Vocabulary development through reading: A comparison of approaches

TJ Boutorwick

A thesis submitted for the degree of
Doctor of Philosophy in Applied Linguistics

Victoria University of Wellington, New Zealand

2017

Abstract

This thesis compares two approaches to extensive reading to determine the extent that they facilitate vocabulary development. The first approach is a traditional reading-only approach, and the second approach is a task-based approach which supplements reading with post-reading meaning-focused discussions. These two approaches are compared using a battery of tests, most notably a measure for productive knowledge of word associations.

For years, scholars have believed that word associations have potential to reveal important information about a person's language proficiency. One reason word associations are intriguing is that a large amount of a person's lexicon can be assessed (Meara, 2009). This is possible because a large amount of data from the learner can be gathered in a short period of time. Another intriguing aspect of word association data is that it is one aspect of vocabulary knowledge that is not based on correct performance. This raises the question of an appropriate means of assigning value to the associations, a question which still hinders research to this day. Recent research has made progress in this area with a multi-level taxonomy (i.e., Fitzpatrick, 2007), creating a picture of the types of associations which exist in a learner's lexicon. However, this taxonomy does not address the *strength* of the association. Wilks and Meara (2007) have attempted to tackle association strength through the use of self-report measures, whereby a test-taker reports strength of association on a four-point scale from weak to strong. This has left them with "...problems which we have not yet solved, notably a tendency for some test takers to claim that most associations are strong, while others appear to be very reluctant to identify strong associations..." (Meara, 2009, p. 80). In other words, the question of how to appropriately determine association *strength* is still unanswered.

In the current study lexical development, in the form of word association knowledge, was measured using a multi-response word association test. Participants were assessed on their knowledge of 60 target words which occurred in five graded readers that they read over the course of the study. The learners first self-reported their knowledge of

the 60 target words in terms of *no* knowledge, *form* knowledge, or *meaning* knowledge. The students provided up to five associations for each word that they reported at either the *form* or *meaning* levels. They did this once before reading the five graded readers, and again after finishing the graded readers.

The associations provided by the students were analyzed using Latent Semantic Analysis, a method for computing semantic similarity between words (Landauer & Dumais, 1997). The associations a learner provided for each target word were assigned a similarity value representing how similar they were to the target word to which they were provided. The hypothesis was that the students who engaged in the post-reading discussion activities would show greater increases in associational knowledge of the target words than those students who did not participate in the discussions.

The major finding from this thesis was that the students who struggled with a word during the post-reading discussion and were provided an opportunity to discuss the word with their group developed associational knowledge to a significantly greater degree than those students who did not encounter the words during the discussions. This emphasizes the facilitative role that meaning-focused output activities have on vocabulary development. In addition, the associational knowledge developed at the initial stages of word learning (i.e., from *no* knowledge to *form* knowledge), continued to develop from *form* knowledge of a word to *meaning* knowledge of the word, and was also developing even when words did not change in reported knowledge. This suggests a continual restructuring of the learners' lexicon, exemplifying past research (e.g., Henriksen, 1999). Overall, the findings suggest that an extensive reading approach which includes opportunities for meaning-focused interaction has greater benefits for lexical development when compared to a traditional reading-only approach to extensive reading.

Acknowledgments

The research presented in this thesis seems insignificant when compared to the amount of knowledge and experience I have gained along my PhD sojourn. I am honored to have had the opportunity to exchange ideas with the brightest, most creative minds of our time. First and foremost, I would like to thank my supervisors John Macalister, Irina Elgort, and Anna Siyanova, for providing just the right amount of direction to get me through.

Next, I would like to thank the teachers that allowed me into their classroom, in random order: Cherie Connor, Joanne Tham, Natalia Peterson, Sarah Ann Petus, Eva Jiang, Angela Joe, Kristen Sharma, Shelley Dawson, Joan Callanan, Heather Roberts, Lauren Whitty, Anna Dowling, Liz Covington, and Mark Toomer. I also thank their students who became my participants for allowing me into their minds; without them this thesis would not have been completed.

I thank those who helped me with data collection and analysis: Emily Greenbank, Dr Keely Kidner, Debbie Evans, Dr Diego Gonzalo Navarro, and Matthew Sorola. I also thank Dalice Sim and Lisa Woods for their statistical assistance.

I would like to thank the Von Zedlitz posse and 22KP for reminding me that life exists outside of a PhD. Dr Khadij Gharibi, Shelly, Evan, Pakjira, Natalia, Cikgu, et al., you know who you are.

Finally, a special thank you to my mom and sisters, who keep me pushing.

You have all gotten me through this thesis. Thank you.

Conference presentations derived from this research

1. Boutorwick, T. (2015). Enhanced Extensive Reading and second language learning: Learner case studies. Presented at the JALT conference, Shizuoka, Japan, November 20-23.
2. Boutorwick, T. (2015). Enhanced Extensive Reading and second language learning. Presented at the American Association of Applied Linguistics conference, Toronto, Canada, March 21-24.

Dedication

This thesis is dedicated my father, Thomas Edward Boutorwick, who died November 29^{th}, 2003 after a year-long battle with cancer. I made it this far because of him.

Contents

1 Introduction

1.1 *Vignette*: Language learning experience

My first exposure to a second language occurred around the age of ten. A family of four from Japan moved in to my neighborhood, including two children about the same age as I was. Our families soon became friends and we would often visit each other. In addition to the mesmerizing television programs and the delicious food, I was fascinated by the language they used and how it seemed so different to English. I was interested in finding out more and began formal Japanese language study. That is when I discovered Kanji, the Japanese logograms borrowed from Mandarin Chinese. I was captivated by the idea that words were represented as pictures, not as groups of letters like English. I was intrigued by the components which combined to make each of these Kanji characters; they themselves contained information pertaining to the meaning of the character. This Kanji fascination led me to spend many hours reading books, studying vocabulary cards, and paging through dictionaries. As my studies furthered, and I decided to pursue a Bachelor of Arts degree in the Japanese language, my free time was spent in a cafe in the basement of the university library engulfed in the language. In addition to reading and studying vocabulary cards, I discovered the joy of being able to use the words I was learning. I enjoyed trying to incorporate newly-acquired words into conversations with others in Japanese, and remember the rewarding feeling I got when succeeding. I also remember the less than rewarding feeling of not succeeding, of not conveying the intended meaning, but appreciated the opportunity it presented to receive assistance from my peers so that my intended message would become clear.

Learning Japanese made me aware of how words carry meaning. This experience of second language vocabulary acquisition through a combination of input-based activities especially reading, and through output-based activities such as discussions, both contributed in important ways to my understanding of vocabulary learning. It is from this perspective that the current research is implemented, investigating the extent that input-based learning and output-based learning affect second language vocabulary de-

velopment.

1.2 Aims of the study

This thesis compares two approaches to extensive reading to determine the extent that they facilitate lexical development. The first approach is a traditional reading-only approach, and the second approach is a task-based approach which supplements reading with post-reading meaning-focused discussions.

1.3 Background to the study

Lexical development has become a central focus in the field of second language acquisition (Llach, 2011). Researchers have examined vocabulary knowledge from different perspectives and with different criteria, yet they all agree that word knowledge is a multidimensional construct (e.g., Laufer, 1997, Nation, 2013a, Webb, 2005). This means that learning a word is an incremental process that begins when the word is initially encountered, and can continue after other aspects of the language (e.g., grammar) have been learned making vocabulary knowledge an intriguing language learning phenomenon (Schmitt, 1998).

