been challenged, most notably perhaps by Green (2005) who, after examining the mandated practice of ER in Hong Kong secondary schools, described it as resembling a "monastic detention session with teachers sitting at the head of the class enforcing a rule of silence" (p. 308). To avoid this, Green suggested that teachers should instead incorporate ER as part of a task-based approach, to allow for further engagement with the books that the learners read. His article inspired a response from Macalister (2008) who reported on the successful implementation of a traditional, stand-alone ER programme in a tertiary context, and who concluded that "while the way in which extensive reading is incorporated in an EAP programme will vary from classroom to classroom, … extensive reading definitely can have a place in such four-skills teaching programmes" (p. 255). The nature of the programme will be determined by the context.

Literature Review

Dimensions of Extensive Reading

Recognition of the role context plays in determining how ER is implemented led Waring and McLean (2015), in a recent meta-review, to examine the ER literature to determine what the core dimension(s) of ER may be. This examination led to a distinction between core and variable dimensions. The authors note that the core dimension of ER includes fluent comprehension of text, read at a high speed and in large amounts, with a focus on meaning. One difficult question that arises from this definition of ER is how many books constitute 'large amounts'? Taguchi, Takayasu-Maass, and Gorsuch (2004), investigating the effects of ER on reading fluency, had their participants read graded readers over a 17-week period, amounting to 205 pages of text. In a more recent study, Sakurai (2015) investigated the influence of translation on reading amount, proficiency and speed in ER, by having participants read over a 15-week period. The participants were grouped according to their second language (L2) proficiency and assigned a baseline amount to read. Lower-proficiency students were told to read at least 15,000 words, while higher-proficiency students had a baseline of 35,000 words. These two studies help to give an idea of the amount of reading that researchers use to ensure their participants are reading in accordance with the core dimension of ER. The variable dimension Waring and McLean (2015) discuss can be categorized as pedagogical decisions. Examples of this dimension include the location where the reading takes place, who selects the texts, and whether or not there will be supplementary activities used in conjunction with the reading. The current study adopts Waring and McLean's (2015) core definition of ER as the fluent comprehension of a large amount of text with a focus on meaning.

In a recent meta-analysis of research investigating facilitating effects that ER can have on language development, Nakanishi (2015) highlights that the dimensions of ER have been utilized to varying degrees in research. In his meta-analysis, he selected 34 studies and examined the nature of the research conducted. This included research design (e.g., control group versus experiment group; pretest and post-test), assessment area (e.g., vocabulary knowledge, reading comprehension), and assessment instrumentation (e.g., multiple-choice tests). The meta-analysis revealed that some of these studies included a variety of supplementary activities such as book reports and small-group

discussion (e.g., Beglar, Hunt, & Kite, 2012; Horst, 2005; Yamashita, 2008), others devoted class time to completing worksheets and filling in vocabulary notebooks based on the books the learners read (e.g., Horst, 2005), while other studies included pre-reading and post-reading activities such as teacher-led instruction on vocabulary-learning strategies (e.g., Al-Homoud & Schmitt, 2009).

One result of this variety in implementation is that the development assessed comes from a combination of reading and supplementary activities. To that extent, it becomes difficult to discern the degree that each component contributed to language development. Was learning a result of reading, a result of the activities, or a combination of the two? To gain a clearer understanding of the effects that ER and supplementary activities can have on language learning, the current study focuses on vocabulary development and compares two approaches to ER: a traditional ER-only approach, and an ER-plus approach which supplements ER with post-reading discussion carried out in small groups.

Developing vocabulary knowledge through small-group discussions

As an activity which allows learners to produce language collaboratively, small-group discussion can facilitate language development since language production compels learners to undertake full processing of the language they possess (e.g., Gass, 1988; Joe, 1998; Swain, 1985). Some benefits of language production include:

- comprehensible input generated in the form of learner feedback,
- a necessity for processing language syntactically, and
- a means to test hypotheses a learner has about the target language (Skehan, 1998).

Activities such as small-group discussion are beneficial for language learning because they provide an environment which allows learners to modify their output, leading to increased comprehensibility, complexity, and accuracy (de la Fuente, 2002; Pica, 1996). The opportunity to modify is often manifest during breakdown in communication which can occur during a discussion. When a breakdown occurs, it is usually due to perceived problems with comprehension or language production. This discursive detour, a language-related episode (LRE), necessitates negotiation of meaning by focusing learner attention away from the topic of the discussion temporarily, to the language that the learners are producing (Long & Robinson, 1998; Pica, 1996). LREs can help to resolve misunderstandings in a number of ways, including:

- identifying communication breakdowns,
- clarification questions and comprehension checks, or
- repairing a breakdown through modified output (Pica, 1996; Skehan, 1998).

These strategies are important for producing language that is conveyed precisely, coherently and appropriately, contributing to language acquisition (Swain, 1985; Swain & Lapkin, 1998). The current study hypothesizes that learners who are provided with the opportunity to engage in post-reading discussions will experience greater gains in vocabulary knowledge than those learners who read as a stand-alone activity. This hypothesis assumes that both groups of learners spend an equal amount of time on task, an issue which is addressed in more detail below.

Measuring vocabulary development through word association knowledge

The majority of ER research has tended to measure vocabulary knowledge on a dichotomous scale, that is, either correct or incorrect knowledge. This is seen in Nakanishi's (2015) meta-analysis; the

most commonly used test was the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001). This test assesses understanding of a word's meaning using a multiple-choice format. A test taker receives a mark for each correct answer they select.

Word associations, on the other hand, measure vocabulary knowledge on a scale that is not based specifically on correct language performance (Wolter, 2002). For example, the association between *ship* and *vessel* is not about being correct or incorrect, but is rather a question of degree of similarity. The relationship between *ship* and *vessel* is likely to be stronger than the relationship between *ship* and *keyboard*. Meara (2009) investigated strength of word associations using *V_links*, a computerized word association test. The test consists of 20 sets of words, each set consisting of 10 words from the first 1,000 most frequent words in English. The test-taker selects associated pairs from the 10 words, and after selecting each one, reports the strength of the association on a scale from one (weak association) to four (strong association). Meara found that this method created further issues, one issue being that non-native speakers (NNS) often selected word pairs that were not selected by native speakers (NS). Another issue Meara found was that some NNSs would report certain associations as strong, while other NNSs would report the opposite. Appropriately assigning association strength is an area still under investigation (Meara, 2009).

The current study addresses the issue of word association strength by using Latent Semantic Analysis (LSA), a statistical method used in computational linguistics for representing meaning in large corpora (Landauer, Foltz, & Laham, 1998). LSA is based on the idea that a word's meaning is the sum of all of the contexts in which it does and does not occur (Landauer & Dumais, 1997). LSA computes word-context comparisons to determine how closely the meaning of a word is related to the meaning of the context in which it occurs. For example, in a context (e.g., a paragraph) about war, the word *battle* would likely have a higher semantic overlap with the text (resulting in a higher LSA score) than the word *carpet*. Although LSA has been seen in L2 literature since the mid-1990s (e.g., Ellis, 1994), it has not been widely adopted. In one of the few studies that has used LSA, Crossley, Salsbury, McCarthy, and McNamara (2008) used it to determine whether learners' speech became more cohesive over the course of one year of language study. They recorded interviews with their participants over the course of an academic year. Using LSA, they compared each pair of adjacent sentences from each participant's interviews. The results of their study revealed that the LSA values increased as a function of time leading the authors to conclude that language learners' speech becomes increasingly cohesive as the time they spend studying the language increases. Their study provides evidence that as L2 learners develop their linguistic knowledge, they develop stronger semantic relationships in their mental lexicon. The authors note that the more experience learners have with vocabulary in context, the more likely they are to develop connections between words.