If vocabulary knowledge develops through exposure to language, then learners should be exposed to a large amount of the language so that the knowledge can develop. One way to achieve this is through extensive reading, the reading of easy, enjoyable material in large quantities (Day & Bamford, 2002). Extensive reading allows learners to engage with words in a contextualized, authentic environment, providing a wealth of meaningful input. Empirically, research has shown that the environment provided by extensive reading does facilitate language learning. A seminal study by Elley and Manghubai (1981) revealed the power that reading can have on language development. In their study, primary school students in Fiji engaged in extensive reading for approximately eight months. At the end of the reading, the participants sat a battery of tests to assess any learning which occurred. They found that the students increased their

knowledge of multiple areas of proficiency including spelling, writing, grammar and meaning knowledge. In some cases, the students who read developed almost twice as fast as students who did not engage in extensive reading.

However, while extensive reading is beneficial to vocabulary development, it is not the only way that learners develop vocabulary knowledge. They can also increase their vocabulary knowledge through deliberate study. Deliberately learning vocabulary involves focusing attention specifically on learning vocabulary. Research has shown that deliberate study allows for new information to be integrated with existing knowledge (e.g., Joe, 1998; Schmidt, 1990) and can be an effective means to learn words. In addition to the input provided by extensive reading, as well as through deliberate study, output or language production has also been shown to have a facilitating effect on vocabulary development (e.g., Dobao, 2014; Swain & Lapkin, 2002). Language production is important for language learning and especially for vocabulary learning. This is because producing language can lead to instances where a learner questions the meaning of a word they said, questions the correctness of a word's pronunciation, or receives a corrected word form which they were unable to produce correctly (Nguyen, 2013; Williams, 1999).

In short, research shows both input and output to be conducive to language learning, and both should be taken into account when assessing lexical development. In a recent meta-analysis, Nakanishi (2015) assessed 34 studies carefully chosen from 1989 to 2012 to investigate the overall effectiveness of extensive reading. He assessed the 34 studies by grouping them based on similar design features, such as the area of proficiency examined (e.g., vocabulary knowledge), and the instruments used to measure learning (e.g., multiple-choice tests assessing word meaning). Interestingly, a closer look at the 34 studies reveals that in many instances reading was not the only activity the participants engaged in. They also took part in a variety of activities which provided opportunity for language production. That these activities were included in the original research but not taken into account in Nakanishi's (2015) meta-analysis may give the impression that the studies incorporated a traditional reading-only approach to extensive reading.

This is misleading, and the addition of the activities, and the lack of investigation into their effects on development means that it is not possible to know the effects of the reading by itself since it is possible that knowledge was acquired during the activities in addition to reading.

To that extent, the current study includes a detailed analysis of the post-reading discussion activities to determine their effect on lexical development. At the theoretical level, the research in this thesis can provide insights into the combined effects of input and output on vocabulary knowledge development. At a more practical level, the research presented herein can talk to those practitioners who believe both input and output to be more fruitful in terms of developing second language vocabulary knowledge.

1.4 Research questions

The following research questions motivate this thesis:

1. How do different approaches to ER affect L2 vocabulary development, specifically an ER-only approach and an ER-plus approach?
2. How do different types of interaction following ER affect L2 vocabulary development, specifically spoken interaction and written interaction?
3. What additional aspects of language do learners focus on during the interactions?

1.5 Organization of the thesis

This thesis consists of six chapters. Following this brief introductory chapter, Chapter Two gives a detailed account of research and key concepts about second language vocabulary learning. This review includes conundrums challenging current research, issues relating to the facilitating nature that extensive reading has on lexical development, and methods for assessing this development. Chapter Three reports the methodology used in the current study and Chapter Four presents the results of the study. Chapter Five discusses the findings of the study in terms of the approaches to extensive reading and the degree that they facilitated development. This chapter also posits reasons for

this facilitation drawing on the literature from Chapter Two. Chapter Six concludes this thesis by summarizing the main findings and discussing the contributions of this thesis. It also discusses the limitations of the research and suggests avenues for future investigation.

The next chapter presents relevant literature regarding lexical development in a second language.

2 Literature

2.1 Introduction

This chapter reviews literature on second language vocabulary acquisition. The field has grown immensely in scope since its birth, and only the areas directly relevant to the study are dealt with. The chapter begins by explaining the importance of researching second language vocabulary acquisition in section 2.2. The next section, section 2.3, discusses some of the major questions that researchers deal with when conducting research into vocabulary learning. These include defining a word (section 2.3.1), determining how many words need to be known and what it means to know a word (section 2.3.2), how many times it takes of seeing a word before learning occurs (section 2.3.3), and how words are learned (section 2.3.4). Section 2.4 introduces extensive reading, one major way for learning vocabulary in a foreign language. The following sections describe the relevant limitations found in research surrounding extensive reading (sections 2.4.1 through 2.4.5). Next, section 2.5 mentions that ER can be used in a way that allows learners to use and remember vocabulary, rather than read and forget. The next section introduces Latent Semantic Analysis, a technique which is used in this thesis to assign association strength to word associations. Section 2.6 summarizes the issues to be addressed, and section 2.7 introduces the research questions which motivate the research carried out in this thesis.

2.2 Why is L2 vocabulary acquisition worth investigating?

Approaches to teaching and learning a foreign language have traditionally de-emphasized vocabulary and focused on learning grammar structures, for example through grammar translation, limiting vocabulary learning to memorization techniques and decontextualized word lists (Kelly, 1969; Nation, 2013a; Richards & Rodgers, 2001). This de-emphasizing of lexis was in large part due to political changes in Europe in the sixteenth century. At that time, Latin had been the most widely-studied language. However, as

the importance of Spanish, French, and English increased, Latin became reserved for the study of the classics, for examining the grammar and rhetoric of classical works. As these new languages rose in importance, the grammar-centric methods applied to studying classical Latin were adopted for learning the new languages (Richards & Rodgers, 2001). This grammar-centric model of language learning would become the foreign language learning model for the next three centuries.

Grammar is an important part of language learning, and very little can be conveyed without it, however without vocabulary nothing can be conveyed (Wilkins, 1976). Vocabulary represents the building blocks from which meaning is communicated, and proficient speakers have relatively little difficulty in remembering the thousands of words needed for conversation, nothing short of a considerable feat (Aitchison, 1987). The process of learning words, whether in a first language or second, begins when the language is initially encountered, and continues long after other aspects of the language system (e.g., grammar) have been mastered. In this way, vocabulary is one of the more intriguing puzzles in language learning (Schmitt, 1998).

The systematic investigation of vocabulary knowledge has become a central tenet in second language acquisition (Llach, 2011), and vocabulary research is at an all-time high. Nation (2013a) calculated that approximately 30% of vocabulary research in the last one hundred years has occurred in the last 10-15 years, meaning vocabulary is no longer a neglected aspect of language learning.

2.3 Conundrums challenging second language vocabulary research

2.3.1 What is a word?

A word can be defined in a number of ways, and the definition ultimately depends on the purpose for which the words will be used or counted. For example, to determine the number of words in a book, or how many words someone can type in a minute, a word can be defined as every grouping of characters surrounded by space, or a word

token. On the other hand, if the same person were to attempt to count how large their vocabulary is, using tokens would not be appropriate. Instead, a word would be better defined as each unique grouping of characters surrounded by space, or a word *type*. For example, in the phrase "a dog and a girl" there are five word *tokens*, but only four word *types* since "a" occurs twice. Using word *types* as the definition of a word means that the plural of a word, e.g., *footbags*, is seen as unique to the singular form *footbag*, and this may not be desirable. Instead, it might be more appropriate to include inflections (e.g., *shred*, *shreds*, *shredded*) and reduced forms (e.g., can't -> *cannot*), i.e., a word *lemma*. Underlying the use of the *lemma* is the idea of learning burden, or the amount of effort required to learn a word (Nation, 2013a; Swenson & West, 1934). Once a learner understands the inflectional system, the learning burden of word types is reduced. This means that, for example, learning *footbags* after learning *footbag* becomes negligible. Word lemmas tend to have the same part of speech (e.g., the lemma *run* includes *runs* and *running*), yet there are other affixes, i.e., word derivations, which also attach to words. The inclusion of these derivations creates what is known as a word family (Bauer & Nation, 1993; Nation, 2013a), and accordingly, *active*, *actively*, *activity*, and *activities* all become part of the word family *active*.

2.3.2 How many words need to be known, and what it means to know a word?

How many words a learner needs to know depends on the purpose for wanting to learn them. Those wanting to travel to a foreign country for a short vacation may survive on 100-200 word *types*, for example greetings and words used for directions (Nation & Crabbe, 1991). However, those pursuing higher education can find themselves needing more for survival. These students, met with academic texts on a daily-basis, may need to know up to 8,000 word families (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006). But what do these students need to know about these words? Put another way, what is involved in knowing a word? At the most basic level, knowing a word involves recognizing that a string of letters is a word, and not a random arrangement of

characters (Daller, Milton, & Treffers-Daller, 2007). Of course, this by itself would not be useful for most circumstances without also knowing the meaning of the word. This form-meaning combination is the basis for investigating vocabulary size, or *how many* words someone knows.

Vocabulary size is important for various reasons. First, it is an indicator of L2 proficiency, and so can be used for diagnostic purposes (Nemati, 2010). English proficiency programs often place learners into classes by proficiency level, and determining the size of a learner's vocabulary can assist in accurate placement. How is the vocabulary size of an L2 learner measured? One way is using a yes/no format (e.g., Meara & Jones, 1988). The test is made up of 100 words, in sets of 10 from differing frequency bands. The test-taker ticks all of the words that they know. Another test which measures vocabulary size is the Vocabulary Size Test (Nation and Beglar, 2007). Similar to Meara and Jones' (1988) yes/no test, the Vocabulary Size Test consists of target words of varying frequencies of occurrence. Unlike the yes/no test however, the test-taker in the Vocabulary Size Test chooses one of four multiple-choice options which define each of the target words. Both the yes/no test and Vocabulary Size test measure the amount of words which the test-taker knows the meaning of. The form-meaning aspect of word knowledge is one of the most important relationships in vocabulary acquisition since words are units of meaning (Laufer & Goldstein, 2004), and this aspect is one of the initial aspects which tend to be acquired. However, knowing a word involves more than the form-meaning relationship (Cronbach, 1942; Nation, 2013a; Richards, 1976; Webb, 2005). Knowing a word also entails depth of knowledge, or the quality of vocabulary knowledge (Read, 1993). For example, if two learners both knew the meaning of the word *act*, as in *she was acting strange last night*, but the second learner also knew that *act* can refer to a division in a play, the second learner would have greater depth of knowledge of the word *act*. Nation (2013) provides one of the most comprehensive taxonomies of word knowledge, seen in Table 1. At the most general level, the different aspects of word knowledge are categorized into form, meaning, and use. The form-meaning relationship can be seen under the meaning category, in the

subcategory labeled *form and meaning*. It can be also be seen that this is only one of a multitude of aspects of word knowledge; Table 1 depicts 18 aspects, or dimensions of word knowledge which are displayed in Table. This taxonomy views vocabulary knowledge from a dimensions approach to vocabulary knowledge (Dóczi & Kormos, 2016), which addresses each part of word knowledge as aspects. The depth of knowledge of a word develops as a result of accumulating these knowledge dimensions.

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

Table 1: What is involved in knowing a word (Nation, 2013a, p. 49)

Another way to view depth of vocabulary knowledge is to envision the knowledge of a word on a developmental scale. At one end of the scale, there is complete lack of word knowledge. At the other end, there is full mastery of a word (e.g., Wesche & Paribakht, 1996). During the initial stages, acquiring form-meaning knowledge of words means that learners may be able to acknowledge that a word is a word, or be able to accurately select a word's meaning from a set of multiple choice options. As

depth of knowledge increases, and a word moves along the scale towards full mastery, the learner becomes able to use the word in a semantically appropriate manner. One example of a scale used in this respect is the Vocabulary Knowledge Scale (Wesche & Paribakht, 1996). Table 2 shows an example of possible stages of knowledge in this scale, from no knowledge (the lowest level) to accurate usage (the highest level).

	Self-report categories
I.	I don't remember having seen this word before.
II.	I have seen this word before, but I don't know what it means.
III.	I have seen this word before, and I *think* it means _____. (synonym or translation)
IV.	I *know* this word. It means _____ (synonym or translation)
V.	I can use this word in a sentence: _____. (If you do this section, please also do Section IV.)

Table 2: The Vocabulary Knowledge Scale (Wesche & Paribakht, 1996)

A third approach used to categorize depth of knowledge is through lexical networks (Henriksen, 1999). From this point of view, a learner's lexicon is conceptualized as a semantic network with words connected to each other through the relationships between them (Haastrup & Henriksen, 2000; Read, 2004; Waring & Nation, 2004). A learner's network develops as the relationships between the words increase, and as a result the network restructures, creating a well-developed network and increasing depth of knowledge. A well-developed network allows for lexical properties to be represented economically, and the connectivity allows for inferences to be drawn and generalizations to be made (Ellis, 1994; Murphy, 2004).

2.3.3 Frequency of occurrence

On top of learning thousands of words, and several aspects of them, this learning is not likely to happen without repeated exposure to the word. This is because initial

exposures to a novel word tend to create lexical knowledge that is fragile and incomplete (Elgort et al, 2016; Landauer & Dumais, 1997). While research is not clear on a specific number of occurrences required for acquisition to occur, some research (e.g., Bolger et al., 2008) has shown that three or four exposures are sufficient for some aspects of word knowledge to develop, while other research has shown that closer to six times is needed (Rott, 1999) and yet other research suggests eight to 10 exposures is the point where incidental learning begins to accelerate (e.g., Horst, Cobb, & Nicolae, 2005; Pigada & Schmitt, 2006; Waring & Takaki, 2003).