The current study compares word association knowledge development in two approaches to ER: a reading-only approach and an approach whereby reading is supplemented with post-reading discussion. The research question motivating this study is how does a reading-only ER approach compare with a reading-plus-discussion ER approach in developing L2 word association knowledge?

Method

Participants

The study took place at a New Zealand university. The participants were enrolled in a full-time, 12-week EAP programme. The programme prepares students for participation in a New Zealand

university academic community, at both undergraduate and graduate levels.

A total of 63 students participated in the study (29 female, 34 male). The students came from a variety of countries, with the majority from China (42). Smaller numbers came from Japan (5), Colombia (3), and Myanmar (2). One student each came from Iran, Iraq, Russia, Saudi Arabia, and Vietnam. Demographic information was not available for six of the participants. The participants were 22 years old on average (*SD* = 3.5 years).

At the university where the research took place, students are placed into classes based on their English proficiency. Their proficiency is determined by an in-house placement test administered at the beginning of the programme. The participants' average score on the placement test was 134 (*SD* = 12.8). The threshold separating intermediate and advanced students is 130 points out of a possible 260 on the placement test (English Proficiency Program coordinator, personal communication, December 8, 2015). Additionally, in the first week of the course, students sit the Vocabulary Levels Test (Schmitt et al., 2001) in order to give teachers an idea of the vocabulary learning requirements students have. Table 1 shows descriptive statistics for the participants, including average proficiency test score and VLT results. Mastery of a VLT level is reached with a score of 26 (Schmitt et al., 2001). Table 1 shows descriptive statistics for the participants (Standard Deviations are in parentheses). Table 1 suggests that the participants are approaching mastery of the 2000 band, and decrease in knowledge as a function of frequency band level.

Table 1. *Participants' descriptive statistics*

N	Proficiency score	Vocabulary Levels Test				
		2000	3000	Academic	5000	10000
63 (29 females, 34 males)	134 (12.8)	24 (3.6)	20.7 (4.4)	19.7 (4.7)	15 (5.4)	4.4 (3.6)

Materials

Graded Readers. All participants read the same five graded readers (GRs) over the course of the study. The GRs were chosen by the first author before the study began. Two criteria were used to choose the GRs. First, it was important to maximize the likelihood that the GRs would maintain reader interest so that the participants would enjoy reading. The Language Learner Literature Awards were consulted. The Language Learner Literature Award is given by the Extensive Reading Foundation (http://erfoundation.org/wordpress/) to books for their overall outstanding quality and likely enduring appeal. The five GRs were all award-winning books.

The second criterion used to select the GRs was difficulty level, determined by coverage rate, or the percentage of words in a text known to a learner. A 95% coverage rate was adopted (Hu & Nation, 2000; Schmitt, Jiang, & Grabe, 2011). In addition to providing adequate comprehension, a 95% coverage rate creates conditions for incidental vocabulary learning to occur (Nation, 2013). At the time the GRs were purchased the participants had not yet enrolled in the EAP programme meaning their VLT scores were not available. Instead, three years' worth of previous cohorts' VLT scores were used. The students in these previous cohorts scored 71% on average in the second thousand most frequent word band. This amounts to a vocabulary size of around 710 words for this band. The GRs were chosen to be within this level. Table 2 displays information about the GRs selected for the study, in the order that they were read. The appropriacy of these GRs will be addressed at the end of this article.

Table 2. Graded readers used in the study, in the order they were read

Title (Publisher)	Level*	Pages	Headwords	Tokens	Lexical Profile**
Jojo's Story (CUP)	High Elementary	46	800	9,556	K1: 92.21% K2: 4.69%
Dead Cold (CUP)	High Elementary	48	800	9,744	K1: 85.38% K2: 4.98%
Billy Elliot (Penguin)	Mid Intermediate	49	1,200	13,445	K1: 87.21% K2 6.56%
Land of My Childhood: Stories from South Asia (OUP)	High Intermediate	72	1,400	18,154	K1: 86.52% K2: 6.86%
A Kiss before Dying (Macmillan)	Early Upper Intermediate	86	1,600	19,816	K1: 85.15% K2: 4.96%

*Level is according to the ER Foundation Grading Scale **Profile computed using Lextutor

Target words. Knowledge of sixty target words was measured over the course of the study. The target words were selected based on a number of factors, the first of which was frequency of occurrence of the words in the GRs. Frequency of occurrence is a major determinant of vocabulary learning (e.g., Horst, Cobb, & Meara, 1998; Nation, 2013) and due to its relationship to vocabulary learning, three different frequency bands were created based on word *type* (so all unique forms of a word family were counted as different words). The first band, the *high-frequency* word band, consisted of twenty words which occurred more than 30 times in the GRs. The *mid-frequency* word band included words which occurred seven to 29 times, and in the *low-frequency* word band were words which occurred from one to six times. These frequency bands were then checked against Nation's (2004) British National Corpus frequency list, and conformed to Nation's (2013) classification of frequency bands. The *high-frequency* word band included words in the first 1,000-word frequency band. The *mid-frequency* word band included words found in the fourth to the eighth 1,000-word frequency bands. The *low-frequency* word band included words in the ninth frequency band and beyond.

The second factor considered when selecting the sixty target words was word length. The longer a word is, the more there is to be remembered, and thus the more room there is for error in remembering (Ellis & Beaton, 1993). The language used in GRs is by definition graded, meaning there were not enough words to limit the length of the words in each frequency band to one specific length. Instead, a range of lengths was set from three to nine characters, and all words which fell within this range were considered candidates for selection. In this way, extremely short or extremely long words, which may draw special attention from the participants while reading, were excluded.

The remaining factors used for target word selection were psycholinguistic traits retrieved from the Medical Research Council psycholinguistic database (http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm). First was word familiarity, or the degree that a person is exposed to the word daily (Gilhooly & Logie, 1980). The more familiar a word is, the less difficult it tends to be (Leroy & Kauchak, 2014). The second psycholinguistic factor was meaningfulness, or the degree of associational relationships linked to a word. The more meaningful a word is, the more relationships a word has, and the more avenues that are available for the word to be learned (Ellis & Beaton, 1993). The third psycholinguistic factor

taken into account was concreteness. Words referring to objects, materials or people tend to be more concrete and are typically easier to learn (Gilhooly & Logie, 1980). The final psycholinguistic factor considered was imageability, or the degree to which a word conjures imagery (Fitzpatrick & Izura, 2011). Words with higher imageability tend to be remembered more often than words with lower imageability (Atkinson & Raugh, 1975). It was impossible to select words which had exactly the same value for these four psycholinguistic traits due to the graded nature of the language used in GRs. Instead, a range of values for each of the characteristics was created and words which fell into the range were included as possible target words. The process of defining appropriate ranges was methodical, starting with as narrow a range as possible, and expanding only until enough potential words were available so that there were 20 words in each of the frequency bands. When enough words were available, those words were used. The *high-frequency* word band had a larger number of words to choose from and the words were chosen at random. Proper nouns and function words were excluded. Part of speech was not controlled for due to the limited number of words available, especially in the *mid-frequency* and *low-frequency* word bands. The *high-frequency* word band had four nouns, eight adjectives, six verbs, and two adverbs. The *mid-frequency* word band had 17 nouns, one adjective and two verbs. The *low-frequency* word band had 17 nouns, one adjective, one verb and one preposition. Appendix A shows summary statistics for each target-word frequency band. Appendix B lists detailed information for each target word.