2.3.4 How are words learned?

With thousands of words to learn, multiple aspects to be learned about each word, and repeated exposure facilitating acquisition, how is it that words are learned? One way is through deliberate study. Some scholars say deliberate learning may be the main avenue of learning since it creates opportunities for language learners to notice new vocabulary (Long & Robinson, 1998; Schmidt, 1990). Deliberate learning involves focused attention on vocabulary, allowing for new information to be integrated with existing knowledge (Craik & Tulving, 1975; Joe, 1995, 1998). One deliberate learning activity is the Keyword Technique (Nation, 2013a). This technique links a first-language word to a second language word which it sounds like, and then the two words are used to conjure an image. For example, for a Japanese learner to learn *sorry*, the learner thinks of a word which sounds like *sorry*, *sori* (razor), and then creates an image of a hairdresser cutting the cheek of a bearded customer, apologizing immediately after. Language classrooms are one place that deliberate learning often occurs. Nation (1980) for example, found that learners learned 30-100 words per hour when using bilingual vocabulary cards, with the second-language word on one side of the card, and its first-language translation on the other side. Using Nation's 30-100 word estimate, approximately two hours of deliberate learning would be needed to learn the survival vocabulary required for the short-term travel mentioned in section 2.3.2. For the students who plan to enter tertiary education, where upwards of 8,000 word families are required, this equates

to approximately 123 hours. Two hours of deliberate learning spread over the length of a course is more feasible than spreading 123 hours, and with class time often too restricted to provide sufficient opportunities for deliberate learning (Hunt & Beglar, 2005), deliberate study is not the only means from which vocabulary knowledge is learned; it can also be learned incidentally, while attending to something other than word learning itself.

Some scholars believe that incidental learning of vocabulary during reading may be the easiest and single most powerful means of promoting large-scale vocabulary growth in an L1 (e.g., Krashen, 1998; Mckeown & Curtis, 1987). This is because the primary goal of reading is to understand the meaning of what is being read and not to learn vocabulary. Since the focus of attention is on the story, any vocabulary learning which takes place occurs as a by-product, without the intent to commit the knowledge to memory (Hulstijn, 2013). In a seminal study, Nagy, Herman and Anderson (1985) measured the effectiveness of incidental learning on 57 junior high school students' vocabulary knowledge. The students were placed into one of two reading conditions. In the narrative condition, students read a mystery-style story, and in the expository condition the students read an excerpt from a junior high school earth science textbook. Both texts were about 1,000 words in length. Among the words on a post-reading test, 30 were pre-selected from the texts as target words, 15 from each text. To assess the extent that incidental learning took place, the participants were asked to explain the meaning of these words during an interview. After the interview the participants completed a multiple-choice test. The results of the study revealed a significant effect of incidental learning through reading in the interview test ($F = 75.8$, $p < 0.01$), as well as the multiple-choice test ($F = 34.3$, $p < 0.01$). The authors emphasize the fact that 23 of the 30 target words occurred as hapax legomenon. This means that in many of the instances, it took only one exposure to learn something about the word. They synthesize their results with previous research (e.g., Anderson & Freebody, 1983) to conclude that a typical junior high school student learns between 750 and 5,500 words per year from incidental learning through reading.

The benefits of incidental learning through reading also hold true in an L2 because from a cognitive point of view there is no essential difference between fluent L1 and fluent L2 reading (Day & Bamford, 1998; 2002). Advocating this idea, reading has become a main means of measuring incidental vocabulary acquisition in an L2 as an important source of incidental learning (Widdowson, 1979). Incidental learning in second language vocabulary acquisition research refers to the acquisition of vocabulary knowledge without a conscious intent to commit it to memory (Hulstijn, 2013). This type of learning typically happens in a slow, incremental manner, in part due to the amount of knowledge involved in knowing a word. Some scholars posit that vocabulary learning is in large part the result of implicit associations made between words, not the explicit learning of their meanings (Landauer & Dumais, 1997), and others suggest that except for the first few thousand words, vocabulary learning predominantly occurs through extensive reading (Huckin & Coady, 1998).

2.4 Extensive reading in a second language

At its core, extensive reading is the reading of easy, enjoyable material in large quantities (Day & Bamford, 2002). In a seminal study of primary school students in rural Fiji, Elley & Mangubhai (1981) revealed the power that reading can have on language development. Twelve rural Fijian schools participated in the study and were each assigned to one of three conditions: a shared book condition, a silent reading condition, and a control group. Each of these three conditions was comprised of four schools. In the shared book condition, the teacher chose an interesting story to read aloud to their class. Before reading, the teacher had whole-class discussions related to the pictures in the book, asked what the students thought plot likely entailed, and also attended to any words unknown to the students. After reading, the teacher carried out post-reading activities included drawing, role playing and writing. In the next group, the silent reading group, teachers displayed the books in a way that would attract students' attention, and encouraged up to 30 minutes of in-class reading a day. The third group,

the control group, kept to their normal English program.

The findings in Elley and Mangubhai (1981) revealed that both reading groups outperformed the control group in reading and listening comprehension; Some of the students who read made 15 months growth in only eight months, twice the rate at which some students in the control group progressed. Moreover, the gains found in the reading groups were not limited to one area of proficiency, but included reading, writing, vocabulary and grammar.

Subsequent research has investigated specific aspects of language proficiency, in addition to a comprehensive approach used such as in Elley and Mangubhai (1981). Janopoulos (1986), for example, surveyed 79 foreign graduate students about the extent to which they engaged in extensive reading. The students were asked to self-report the amount of time they spent reading. They also gave a sample of their writing to determine their proficiency level. The learners' proficiency was measured by an obligatory hour-long writing sample produced upon entrance to the university which the learners were enrolled. The results revealed a significant positive correlation between L2 reading and writing quality, leading Janopoulos (1986) to the conclusion that avid L2 readers were more likely to be proficient in L2 writing than those who were not avid readers.

In another study, Hafiz and Tudor (1989) assessed the effects of extensive reading on L2 reading and writing proficiency, determined by the National Foundation for Education Research's (NFER) test of proficiency in English. A group of 15 learners aged 10-11 read extensively one hour a day, five days a week for a period of 12 weeks. This group was compared to two control groups of 15 students: a group from the same school, and another group from a different school in the same city. The participants were pretested and then post-tested after the reading treatment finished. The results revealed that the ER group made significant gains in reading and writing, while the control groups did not make a significant increase. Hafiz and Tudor (1989) conclude that ER substantially affected the participants' language proficiency in terms of reading and writing proficiency. Tudor and Hafiz (1989) carried out a subsequent analysis on

one set of the data collected, the learners' written compositions they produced for the writing sections of the NFER test. The authors were interested in investigating the nature of the development found in their study. The students' compositions were scored according to four criteria:

- *Writing Readiness* (e.g., total tokens produced),
- *Vocabulary range* (e.g., total types produced),
- *Syntactic maturity* (e.g., sentence length), and
- *Accuracy of expression* (e.g., spelling correctness).

The results showed significant increases in writing readiness since the participants wrote significantly more on the post-test than the pretest. More notable however was the significant increase in accuracy of expression, with semantically acceptable sentences showing the greatest gains, from 26% acceptable on the pretest, to 67% on the post-test. The authors conclude that ER has positive benefits for L2 development, and this can feed through to written production as well.

Research conducted in the 1980s and 1990s revealed the facilitative effects that ER can have on L2 proficiency. Amalgamating this body of research, Day and Bamford (2002) formulated ten principles they promoted as elements of a successful ER program:

1. Students read as much as possible, in and out of the classroom.
2. A variety of materials on a variety of topics, to encourage reading for various reasons and in various ways.
3. Students select what they want to read and have the freedom to stop reading material that fails to interest them.
4. Reading is done for pleasure, information, and general understanding.
5. Reading is its own reward. There are few or no follow-up exercises after reading.
6. Reading materials are within the linguistic competence of the students regarding vocabulary and grammar.
7. Reading is individual and silent, with students reading at their own pace.

8. Reading speed is usually faster rather than slower.
9. Teachers explain the goals of the ER program to the students.
10. The teacher is a role model for students, actively taking part in reading.

Day and Bamford's (2002) list has provided a principled approach to implementation of ER research for decades. However, this research has typically been conducted independently of other research, resulting in difficulties determining how effective extensive reading actually is (Nakanishi, 2015). To combat this, Nakanishi (2015) conducted a meta-analysis of 34 carefully-chosen ER studies between 1989 and 2012. Research that met the following criteria was included in his study:

- The study concerned extensive reading, pleasure reading, or graded readers,
- The participants had to engage in ER for a certain period of time,
- The study was experimental or quasi-experimental,
- Means and standard deviations, or t-values or F-values were presented,
- The participants were children or older (p. 12-13).

Nakanishi (2015) analyzed the studies to determine how many had used a control group(s), how many adopted a pretest/post-test design, which areas of proficiency were being examined (e.g., vocabulary), and which instruments were used to measure any development. The results showed that the top three areas which the ER research assessed were reading comprehension (74%), reading speed (40%), and vocabulary knowledge (35%) (These percentages will not sum to 100% since some studies fit more than one category). In order to investigate these areas, the studies typically used reading comprehension tests (60%) including the Test of English as a Foreign Language (TOEFL), and the Edinburgh project on extensive reading (EPER) placement/progress tests. Reading rate tests and cloze tests were also used frequently (40% and 37%, respectively).

2.4.1 Issue I: Additional activities not given attention

The overall conclusion from Nakanishi's (2015) meta-analysis is that ER facilitates L2 learning. However, the extent that ER is responsible for the learning found in these

studies may need to be examined more carefully. Some of the research suggests that it is the quality of the context which determines word learning, not repetition (Webb, 2008). In Webb's study, he examined the quality of a context in terms of the amount of information a sentence provided about a target word. Fifty Japanese EFL university students participated in the study, and were assessed on 10 target words. The students read three sets of 10 sentences, with each target word appearing once in each set for a total of three occurrences for each target word. The sets varied in the amount of information they provided about the target word contained in it. A more informative condition was compared to a less informative condition to determine the extent that the words were learned from the contexts. The students in the more informative condition received sets with informative contexts, while those in the less information group received sets with less informative contexts than the more informative condition. A surprise post-test including four types of tests conducted immediately following the reading showed that those students in the more informative context group scored higher on the recall of meaning as well as recognition of meaning. How informative a context was did not predict recall or recognition of form. Webb (2008) concludes that the quality of context may outweigh frequency of occurrence for acquiring knowledge of meaning, while frequency of occurrence may be a better predictor of orthographic learning.

The second reason interpreting these results should be done with caution is because many of the studies used in Nakanishi (2015) included supplementary activities in addition to ER. For example, some of the studies used book reports and small-group discussions about the books (e.g., Beglar, Hunt, & Kite, 2012; Horst, 2005; Yamashita, 2008). Other research devoted class time to completing worksheets and filling in vocabulary notebooks based on what was read (e.g., Horst, 2005). In one study, Al-Homoud & Schmitt (2009) compared a group of learners who did intensive reading (line-by-line grammar translation) with a group who did extensive reading. The extensive reading group was encouraged to read, and also completed warm-up activities which had them practicing intensive reading skills, and post-reading activities including vocabu-

lary learning strategies and general discussion on reactions to the books they read. As a result of these additional activities, it becomes difficult to determine the effects that extensive reading had by itself on the development of language.

Learning vocabulary through output That the activities may have impacted the extent that proficiency developed stems from the idea that vocabulary can be learned through the activities, especially those which involve learners producing language collaboratively, as in a post-reading discussion. Advocates of vocabulary learning through output believe that language-production compels the learner to undertake full processing of the language they possess (e.g., Gass, 1988; Joe, 1998; Swain, 1985). Some benefits of output 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).

Output activities are beneficial for language learning because they provide an environment which allows for modification of output, leading to increased comprehensibility, complexity, and accuracy (de la Fuente, 2002; Pica, 1996). During an activity, learners interact with each other in a way that connects input, internal learner capabilities, and output in productive ways. These features of interaction have been a core issue in second language acquisition research since the seminal work of Hatch (1978) found a relationship between interaction and language acquisition. Often when learners communicate with each other the flow of their discourse will break down, or be on the verge of breakdown, due to perceived problems with comprehension or production, thus necessitating negotiation (Long & Robinson, 1998; Pica 1996). As a result of this problem, learner focus is shifted from the message or topic of the discussion, to the language that the learners are producing. The negotiation work which ensues resolves the misunderstanding and is referred to as negotiation of meaning (Ellis, 2003). This negotiation allows for any (or all) of the following communication strategies to occur:

- 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). In a small-group discussion for example, a person can be asked to clarify something they have said. In order to clarify, they must modify their output in an attempt to make their intended message clear. By doing so they are given the opportunity to think deeply about their message, and produce a new message which conveys a similar meaning. In some situations, depending on the linguistic ability of the learner, it may be beyond their capability to modify their message accurately, or at all. In these instances, it is possible to ask for assistance from the group, leading to collective scaffolding as knowledge of the group is pooled together, facilitating effective communication, and increasing comprehension (Donato, 1994; Lyster & Ranta, 1997; Pica, Holliday, Lewis, & Morgenthaler, 1989). Negotiation provides opportunities for mental resources to develop, increased comprehensible input that is co-constructed, and opportunities for learners to explicitly *focus on form* while expressing meaning (Long, 1983; Swain, 1985, 1995).

Focus on form - Episodes of learning The focus on form episodes which occur during interaction allow learners to reflect on their own knowledge and actively refine it (Swain, 1995, 1998; Kowal & Swain, 1994, Swain & Lapkin, 1995). These episodes, also referred to as language-related episodes (LREs) can include instances in which learners:

- question the meaning of a linguistic item,
- question the correctness of a word's spelling/pronunciation,
- question the correctness of a grammatical form, or
- correct their own (or someone else's) usage of a word (Nguyen, 2013; Swain & Lapkin, 1998; Williams, 2001).

Two types of focus on form episodes have been researched. In *planned* focus on form, focused tasks are used to elicit the use of specific linguistic features in the context of meaning-centered language use (Zhao & Bitchener, 2007). A focused task could be an information gap activity, where the target words of interest are specifically elicited. The second type of episode, *incidental* focus on form, occurs during unfocused tasks, or communicative activities designed to elicit general samples of the language rather than specific features (Ellis, 2001; Ellis et al., 2002). An example of incidental focus on form would occur during a post-reading discussion, when learners were discussing a book that they read, and one of the group members forgets how to say a certain word. Incidental focus on form integrates meaning-focused and form-focused activities, a balance between the naturalist approach (e.g., Krashen & Terrell, 1983), and planned focus on form (Long & Robinson, 1998).