Measures

Three measures were used to assess vocabulary knowledge development: a self-report test, a word association test, and a post-test interview. Each are explained below.

Self-report test. In the self-report test, each target word was presented by itself on a computer screen, one-by-one to each participant. Below each word the participant selected their level of familiarity with the word from the following options:

1. I have never seen this word.
2. I have seen this word, but don't know what it means.
3. I know the meaning of this word.

Words which were reported at the second and third levels were presented to the students in the word association test immediately after they reported knowledge at one of these two levels. The words which were reported as unknown were not included in the word association test.

Word association test. The current study adopted a multi-response word association format, with learners providing up to five responses for each target word they reported having knowledge of at either the form or meaning level (the second and third levels in the self-report test section above). The instructions told the participants to provide up to the first five words they thought of for each target word. There was no time limit for the test, however the learners were instructed to come up with word associations as quickly as they could. The test took approximately 35 minutes to complete.

Post-test interview. Within one week of finishing the post-test, participants completed a one-on-one interview with the first author, lasting approximately 15 minutes. Five areas were addressed during each interview. The first area involved aspects participants found (un)appealing about the treatment, and whether or not they felt the treatment was conducive to learning (Foster & Ohta, 2005). The second area addressed the appropriateness of the GRs. The third area focused on the reading process that the participants went through, with the purpose of establishing individual reading habits. The

fourth area addressed in the interview was the Say-it activity (the Say-it activity is detailed below). The purpose of this set of questions was to get a view of the supplementary activity from the participants' perspective. The fifth area was included to investigate reasons for the associations produced by the participants.

Procedure

The participants in the study were assigned to one of two groups. The first group was an ER-only group engaged solely in Extensive Reading (n = 37). The second group was an ER-plus group (n = 26). This group completed the Say-it activity (Macalister, 2014) after finishing each GR. The Say-it activity is a post-reading small-group discussion task. During the 15-minute task, learners form triads and take turns choosing prompts from a three-by-three grid for another group member to discuss. By designing the prompts in a way which promotes discussion of the characters and events in the story, learners may have the opportunity to use the vocabulary they were exposed to in the books.

To familiarize the ER-plus group with the nature of the Say-it activity, a two-minute video was created and shown to the students in the ER-plus group before the first Say-it activity. The video consisted of three L1 English speakers demonstrating the Say-it activity using the supplementary materials found in Macalister (2014). The pretask was conducted to help reduce the cognitive load on the learners (Ellis, 2003; Skehan, 1996; Willis, 1996). After watching the video, participants were assigned to a triad and remained in that triad for the entire study. Where possible, each triad was composed of one higher-proficiency student, one average proficiency student, and one lower-proficiency student based on the placement test that the participants sat at the beginning of the course. Each triad had at least one female and one male student.

While the ER-plus group was engaged in the Say-it activities, the ER-only group read a chapter from a short-story book. The short-story books selected for this purpose were borrowed from the Language Learning Centre at the university where the research took place. The short-story books that were chosen had between 800 and 1600 headwords, within the range of headwords making up the five GRs used in the study. Short-story books were used since they were comprised of stories that could be read in a short amount of time (i.e., 15 minutes). Each person in the ER-only group chose a different short story book. After choosing a book, they selected a story and read for 15 minutes.

Figure 1 illustrates the design of the study. The pretest consisted of a self-report test and a word association test. The participants began reading a few days after the pretest. The reading took place for 15 minutes in class every day. Students were given a reading schedule which had a number of chapters to read each week, amounting to approximately one chapter each weekday, and three chapters over the weekend. All of the books were completed in approximately 7 to 9 days to allow for repetition and reinforcement of new input (Nation & Wang, 1999).

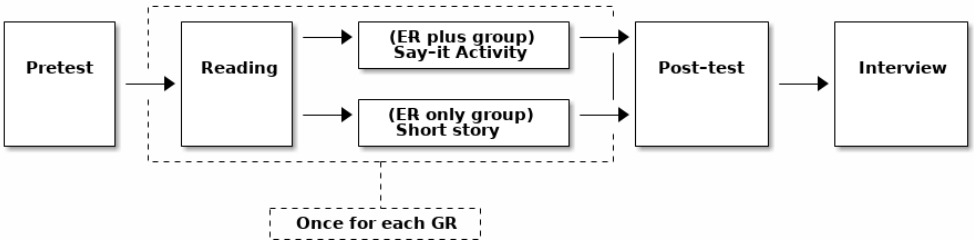

Figure 1. Research design of the study.

After the fifth and final Say-it activity, all participants sat the post-test which was the same as the pretest. The post-test took between 30 minutes and 50 minutes to complete depending on the student. When each participant finished the post-test, they scheduled a time for their one-on-one interview and left the test room. All of the interviews took place within one week of finishing the post-test. The pretest and post-test data was collected electronically using the online software Qualtrics (https://www.qualtrics.com).

Analysis

Data analysis was performed using the R software environment (version 3.5.1, R Core Team, 2018). Mixed effects models were fitted to the data using the *lme4* (version 1.1.17, Bates, Mächler, Bolker, & Walker, 2015) and the *lmerTest* (version 3.0.1, Kuznetsova, Brockhoff, & Christensen, 2017) packages in R. Models were built using backwards elimination and were compared using ANOVAs. The findings section reports the best-fitting models. Multiple comparisons were conducted using the *emmeans* package (version 1.3.0, Lenth, 2018). Omega squared (ω^2) was used to calculate effect sizes (Xu, 2003).

Self-report analysis. In order to examine whether the two participant groups had similar levels of learning potential, the total words reported in each of the three categories were compared statistically. To do so, the self-report data was first re-coded. The first level (number one in the *self-report test* section above) was recoded as "none", the second level was recoded as "form" and the third level was recoded as "meaning". Next, the target words were split into their respective frequency bands, and self-reported knowledge was totalled for each of the three self-report levels within each band. This resulted in three sets of values for each student, each set comprising one frequency band.