Incidental focus on form has received less empirical attention compared to planned focus on form (Zhao & Bitchener, 2007). Earlier research examining incidental focus on for was primarily descriptive in nature and revealed relationships between the focus on form episode and accurate language performance (e.g., Ellis et al, 2001; Lyster & Ranta, 1997; Williams, 2001). The language facilitating aspects of the focus on form episodes lie in the nature of the episode meaning it is essential to examine the episodes themselves (Williams, 2001). Loewen (2005) investigated focus on form episodes to better understand the extent to which they facilitated learning. Twelve intact classes at a private language school participated in the study. Four lessons were observed in a one-week period for each of the 12 intact classes. Each of the teachers wore a clip-on microphone, recording all teacher-learner interaction. It did not capture learner-learner interaction when the teacher was not in the vicinity. The episodes were located using a 17-hour subset of the recordings resulting in a total of 473 episodes. Each learner was tested on their respective episodes. Individual test items for each student were created based on the linguistic items that arose in their discourse. The type of test item depended on the focus of the episode and resulted in three question types. The first type was classified as *correction* and required learners to improve sentences

that they had produced incorrectly during an episode. This question type focused on grammatical knowledge. The second question type, *suppliance*, required learners to provide linguistic information. Suppliance questions focused on word meaning and spelling. The third question type, *pronunciation*, focused on the pronunciation of words mispronounced during an LRE. The results are summarized in Table 3, showing only significant predictor variables of learning from the LREs for each question type. Note that the p-value was set at 0.15 as Loewen notes that when using forward step-wise regression, as he did, a p-value of 0.05 is too stringent and might exclude important variables (Hosmer & Lemeshow, 2000).

Question type	Linguistic Focus	Predictor variables	Odds ratio	p value
Correction	grammar	Uptake	0.522	0.146
		Successful uptake	2.149	0.033
Suppliance	meaning; spelling	Uptake	0.208	0.010
		Successful uptake	8.797	0.000
Pronunciation	pronunciation	Complexity	0.309	0.007
		Source	2.997	0.020
		Successful uptake	1.906	0.127

Table 3: Results of Loewen's (2005) LRE analysis (adapted from Loewen, 2005).

Table 3 reveals a number of facilitative characteristics of the language-related episodes. The two most common predictors of learning grammar and word meaning were successful uptake and uptake. Successful uptake occurs when a learner incorporates the linguistic item they had difficulty with into their own production (Ellis et al, 2001). The fourth column of Table 3 ratio labeled *Odds ratio* shows that the episodes focusing on grammar and pronunciation which included successful uptake were approximately twice as likely to lead to learning compared to episodes without successful uptake. For those episodes focusing on meaning and spelling, the effect was more profound; an episode with successful uptake was nine times as likely to result in learning than an episode without successful uptake. That successful uptake significantly pre-

dicted learning regardless of which aspect of language was targeted emphasizes the role of language production in acquiring linguistic knowledge.

The second predictor of learning, uptake, refers to student acknowledgment that they had been provided assistance (Shen, 2008). As Table 3 reveals, the grammar-focused episodes which included uptake of this kind were half as likely to lead to subsequent learning compared to episodes without uptake. Likewise, episodes focusing on meaning or spelling which included uptake were one-fifth as likely to lead to subsequent learning compared to episodes without uptake. These two predictors highlight the idea that the presence of uptake may not be as important as the quality of the uptake which occurs (Loewen, 2005).

Another finding to mention is the different predictor variables which occurred only for the pronunciation question types. The first variable, complexity, refers to the number of response moves in an LRE. In Table 4, which depicts an example episode from Loewen (2005), a learner (L) and teacher (T) are talking. In the teacher's first turn they correct the learner's error of double pluralizing the word *children*. After the learner repeats the correct form, the teacher responds again, this time providing a meta-linguistic explanation. Thus, this episode is complex since the teacher has responded more than once. The results in Table 4 revealed that complex episodes were one-third as likely to result in subsequent learning than simple episodes which involve only one response move. With respect to pronunciation of a word, this result suggests that if a learner makes a pronunciation error and corrects it in the same turn, they may have some control over the phonological feature. However, if they struggle with the word and are not able to correct it immediately, they may need more time explicitly focusing on the word.

The other variable found to predict learning in only the pronunciation items related to the source of the episode, or in other words the reason for instigation of the episode. In the example presented in Table 4, no communication breakdown has occurred because the teacher understood what the learner was trying to communicate. As a result, the source of the LRE is *code-related* since it was only the inaccurate use of the plural in

L: she works full time and uh *pay take care of her childrens until she...*
T: *children*
L: *children*
T: *no S*
L: *yeah children (.) and she get no finance finance <finance-al> help...*

Table 4: Complexity and source aspects of an LRE (adapted from Loewen, 2005)

the word *children* which was being addressed, not the meaning of the message conveyed by the learner. However, the results in Table 3 reveal that *message-related* LREs were approximately three times more likely to result in learning than *code-related* LREs. Loewen (2005) speculates that phonological problems interfering with the meaning of a message may have been closer to the learners' ability to improve than the phonological items which focused on accuracy. That the predictors differed depending on the question type suggests that different kinds of focus on form may be necessary for different aspects of language (Nicholas et al., 2001). From these results, Loewen (2005) concludes that incidental focus on form can facilitate linguistic accuracy when learners are engaged in meaning-focused activities.

2.4.2 Issue II: Time on task is unbalanced

In another study included in Nakanishi (2015), Tanaka & Stapleton (2007) looked at two groups of Japanese high school students over five months: an extensive reading group and a control group. In the extensive reading group, learners were assigned teacher-produced readings and also were encouraged to read graded readers that their instructor brought to each class. The instructor read the texts aloud, as well as checked their understanding of keywords by asking for the meaning of the words. The reading concluded with comprehension questions. The control group did not engage in the ER, but rather kept to the normal class program. The students were assessed using the reading comprehension section of the Society for Testing English Proficiency Test in practical English proficiency (the STEP test). The results revealed that the ER-only

group outperformed the control group on the post-test ($t = 2.5$, $p = 0.01$), despite there being no difference on the pretest between the two groups. The authors also analyzed the data by omitting those students in the extensive reading group who also read graded readers (since not everyone did), to determine the extent that ER facilitated learning. This time, they found no significant differences compared to the control group, concluding that it was the exposure to the graded readers which lead to the gains.

One issue with Tanaka & Stapleton (2007) is that the eighteen students who read the graded readers spent more time reading than the other students. As a result of this increased exposure, it is perhaps not surprising that they made the greatest gains. Beglar, Hunt, and Kite (2012) found a positive correlation between time on task and language learning. In their study, four intact classes each formed one of four reading groups. The first group acted as a control group and reading intensively. The second group also read intensively, but in addition they also read graded readers. The third group only read extensively and no intensive reading was done as in the previous two groups. The fourth group also only read extensively, but read more than the third group. The effects of the reading treatments were measured via a reading rate test, a 32-item test consisting of four 400-word passages followed by the 32 multiple-choice comprehension questions. The results revealed that the groups who read more made greater gains in reading rate, or in other words the more time spent on task, the greater the gains will be.

2.4.3 Issue III: Greater learning potential in the ER-only condition

In one of the studies included in Nakanishi (2015), Smith (2006) compared three reading conditions: an intensive reading condition, an extensive reading condition, and an extensive reading + activities group, where the reading was supplemented with post-reading reaction reports. Participant knowledge was measured using the Edinburgh Project of Extensive Reading Placement/Progress Test (versions A and B), and the College Students English Proficiency Test. Version A was administered at the beginning and end of the school year, while version B was given midway through. The College

Students English Proficiency Test was given one month before the treatment started, and five months after the treatment ended. The results showed that the ER group (the group without the activities) made the biggest gains in both measures, leading to the conclusion that extensive reading was more effective in producing gains in English competence. It was not mentioned however that the group which made the biggest gains, who happened to be the extensive reading only condition, also had the lowest scores on the pretest. This means they had the most potential for learning since there was more for them to learn. It is thus possible that this group showed the greatest improvement (at least in part) because they had the most to learn, not necessarily because of the ER treatment.

2.4.4 Issue IV: What is meant by extensive reading?

Nakanishi (2015) ends by stating that extensive reading can be an effective means of improving language proficiency, but to what extent is this true? In a surprising amount of the research included in the meta-analysis, activities based around the extensive reading were also being implemented, meaning it becomes difficult to know the extent to which extensive reading by itself contributes to learning (Yamashita, 2008). With the inclusion of these activities, to what extent can it be said that ER is responsible for the changes observed? The answer is that it depends on what is meant by ER. As mentioned earlier, extensive reading essentially involves reading easy, enjoyable material in large quantities. Bruton (2002) suggests that extensive reading is simply reading extensively. This is in contrast to Day and Bamford's (2002) *approach* to reading as seen through their ten principles. Nakanishi's (2015) meta-analysis gives the impression that ER is a standalone phenomenon, with the activities present in the individual studies not worthy of empirical attention. However, ER does not have to only be silent reading, an idea which is gaining momentum (e.g., Green, 2005) and brings to attention the fact that Day and Bamford's (2002) principles can be thought of as guidelines rather than commandments (Macalister, 2015). For example, it may not be feasible to obtain a variety of books on a wide range of topics, especially in resource-poor environments

(Macalister, 2014), thus violating Day & Bamford's second principle. Additionally, it may not be possible that students can select what they read in these resource-poor environments, contradicting the third principle. Similarly, the integration of ER and supplementary activities means that reading may not be done for its own reward, but as a platform for supplementary material, violating the fifth principle.

The opportunities the activities provide to learners to share opinions about the stories read, to have focused, stimulating discussion about the content of the books goes unnoticed. These opportunities are unique since they arise when learners are given opportunities to *use* language, focusing on meaning, to discuss the stories they have all read, as is the case in post-reading discussions (e.g., Macalister, 2014). How can these activities be better integrated? One way could be to see the ER as part of a *task*, i.e., an activity which requires learners to use language, with emphasis on meaning, to attain an objective (Bygate, 2001). In this way, the ER could become a springboard for a post-reading discussion. A task-based approach such as this allows for the content in the books to be discussed and used, rather than read and forgotten. In this respect, learners' interest is engaged as they focus primarily on meaning to complete an objective (Willis & Willis, 2007).

2.4.5 Issue V: Measuring knowledge of meaning may be limited

To measure the extent that extensive reading facilitated increases in proficiency, Nakanishi (2015) looked also at the types of instruments used across the studies included in his meta-analysis. He found that of the 34 studies, a total of 20 different tests were used to measure changes in proficiency. Nakanishi split the studies according to their research design and then analyzed each category separately. In the first category were those studies which adopted a pretest/post-test design (n = 21). The results revealed a significant effect of ER on vocabulary learning ($d = 1.25$, CI = 0.14 - 2.35), and Nakanishi concludes that the overall effect size was large enough that ER was responsible. The second group he analyzed included the studies which compared an ER group to a control group. The results from this group of studies revealed no significant effect of

extensive reading on vocabulary learning (d = 0.18, CI = -0.60 - 0.96).

There are two issues that may be making the results in the previous paragraph difficult to interpret. The first issue is that there was no mention of possible effects that the supplementary activities may have had on learning; in addition to the treatment, students in the studies were also attending English classes, and their learning was thus not limited to extensive reading. This point was made in section 2.4.1. The second issue is that the tests used to measure learning may not be optimized for the type of learning which takes place through reading. The most commonly used vocabulary test in Nakanishi's (2015) meta-analysis was the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001), used in 4 of the studies. The underlying construct of the test is that it measures initial receptive knowledge of the meaning of a word (Schmitt, Schmitt, & Clapham, 2001). Inasmuch as this type of knowledge is critical for development of vocabulary, learning a word involves more than just one aspect of knowledge (see Table 1 in section 2.3.2 for an example of what is involved in knowing a word). Similarly, words can have more than one meaning sense, as Table 5 shows. The table depicts a typical question format present in many of the general proficiency tests, e.g., the Test of English as a Foreign Language (TOEFL):

Humans have a innate ability to recognize the taste of salt because it provides us with sodium, an element which is essential to life. Although too much salt in our diet may be unhealthy, we must consume a certain amount of it to maintain our wellbeing.
Question: What is the meaning of consume in this text? a. use up completely b. eat or drink c. spend wastefully d. destroy

Table 5: Context-dependent vocabulary test question example (adopted from Read, 2000)

The word *consume* is being assessed on one aspect of its meaning, which in this case, refers to option *b*, to eat or drink. This represents one aspect of meaning for *consume*, yet *consume* can also refer to any of the other options provided, depending on the context. For example, in the sentence "The blackhole consumed everything society had built", *consumed* is being used to mean *destroy*. By focusing on one meaning sense, as test items such as the one in the table tend to do, no attention is given to the other meaning senses. That is not to say that measures testing one aspect are faulty; tests with similar formats are reliable indicators of various aspects of proficiency (e.g., Nation & Beglar, 2007). But it would be interesting to know if a learner was familiar with *consume* in its other meaning senses as well. Or, since these meaning senses make up a more holistic semantic representation of *consume*, it would be beneficial to examine the extent to which a learner is familiar with this representation. This could be exploited through ER as if provides exposure to words in a variety of contexts (Nation, 2013a). Each of the studies measuring vocabulary knowledge in Nakanishi's (2015) meta-analysis used instruments which measured one meaning sense of a word. It is possible that through ER, a learner learns three of the four meaning aspects of *consume* in Table 5, all but *to eat or drink*. That they learned the other three meaning senses goes unmeasured in this test, and were they to answer this question, they would likely answer incorrectly.

2.5 Measuring semantic knowledge of vocabulary

In what ways would it be possible to test multiple meaning senses of a word? One approach could be to test each word on all of its meaning senses; If a word had seven senses, seven questions would be administered. One problem with this is the rapid increase in questions needed to assess all meaning aspects of a word; In order to test 10 words, each with five meaning senses, 50 questions are necessary. This seems like a lot of test time for a small amount of words. A slightly different approach was used by Ishii and Schmitt (2009). In their study, Japanese EFL university students sat a *multiple*

meanings test. The test questions were made up of one English target word, and five Japanese options below it, two of which were synonymous with the English word. The remaining three options were semantically distant distractors. The objective of the test was to select the two correct translations of the English word. For example, the word *act* can mean *a thing done* as well as *a division of a play.* Accordingly, in the test, the test-taker would choose these two translations. This test is more practical than the example in Table 5, since each question measures two meaning senses of a word instead of one meaning sense. Yet, extending the *multiple meanings* test format to include all meaning senses of a word, as well as an appropriate number of distractors, leads back to doubts in feasibility.

Another way to assess semantic knowledge would be to have learners produce words included in the concept. But what is involved in conceptual word knowledge? According to Nation's (2013) word knowledge taxonomy (see Table 1), conceptual knowledge is a subcategory found in the meaning aspect of vocabulary knowledge, it is semantic in nature, and it consists of the words that the concept can refer to. One test which can provide the opportunity for a test taker to produce words which a concept refers to is a word association test. A word association test presents its test-taker with a set of prompts, one by one, and they are to provide the first word (or words) that comes into their head (Read, 2000).