Word association analysis. The word association data was prepared for analysis in the following manner. Each response that the participants produced was corrected for spelling if it was incorrectly spelled but was immediately recognizable as the intended word (e.g., *pletend* instead of *pretend*) (Fitzpatrick, Playfoot, Wray & Wright, 2013; Wolter, 2002). Next, the spelling of any response which had a different spelling in American and British English was changed to the British English form (e.g., *color* was changed to *colour*) because the majority of the GRs were written in British English. The third step was to reduce all multi-word responses to single-word responses. This entailed omitting all words except for the headword of the response (Laufer & Nation, 1995; Wolter, 2002). For example, if a student's response to the prompt *move* was *to go*, the *to* was omitted. This resulted in the deletion of determiners, function words, *not*, *opposite of*, and in some cases prepositions (Fitzpatrick et al., 2013). The final step was to change all responses to their lemma form. Changing the responses to their respective lemmas reduces statistical noise by merging semantically similar word forms (Kantrowitz, Mohit, & Mittal, 2000; Lifchitz, Jhean-Larose, & Denhiere, 2009). Leech, Rayson, and Wilson's (2001) word lemma list was used to lemmatize the responses (available at http://ucrel.lancs.ac.uk/bncfreq/lists/1_1_all_alpha.txt) and is under the Creative Commons Attribution-Share Alike 2.0 UK: England & Wales License. Next, electronic copies of the GRs, retrieved with permission from their publishers, were prepared using the same method used for formatting the learners' word association responses.

LSA determines the strength of relationships between words in text by creating a weighted word-by-document matrix, known as a semantic space. In the current study, a semantic space was created using the combined text from the five GRs. Each row in this space represented one lemma occurring in the GRs, and each column represented one paragraph of text from the GRs. The paragraphs were kept in the order that they were read. For example, the first column corresponded

to the first paragraph of the first GR. Similarly, the last column corresponded to the last paragraph of the final GR. Paragraphs were chosen to represent each column in this space since they tend to consist of one self-contained idea (Lifchitz et al., 2009). In this lemma-by-paragraph matrix, each cell contains a number representing the frequency of occurrence of a lemma in a document. This number is weighted to reflect its importance in the semantic space. These weights emphasize words which are unique in a paragraph and de-emphasizes common words (Landauer & Dumais, 1997). Both of the weight functions were part of the R package *lsa* (Wild, 2015). The first weight, applied to each column of the matrix, was the inverse paragraph frequency and was applied using the *gw_idf* function in R. Paragraph frequency refers to the number of paragraphs in which a word occurs. The second weight, applied to each row of the matrix, was the logarithm of the lemma's frequency in each paragraph. This weighting was assigned using the *lw_logtf* function in R. The next step in creating the semantic space was to identify and remove stopwords, or words which occur with extremely high frequency, and carry little meaning (Wild, 2015). Following Lifchitz et al. (2009), the following words were removed from the semantic space due to their low global weighting: *a, and, are, at, be, but, do, for, it, no, not, so, the,* and *to*. An updated lemma-by-document matrix was created without the stopwords, using the aforementioned process. This updated matrix consisted of 2,440 rows and 1,108 columns. Each row represented one lemma found in the GRs. Each column represented one paragraph from the GRs. Each cell contained the weighted frequency value for the corresponding lemma in the respective paragraph.

The resulting lemma-by-paragraph matrix is sparse; the majority of the cells will not contain a value. This is because only a limited number of words are present in each paragraph. Comparing words using this sparse matrix is not ideal (Landauer, McNamara, Dennis, & Kintsch, 2007). To alleviate this issue, the final step in creating the semantic space was to apply singular value decomposition to the matrix. Singular value decomposition is a statistical method which is used to reduce the number of elements (or dimensions) in the matrix. By doing so the matrix is reduced and reconstructed producing a semantic space with non-zero values in all cells.

The cosine similarity measure was used to determine the degree of similarity between each target word and the set of associations participants produced. The similarity in meaning of two words is measured as the cosine angle between two vectors (i.e., rows in the semantic space). The cosine similarity measure determines the extent that vectors are pointed in the same direction in a multi-dimensional space. The exact mathematics used to derive the cosine similarity value is outside the scope of this article (but see Landauer et al., 2007), however the value can range from -1 to 1, with larger values indicating a greater degree of similarity and lower values indicating a lesser degree of similarity. Cosine similarity was computed between each target word and the set of associations produced by a participant for that word. This was done twice, once using the associations learners produced on the pre-test, and again for the associations they produced on the post-test. To determine the degree of change, pre-test values were subtracted from post-test values. A positive difference refers to an increase in similarity from the pretest to the post-test, while a negative difference refers to a decrease in similarity. The difference in cosine similarity values was used as the dependent variable in a linear mixed effects model to compare the two ER groups. Responses given by participants which were not in the semantic space were omitted from analysis because there is no way to calculate a similarity value for words which do not occur in the semantic space.

For example, student W was a participant in the ER-plus group. She reported having form knowledge of the mid-frequency word *kite* on the pretest. She provided two associations: *food* and *name*. The cosine similarity value for these two associations and kite was .012. On the post-test, W reported meaning knowledge of kite and provided five associations: *sky, fly, story, children,* and *competition*. These values had a cosine similarity value of .255. The difference between her pretest

and post-test values represents an increase in semantic knowledge for the word *kite*. The word *kite* appears 16 times in the fourth GR, *Land of My Childhood: Stories from South Asia*. It does not appear in any other GR used in the study. *Kite* occurs in a context about an annual kite festival in India. During the festival children fly kites in the sky, fighting other kites until there is only one kite left in the air. The cosine similarity values suggest that W's associations given on the post-test are more similar to *kite* than the values she provided on the pretest.

The Say-it activity analysis. The Say-it activity was completed in triads after the participants finished each GR. Nine triads participated, resulting in a total of 45 Say-it activities. However, three students were absent for the first Say-it activity. In addition, technical difficulties with one of the audio recorders during the second Say-it activity resulted in loss of data from one group. Accordingly, data from 43 Say-it activities was available for analysis, amounting to nine hours and forty-five minutes of transcription data.

After the Say-it activities were transcribed, LREs were located and categorized as either lexical or grammatical. An LRE was considered lexical if it centered on word meaning, spelling, or pronunciation. An LRE was considered grammatical if it centered on other linguistic phenomena (e.g., word tense). These LREs occurred from the point in a triad's dialogue when attention shifted from the discussion of a Say-it activity prompt to the language being produced by the triad. Similarly, the episode finished when the triad's focus returned to discussion of the Say-it activity discussion prompt. This shift in attention to the language being produced was the result of either a participant producing a linguistic error, asking about something that was said by another participant in the triad, or asking about how to say something in English (Loewen, 2005). To ensure that the LREs had been accurately located and categorized, three post-graduate student raters independently rated the same 10% of the LREs in two stages. In the first stage, the raters were told to locate all LREs in the transcript. The researcher (the first author of this article) explained what an LRE was before the raters were asked to do this. When they had finished locating the LREs, the researcher discussed with the raters what they located and all inconsistencies (e.g., LREs that were not located) were resolved. In the second stage, the raters were asked to categorize each of the LREs as either lexical or grammatical. They categorized the LREs in the same transcript excerpt they used for locating the LREs. Rater reliability was computed for the combination of stages one and two, that is, location and categorization of the LREs. This was determined using Maxwell's RE and the value was 0.72. After rating, the raters met with the first author to resolve all inconsistencies. The researcher located and classified the remaining LREs.

Results

The research question posed in the current study seeks to investigate the extent to which post-reading discussion facilitates vocabulary development as measured by word association knowledge. In order to address this question, it was first necessary to ensure the two groups were comparable. This was determined using the two groups' proficiency test results. The two ER groups' proficiency test scores were compared statistically to determine whether the groups were comparable. A Shapiro-Wilk test revealed a normal distribution of the proficiency scores ($W = 0.96$, $p = 0.08$). The test results were analyzed using a two-sample *t* test. The results confirmed that the two groups were of similar proficiency $t(56) = 1.77$, p-value = .082, CI = [-0.73, 11.79].

On the pretest the ER-only group provided a total of 4,243 word association responses. Of these responses 1,126 (27%) were not in the semantic space and were omitted from analysis. The remaining 3,117 responses consisted of 765 unique words. The ER-plus produced 2,948 total

responses on the pretest and 778 (26%) responses were not in the semantic space and were omitted from analysis. The remaining 2,170 responses consisted of 675 unique words. On the post-test, the ER-only group produced a total of 4,514 responses, 1,258 (28%) of which were omitted. The remaining 3,256 responses consisted of 752 unique words. The ER-plus group produced a total of 2,968 responses on the post-test, and 887 (30%) were omitted due to not occurring in the semantic space. The remaining 2,081 responses included 642 unique words.

To determine if the two ER groups' word association knowledge was comparable at the beginning of the study, a mixed-effects model was fitted to the pretest data. The dependent variable was the cosine similarity value; the independent variables were group (ER-only, ER-plus) and target-word frequency band (high-frequency, mid-frequency, low-frequency). Student and target word were included as random effects in the model. The results revealed no significant difference between the groups (t = -0.23, p = .82), and no significant interactions were present in the model. Multiple comparisons of the means for the frequency bands revealed that the participants had greater semantic knowledge of the high-frequency words compared to the mid-frequency words (t(50.43) = 4.498, p < .001) as well as the low-frequency words (t(61.32) = 4.788, p < .001). Knowledge of the mid-frequency and low-frequency bands were not statistically different (t(63.4) = .535, p = .85).

As mentioned in the word association analysis section, the development of word association knowledge was measured using the difference in similarity values from the pretest to the post-test for each target word. To determine the extent that the two groups' respective reading treatments facilitated vocabulary knowledge development, the difference in values was compared statistically using mixed effects modeling. The dependent variable in the model was the difference in cosine similarity values. The independent variables were ER group (ER-only, ER-plus) and frequency band (high-frequency, mid-frequency, low-frequency). Student was included as a random effect in the model. The results are shown in Table 3 and reveal an interaction effect between ER group and target word frequency band, however multiple comparisons of the means showed no statistically significant effect. This means that both groups made similar gains in word association knowledge.

Table 3. *Analysis of the pre- and post-associations by treatment for 60 target words: fixed effects*

Parameter	Estimate	Std. Error	df	*t*	*p*
(Intercept)	.002	.006	67.53	.274	.785
Group = ER-plus	-.012	.008	155.4	-1.546	.124
Frequency band = mid-frequency	.017	.008	71.98	2.19	.032
Frequency band = low-frequency	.007	.010	209	.671	.503
Group = ER-plus x Frequency band = mid-frequency	.019	.010	2377	1.927	.054
Group = ER-plus x Frequency band = low-frequency	.030	.015	2465	1.997	.046

Intercept levels: Group=ER-only, Frequency band=high-frequency

If the post-reading discussions were responsible for facilitating vocabulary development, this should be most prevalent in those words which received attention in an LRE. A total of 769 LREs were located during the Say-it activities: 455 (59%) were lexical and 314 (41%) were grammatical LREs. There were ten target words that were the focus of a lexical LRE. Five of these words were *high-frequency* words: *give*, *near*, *need*, *quiet*, and *read*. The remaining five were *mid-frequency* words: *boxing*, *flick*, *kite*, *lorry*, and *shaft*. A mixed effects model was fitted to the data from these ten

target words. The dependent variable was the difference in cosine similarity value for each of these ten target words from the pretest to the post-test. The independent variables were ER group (ER-only, ER-plus) and frequency band (*high-frequency*, *mid-frequency*). Participant and target word were included as random effects in the model. The results of the model are shown in Table 4 and reveal a significant interaction between group and frequency band. Multiple comparisons of the means revealed that the ER-plus group scored significantly higher in the *mid-frequency* word band than the ER-only group ($t = 2.44$, $p = 0.02$, $\omega^2 = .08$). In other words, the ER-plus group made significantly greater gains compared to the ER-only group in word association knowledge of the mid-frequency words which were focused on in an LRE.

Table 4. *Analysis of the pre- and post-associations by treatment for the 10 target words focused on during LREs: fixed effects*

Parameter	Estimate	Std. Error	df	*t*	*p*
(Intercept)	-0.006	0.01	12.014	-0.627	0.543
Group = ER-plus	-0.014	0.014	481.177	-1.013	0.311
Frequency band = mid-frequency	0.040	0.016	18.049	2.528	0.021
Group = ER-plus x Frequency band = mid-frequency	0.059	0.023	484.879	2.556	0.011

Intercept levels: Group=ER-only, Frequency band=high-frequency

Discussion

The expectation was that the Say-it activity would promote opportunities for language learning in the form of LREs. Two examples are presented below and discussed in detail in order to explore how language learning may have been facilitated through the LREs. The examples that follow are used to provide possible explanations for the quantitative findings detailed in the previous section. The first example, shown in (1) below, depicts an LRE focussing on the target word *boxing.* This LRE occurred during the third Say-it activity after the participants read *Billy Elliot*, a story about a boy who wants to learn ballet but whose father insists that he take up boxing. In the story, *boxing* occurred 27 times and *Billy Elliot* is the only GR in which *boxing* occurred. The initial context of *boxing* is as follows:

> *He came up behind me and closed the piano suddenly. He nearly broke my fingers. Then he ran out of the doors after Tony. "I will see you later at the club" he said on the way out. Oh no I thought. Today I am boxing. I hate it when he watches me.*

Example (1) below involves two students: S and J. S's proficiency test score was 123, and J's proficiency test score was 124. During the Say-it activity in (1), S and J discuss how Billy felt after his first ballet lesson. S explains how Billy felt wonderful because he really enjoys ballet. During S's explanation of how Billy felt, S realizes that she does not know how to say *boxing* in English. She asks J in Mandarin (S and J are both L1 speakers of Mandarin Chinese) how to say *boxing* and J responds with the word form in English. In S's second turn, she reformulates her initial utterance to include the word *boxing*.

(1) LRE excerpt for the target word *boxing*.
S: Mm, I feel life is wonderful because I achieve, I have achieved my uh dream. And uh, I don't need to, uh, [asks in L1 Chinese how to say boxing in English].
J: boxing
S: Uh I don't need uh, I don't need take the box, boxing lessons. I think it is wonderful because I'm really interested in the ballet lessons.

On both the pretest and post-test, S reported meaning knowledge of *boxing*. On the pretest, she provided two associations: *tidy* and *mess*. On the post-test, she provided three associations: *sport*, *race*, and *competition*. These results show a possible change in understanding of *boxing*, from that of a container to the physical sport. Coupled with the LRE excerpt in (1), it appears that S did not learn that *boxing* can refer to a sport from the reading alone. Instead the data suggests that the Say-it provided the opportunity for S to further develop her vocabulary knowledge. Unfortunately this possibility could not be explored further as S was not available for a post-test interview.

The data suggests a slightly different picture for J. On the pretest, J reported form knowledge of *boxing* and provided one association: *pack*. On the post-test, J reported meaning knowledge for *boxing* and provided five associations: *fighting*, *race*, *members*, *strong*, and *muscle*. That her understanding of *boxing* changed is evident in the LRE when she produces the word; she would not be able to translate a word which she does not know. This suggests that her understanding of *boxing* changed as a result of the reading, meaning that reading provided enough scaffolding for her to learn a new meaning for *boxing*. When asked about reading during her interview, J mentioned that since "the vocabulary appears again and again so when I read after final stories it is better".

The LRE in (1) reveals that by including opportunities for learners to engage with each other, a greater number of students are given the chance to reinforce their developing vocabulary knowledge. In the case of S, this meant retrieving the form of a word already known in their L1. For J, this meant producing a newly-acquired vocabulary item to assist her group member.

In another example, excerpt (2) below, a lexical LRE is shown which took place during the first Say-it activity, after reading *Jojo's Story*, a story about a young boy's experience with war. This LRE occurred in the same group as in (1) above. The focus of this LRE is on the target word *lorry*, which occurred eight times in *Jojo's Story* and is the only book in which *lorry* occurred. The initial context that *lorry* occurred in is as follows:

> *There is a sound outside the stable. There is something there, something bigger than a mouse. I do not know what it is and now I can hear another sound. A bigger sound, like a lorry. It is a lorry. A lorry is coming here to the village.*

The LRE begins when T (who scored 127 on the proficiency test) reads a Say-it activity prompt to S containing *lorry* which S is unfamiliar with. S asks for clarification of *lorry* to which J provides a definition. Finally, S acknowledges this assistance and begins addressing the prompt.

(2) Excerpt of an LRE for the word *lorry*
T: You are Jojo. You hear a lorry. What are you thinking?
S: Mmm, lorry?
J: Lorry uh a big car.
S: Oh oh, ok. Firstly, I also really scared and, because I afraid that the man come come to the village again and kill me. So, I I very scared for looking someone's change in the village. Uh, yeah.

This LRE provides an interesting perspective on learning which may have occurred. From S's point of view, reading by itself was not enough to gain an understanding of *lorry*, and this is evident in S's clarification request. The Say-it activity provided an opportunity for S to clarify the meaning of an unknown word by asking for peer assistance, and this opportunity may have facilitated development. On the pretest, S reported no knowledge of *lorry*, and on the post-test she reported meaning knowledge, providing two associations: *car* and *truck*.

While the data suggests that S benefited from the LRE, J's data points to a different source of learning. During the LRE presented in (2), J provided S with a definition of *lorry*. Since J reported no knowledge for *lorry* on the pretest, and was able to provide a definition during the LRE in (2), the data suggests that reading by itself was enough for J to acquire knowledge of *lorry*. During her interview, seen in (3) below, J mentions that she did learn *lorry* from *Jojo's Story*. She mentioned that she did not know the word before reading. She is then asked in the interview what *lorry* means and provides the synonym *big truck.* Later in her interview, she discussed using contextual clues to assist her in guessing the meaning of *lorry*. When she came across the word in *Jojo's Story* she guessed that it meant "...something like a car, a big car, or something like that". To confirm this guess, J mentioned she looked the word up in a dictionary. For her, it seems that reading may have been sufficient for grasping an understanding of an unknown word, and with the dictionary look-up she was able to confirm her developing knowledge. On the post-test, J reported meaning knowledge of *lorry* and provided three associations: *truck*, *big*, and *high*.

(3) Interview excerpt of J discussing the word *lorry*
J: ... and this book [Jojo's Story] is easy, but I learn a word from this. lorry
TT: lorry
J: lorry. I don't know that before.
TT: what does it mean?
J: mmm, big truck
TT: big…?
J: truck. lorry, l-o-r-r-y?
TT: yeah a big what?
J: big truck

The third person in the triad, T, sat as a silent observer during this LRE. While there is no evidence of T's developing knowledge in this LRE, during his interview T mentioned that the vocabulary in *Jojo's Story* "was simple so it was easy to read without a dictionary". Interestingly, on both the pretest and post-test T reported meaning knowledge for *lorry. H*e supplied one association on both tests: *name*. It is interesting because despite exposure to *lorry* in *Jojo's Story*, silently observing the LRE in (2) above, and the ease at which he was able to read the story, T shows no change in his knowledge of *lorry* and indeed no learning from either the reading or the exposure to the LRE. During his interview T mentioned that he was not interested in *Jojo's Story*, although did not provide a reason why. Assuming that T believed *lorry* was a proper noun as his word association responses suggest, the fact that he was disinterested in *Jojo's Story* could have meant that he was not fully engaged during the Say-it activity.

These two examples highlight facilitating effects that post-reading discussion can have on semantic knowledge development. In both examples, more students showed evidence of developing knowledge than would have if the Say-it activity had not been implemented. This suggests that supplementing ER with post-reading discussion increases the opportunities that learners have to develop their vocabulary knowledge. Additionally, the two examples above provide evidence

suggesting that this ER-plus approach allows learners to scaffold knowledge from their peers, which can lead to greater development of knowledge than utilizing ER as a stand-alone activity.

In the current study, *high-frequency* and *mid-frequency* words were dicussed in LREs, while *low-frequency* words were not. As mentioned earlier, the Say-it activity prompts were designed to recall and discuss events in the story. To that extent, specific target words were not targeted directly when designing the prompts. This meant that the words which were more likely to be used during the Say-it activity were in part determined by the events which occurred in the GRs. Even though the prompts were designed in a way to promote discussion of the events in the story, this did not necessarily mean that the students would engage with the words, as was the case with T in (2) above. The *mid-frequency* words focussed on in the LREs above relate to important events in the story which made them more memorable, or more likely to arise during discussion. For example, at the beginning of *Jojo's Story*, the sound of the lorry coming towards Jojo's village is the first event which sets the plot in motion. Likewise, boxing is a central theme in *Billy Elliot* and it would be difficult to discuss the story without discussing boxing. On the other hand, the high-frequency words such as *give* and *need* or the low-frequency words *turban* and *whoosh* may have been less central to the events in the stories, and the participants may not have found it necessary to use them during the Say-it activities.

Limitations

There are potential limitations that should be acknowledged. One issue regards the extent that the participants were engaged in 'Extensive Reading'. ER was defined as the fluent comprehension of text in large quantities. Fluent comprehension of text is possible when reading is carried out at a 95% coverage rate or higher (Nation, 2013). In the current study, it was not possible to test the participants before the materials were purchased to determine which GR level would allow for a 95% coverage rate. As the VLT results in Table 1 reveal, the participants did not have mastery of the 2,000 most frequent words. As a result, it is possible that even the lowest-level GR was not read at a 95% coverage rate (refer to the rightmost column of Table 2 for the lexical profile of each GR). Regarding the second part of the definition of ER - reading large quantities of text - there is currently no amount of reading which can be considered 'large'. Referring back to Table 2, the five graded readers the participants read amounted to 70,715 tokens. Although this is comparable to previous research (e.g., Sakurai, 2015; Taguchi et al., 2004) further research should investigate larger amounts of reading.

The second limitation relates to the relationship between time on task and number of exposures to the target words. The amount of time that the learners spent on their respective interventions was controlled. However it was not possible to control the number of additional exposures to the target words during the short-story reading in the ER-only group, nor during the Say-it activity for the ER-plus group. As a result, the probability that both groups would be exposed to the TWs may have differed. It is difficult to determine the extent that this affected the results and future research would benefit from addressing this limitation.

Another issue to mention is that the development found in the current study, even if small, represents only a portion of the potential learning which occurred. The 60 target words in the study were a small amount of the words which the groups were exposed to during the intervention. This means that it is possible additional learning occurred which was untracked in the study.

In addition, the semantic space created for the current study relied on paragraphs from the GRs. This may be considered problematic, since, due to GRs being fiction, their paragraphs may not be focused on one central idea. In some cases, a paragraph was part of a dialogue which took place and as a result was very short, sometimes one or two sentences. While research suggests that paragraphs tend to be the most appropriate length of text to use with LSA (Lifchitz et al., 2009), this may not be the case with fiction writing. Further research should investigate different genres of GRs.

Finally, the lack of a delayed post-test means it was not possible to measure longer-lasting effects of the two interventions. Future research could include a delayed post-test to investigate the extent that vocabulary knowledge was retained over a longer period of time.

Conclusion

This study has provided evidence (Table 3) supporting both a reading-only approach and a reading-plus-activity approach to ER, in terms of their facilitative effects on the development of word association knowledge. To that extent the current study reinforces Macalister (2008) who also found a reading-only approach successful. The results from the current study also support an approach to ER which supplements reading with post-reading activities such as the Say-it activity. The ER-plus group made greater gains than the ER-only group in their knowledge of mid-frequency target words (Table 4). This finding highlights the importance of learner-learner interaction, specifically the role that scaffolding of peer knowledge can play in the development of vocabulary knowledge. Supplementing ER with additional activities such as the Say-it activity can also help to mitigate the "monastic detention session" that Green (2005) has urged practitioners to avoid. As Macalister (2008) notes, the implementation of ER will vary from classroom to classroom. What is important is that ER has a place in all language-learning classrooms.

References

Al-Homoud, F., & Schmitt, N. (2009). Extensive reading in a challenging environment: A comparison of extensive and intensive reading approaches in Saudi Arabia. *Language Teaching Research, 13,* 383–401. doi:10.1177/1362168809341508

Atkinson, R. C., & Raugh, M. R. (1975). An application of the mnemonic keyword method to the acquisition of a Russian vocabulary. *Journal of Experimental Psychology: Human Learning and Memory, 1,* 126–133. doi:10.1037/0278-7393.1.2.126

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi:10.18637/jss.v067.i01

Beglar, D., Hunt, A., & Kite, Y. (2012). The effect of pleasure reading on Japanese university EFL learners' reading rates. *Language Learning, 62,* 665–703. doi:10.1111/j.1467-9922.2011.00651.x

Bruton, A. (2002). Extensive Reading is reading extensively, surely? *The Language Teacher, 26*(11). Retrieved from http://jalt-publications.org/old_tlt/articles/2002/11/bruton.

Cobb, T. Web VP Classic v.4 [computer programme]. Retrieved from https://www.lextutor.ca/vp/eng/.

Crossley, S. A., Salsbury, T., McCarthy, P., & McNamara, D. S. (2008). Using latent semantic analysis to explore second language lexical development. *Proceedings of the 21st International Florida Artificial Intelligence Research Society* Conference, pp. 136–141.

Day, R. R. (2015). Extending extensive reading. *Reading in a Foreign Language, 27,* 294–301.

Day, R. R., & Bamford, J. (2002). Top ten principles for teaching extensive reading. *Reading in a*

Foreign Language, 14, 136–141.
de la Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary. *Studies in Second Language Acquisition, 24*, 81–112.
Ellis, R. (1994). *The study of second language acquisition*. Oxford Applied Linguistics. Oxford, UK: Oxford University Press.
Ellis, R. (2003). *Task-based language learning and teaching*. Oxford Applied Linguistics. Oxford, UK: Oxford University Press.
Ellis, N., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning, 43*, 559–617. doi:10.1111/j.1467-1770.1993.tb00627.x
Fitzpatrick, T., & Izura, C. (2011). Word association in L1 and L2. *Studies in Second Language Acquisition, 33*, 373–398. doi:10.1017/S0272263111000027
Fitzpatrick, T., Playfoot, D., Wray, A., & Wright, M. J. (2013). Establishing the reliability of word association data for investigating individual and group differences. *Applied Linguistics, 13*, 1–29. doi:10.1093/applin/amt020
Foster, P., & Ohta, A. S. (2005). Negotiation for meaning and peer assistance in second language classrooms. *Applied Linguistics, 26*, 402–430. doi:10.1093/applin/ami014
Gass, S. (1988). Integrating research areas: A framework for second language studies. *Applied Linguistics, 9*, 198–217. doi:10.1093/applin/9.2.198
Gilhooly, K., J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation, 12*, 395–427. doi:10.3758/BF03201693
Green, C. (2005). Integrating extensive reading in the task-based curriculum. *ELT Journal, 59*, 306–311. doi:10.1093/elt/cci059
Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *The Canadian Modern Language Review, 61*, 355–382. doi:10.3138/cmlr.61.3.355
Horst, M., Cobb, T., & Meara, P. (1998). Beyond a Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*, 207–223.
Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*, 403–430.
Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. *Reading in a Foreign Language, 17*, 23–59.
Jeon, E.-Y., & Day, R. R. (2015). The effectiveness of core ER principles. *Reading in a Foreign Language, 27*, 302–307.
Joe, A. (1998). What effect do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics, 19*, 357–377. doi:10.1093/applin/19.3.357
Kantrowitz, M., Mohit, B., & Mittal, V. (Eds.). (2000). Stemming and its effects on TFIDF ranking, Athens, Greece. Conference on research and development in information retrieval. doi:10.1145/345508.345650
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1-26. doi: 10.18637/jss.v082.i13
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284. doi:10.1080/01638539809545028
Landauer, T. K., McNamara, D., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahawah, NJ: Lawrence Erlbaum Associates.
Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*, 307–322. doi:10.1093/appling/16.3.307

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English based on the British National Corpus*. Harlow: Pearson Education Limited.

Lenth, R. (2018). emmeans: Estimated Marginal Means, aka Least-Squares Means (version 1.3.0). Retrieved from https://CRAN.R-project.org/package=emmeans.

Leroy, G., & Kauchak, D. (2014). The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association, 21*, 169–172. doi:10.1136/amiajnl-2013-002172

Lifchitz, A., Jhean-Larose, S., & Denhiere, G. (2009). Effect of tuned parameters on a LSA multiple choice questions model. *Behavior Research Methods, Psychonomic Society, Inc., 41*, 1201–1209. doi:10.3758/BRM.41.4.1201

Loewen, S. (2005). Incidental focus on form and second language learning. *Studies in Second Language Acquisition, 27*, 361–386. doi:10.1017/S0272263105050163

Long, M. H., & Robinson, P. (1998). Focus on form: Theory, research, and practice. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 15-41). Cambridge: Cambridge University Press.

Macalister, J. (2008). Implementing extensive reading in an EAP programme. *ELT Journal, 62*, 248–256. doi:10.1093/elt/ccm021

Macalister, J. (2014). The say-it activity. *Modern English Teacher, 23*(1), 29–32.

Macalister, J. (2015). Guidelines or commandments? Reconsidering core principles in extensive reading. *Reading in a Foreign Language*, 27, 122–128.

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

Nakanishi, T. (2015). A meta-analysis of extensive reading research. *TESOL Quarterly, 49*, 6–37. doi:10.1002/tesq.157

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 3–13). Amsterdam, the Netherlands: John Benjamins.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). New York, NY: Cambridge University Press.

Nation, I. S. P., & Wang, K. (1999). Graded readers and vocabulary. *Reading in a Foreign Language 12*, 355–380.

Pica, T. (1996). The essential role of negotiation in the communicative classroom. *JALT Journal, 18*, 241–268.

R Core Team. (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org

Sakurai, N. (2015). The influence of translation on reading amount, proficiency, and speed in extensive reading. *Reading in a Foreign Language, 27*, 96–112.

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

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

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*, 38–62.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal, 82*, 320–337. doi:10.1111/j.1540-4781.1998.tb01209.x

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–256). New York, NY: Newbury House Press.

Taguchi, E., Takayasu-Maass, M., & Gorsuch, G. J. (2004). Developing reading fluency in EFL: How assisted repeated reading and extensive reading affect fluency development. *Reading in a Foreign Language, 16*, 70–96.

Waring, R., & McLean, S. (2015). Exploration of the core and variable dimensions of extensive reading research and pedagogy. *Reading in a Foreign Language, 27*, 160–167.

Wild, F. (2015). LSA: Latent Semantic Analysis. R package version 0.73.1. Retrieved from https://CRAN.R-project.org/package=lsa

Willis, J. (1996). *A framework for task-based learning*. New York, NY: Longman.

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

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in Medicine, 22*, 3527–3541. doi:10.1002/sim/1572

Yamashita, J. (2008). Extensive reading and development of different aspects of L2 proficiency. *System, 36*, 661–672. doi:10.1016/j.system.2008.04.003

Appendix A

The table below displays means and standard deviations (in parentheses) for the target word characteristics according to frequency band. A hyphen indicates that there were no entries in the MRC database. The mid-frequency words' means and standard deviations for familiarity, concreteness, imageability, and meaningfulness were computed using ten words instead of 20 because ten of the words did not have entries in the MRC database. Only one low-frequency word had values in the MRC database and as such the value was omitted from the table below.

	High-frequency words	Mid-frequency words	Low-frequency words
Frequency in GRs	59 (35)	14 (7)	3 (2)
Range in GRs	5 (0)	1(0)	1 (1)
Length in characters	4 (1)	6 (1)	6 (1)
BNC frequency band	1 (0)	5 (1)	11 (3)
Familiarity	576 (33)	240 (7)	-
Concreteness	388 (63)	272 (288)	-
Imageability	445 (69)	273 (287)	-
Meaningfulness	462 (29)	179 (210)	-

Appendix B

Characteristics for the 60 target words measured in the study, ordered alphabetically within each frequency band. A hyphen indicates that there was no entry in the MRC database.

Word	Band	Frequency	Range	BNC	Familiarity	Concreteness	Imageability	Meaningfulness
back	high	127	5	1	587	540	483	418
dead	high	57	5	1	581	429	520	497
down	high	137	5	1	546	339	459	444
fell	high	31	5	1	546	407	431	403
give	high	50	5	1	595	326	383	465
hard	high	40	5	1	595	425	460	497
long	high	59	5	1	579	381	471	492
name	high	50	5	1	573	405	475	474
near	high	41	5	1	582	337	408	465
need	high	38	5	1	589	314	327	473
quiet	high	36	5	1	577	389	426	451
read	high	45	5	1	568	420	499	467
right	high	53	5	1	599	361	372	413
round	high	35	5	1	563	438	559	489
stop	high	53	5	1	563	308	452	485
stupid	high	34	5	1	550	351	381	487
tell	high	147	5	1	596	306	350	465
thing	high	37	5	1	587	350	358	479
well	high	68	5	1	550	467	522	418
white	high	43	5	1	590	472	566	464
audition	mid	28	1	6	479	370	395	397
blouse	mid	8	2	5	562	640	595	530
boxing	mid	27	1	4	-	-	-	-
campus	mid	13	1	4	-	-	-	-
closet	mid	7	1	6	540	599	525	415
copper	mid	15	1	4	491	547	548	350
cupboard	mid	7	3	5	-	-	-	-
flick	mid	19	1	4	-	-	-	-
gloves	mid	18	2	4	-	-	-	-
goat	mid	7	2	4	469	636	585	402

handsome	mid	28	1	4	-	-	-	-
jeep	mid	7	1	6	564	622	659	477
kite	mid	16	1	6	481	592	624	408
lorry	mid	8	1	8	198	420	383	236
papa	mid	13	1	5	-	-	-	-
pharmacy	mid	15	1	6	-	-	-	-
picket	mid	8	1	6	-	-	-	-
shaft	mid	17	1	4	-	-	-	-
trailer	mid	9	1	5	528	597	587	363
vacation	mid	11	1	5	495	414	559	-
alsatian	low	1	1	13	-	-	-	-
arsenic	low	6	1	9	-	-	-	-
babu	low	2	1	22	-	-	-	-
carapace	low	5	1	12	-	-	-	-
clucking	low	1	1	11	-	-	-	-
envious	low	1	1	9	470	-	361	-
felicity	low	2	1	12	-	-	-	-
gelatin	low	4	1	12	-	-	-	-
hibiscus	low	1	1	13	-	-	-	-
hopper	low	1	1	9	-	-	-	-
jasmine	low	3	1	9	-	-	-	-
prawns	low	4	1	9	-	-	-	-
raffle	low	4	1	9	-	-	-	-
rupees	low	5	1	10	-	-	-	-
sahib	low	5	1	15	-	-	-	-
sans	low	1	1	10	-	-	-	-
saris	low	2	1	11	-	-	-	-
turban	low	2	1	10	-	-	-	-
vats	low	1	1	12	-	-	-	-
whoosh	low	1	1	10	-	-	-	-

About the Authors

TJ Boutorwick is a Learning Adviser at Student Learning, Victoria University of Wellington. Dr. Boutorwick is interested in extensive reading and second language vocabulary development with a focus on word associations. Email: tj.boutorwick@vuw.ac.nz

John Macalister is Professor of Applied Linguistics at Victoria University of Wellington, New Zealand. His research & teaching interests include language learning and teaching, language teacher education, and language curriculum design. Email: john.macalister@vuw.ac.nz

Irina Elgort is Senior Lecturer in Higher Education at Victoria University of Wellington. Dr. Elgort's research interests include approaches and factors affecting contextual word learning from reading, measures of L2 word knowledge and processing, and computer-supported text analysis. Her work has been published in *Applied Linguistics, Language Learning, Studies in Second Language Acquisition, and Language Learning and Technology*. Email: irina.elgort@vuw.ac.nz