This aspect of word knowledge seems to be broad since in theory any number of words could refer to a certain concept. Nation's (2013) taxonomy also includes an *association* subcategory of word knowledge which refers to the words that a certain word conjures, and words which could be used instead of another word. Whether or not word association tests assess this aspect of knowledge as well depends on the type of word association test used. In a controlled association test for example, a test-taker is instructed to choose a response from a particular semantic, grammatical, or conceptual category. As a result, the aspects of knowledge which are measured become more focused, meaning the *association* subcategory of knowledge may be excluded. In a free association test on the other hand, there are no restrictions put on the type of response

that is allowed. This means that in a free association test, including the *association* subcategory of knowledge is a possibility. It is also possible that a free association task includes the *form and meaning* aspect of knowledge, as well as the *collocations* aspect of knowledge. The type of word association test, either controlled or free, determines the aspects of knowledge that are measured. That word associations have the ability to assess various aspects of knowledge is one of the benefits of using this test format.

Another benefit of word associations is that they provide a means of assessing proficiency that is not based specifically on correct language performance (Wolter, 2002). In a test assessing correct performance with items similar to the example in Table 5, a test-taker chooses a correct answer, and if they do not select the correct answer they are incorrect. At the end of the test, the test-taker's total amount of correct answers is summed, perhaps converted to a percentage correct and then used as an indicator of proficiency. With word associations, in contradistinction, the process of assigning value is fundamentally different. For example, if a learner responds to the prompt *lorry* with the word *truck*, to what extent is this a correct or incorrect response? The association produced has a relationship to the prompt, but the relationship is not in terms of being a correct or incorrect associations. What is an appropriate means of assigning value to word association data?

2.5.1 Stereotypy of word association responses

In the early 1900s, when word associations were mainly used in the field of psychology to assess behavior and abnormality (e.g., Jung, 1910), valuation of word association responses was calculated by comparing pre-compiled lists of associations produced by a representative sample of people. Kent and Rosanoff (1910) for example, used word associations to diagnosis a person's sanity. They collected word associations from a group of people diagnosed as insane, and compared their associations to those produced by a group of people who were not diagnosed as insane. In order to determine the extent that a person's associations were pathological, each prompt/response pair was assigned a *stereotypy* value. This percentage score represented the proportion of people who

gave the same response to a certain prompt. For example, if the prompt/response pair *black-white* was given by two out of 10 people, it would receive a stereotypy value of 0.20 since 20% of the participants produced it. The higher the score, the more stereotypical a prompt/response pair was. The authors found that the sane participants produced a small number of unique responses to each stimulus, whereas those who were diagnosed as insane had the tendency to produce a wider variety of responses. This wider variation was thus deemed an indicator of insanity. Comparing associations against a set of standards, or norms, was picked up by second language vocabulary acquisition researchers, to see if it could be used as an indicator of L2 proficiency. Kruse, Pankhurst, and Smith (1987) investigated the extent to which this was true. In their study, 15 native Dutch speakers studying English provided up to twelve associations for 9 English words. To determine the proficiency of the speakers, the participants sat a 50-gap cloze test. The results revealed a modest correlation between stereotypy values and proficiency ($r = 0.547$, $p < 0.025$), and the authors conclude that the findings were "generally disappointing" (p. 152).

The less-than-ideal results in Kruse et al (1987) have been seen in other research as well (e.g., Randall, 1980; Schmitt & Meara, 1997), making it difficult to see the merit in using word association tests. One issue with the early word association studies is the lack of a systematic approach to target word selection. Kent & Rosanoff (1910), for example, do not specify how they arrived at the words they used other than mentioning that 66 of the words were borrowed from Sommer (1901) who in turn chose words based on their ability to access emotions. Wolter (2002) set out to address this by incorporating a strict selection system for the prompt words, in hopes of strengthening the claim that word association knowledge is a predictor of L2 proficiency. He selected words from the Edinburgh Associative Thesaurus (Kiss et al., 1973) a set of 8,400 native-speaker word association norms provided by 100 native English Speakers. A word was selected if:

- The primary response to a word was produced 15% of the time or less according to the Edinburgh Associative Thesaurus, and

- More than 60% of all of the responses given were provided by at least two native speakers (see Meara and Fitzpatrick, 2000, for a similar list of criteria for word selection).

This resulted in 20 verbs to be used as stimulus words in his study. Thirty Japanese EFL university students provided up to three associations for each of these 20 words. The associations were scored using a non-weighted system and a weighted system. Using the non-weighted system a response was given a point if it occurred in the native speaker norm list. Using the weighted system, an association was assigned points in accordance with the number of native speakers who also produced the response. To determine the extent that a relationship existed between the associations and L2 proficiency, the participants also sat a C-test. The results from using both the unweighted and weighted systems revealed correlations almost identical to those found in Kruse, et al (1987). Wolter concluded that word associations in a foreign language are not linked to proficiency in a clear manner.

2.5.2 A clang-syntagmatic-paradigmatic approach to word association data

Until the 1960s, no attention was paid to the *nature* or *type* of associations being produced in word association tests, instead focusing on norm lists used to compare NNS to NS data. With vocabulary knowledge now known to be a multi-dimensional construct (see Table 1 for some aspects of knowledge), the associations in language are highly organized, more so than the classical association theory put forth. During the 1960s, research began looking at new ways to classify associations. One system that emerged from this classified a response based on the nature of the semantic relationship it had to the stimulus word for which it was produced. Three main types of relationships emerged: syntagmatic, paradigmatic and clang relationships. A syntagmatic response represents a syntactic relationship. For example, in the sentence *No one should commit murder*, *commit* and *murder* have a syntagmatic relationship because they co-occur. The other main type of relationship words can have is paradigmatic. This relationship

tends to be more semantic in nature. For example, the words *car* and *boat* have a paradigmatic relationship since they are both types of vehicles, and can be substitutes for each other in a sentence. Finally, a clang association is typically related through phonology, bearing no clear association otherwise. For example, *van* and *can* could be considered clang associates. Early L1 research using this classification system investigated how these three categorizations changed as a function of age. Ervin (1961) gave 184 students a word association test of 46 words. The students in her study ranged from kindergarten (age five or six years) to sixth grade (age 12 or 13). She found that as age increased, students produced fewer syntagmatic and clang associations, and a greater amount of paradigmatic associations. This syntactic-paradigmatic shift was also evidenced in subsequent studies (e.g., McNeill, 1970; Palermo, 1971), giving clout to the idea that a paradigmatic shift equates to increases in semantic knowledge of a word.

The idea of being able to investigate semantic knowledge through word associations raised interest in the types of associations L2 learners would produce. Politzer (1978) for example, examined associations from learners studying French. The students provided associations for 20 French words (in French), and two days later provided associations for the English translation of the same 20 French words (in English). The results revealed that the students produced more paradigmatic responses in their L1 English than syntagmatic responses. The results also revealed a larger amount of syntagmatic responses in their L2 French than paradigmatic responses, and Politzer concludes that syntagmatic associations dominate in the early stages of foreign language learning, as L1 studies had shown to be the case for younger learners with less experience using their L2.

One issue with the clang-syntagmatic-paradigmatic classification system is the subjectivity inherent in determining the type of relationship a prompt-response pair has. With some pairs, for example the pair *dog-cat*, it is reasonable to assume a paradigmatic relationship. But what about the prompt-response pair "light-sun"? Is this relationship paradigmatic since the sun is something that gives off light, or is the re-

lationship syntagmatic, referring to the *sunlight* itself? It is not possible to know the answer without knowing the reason for producing the association. Another issue with the clang-syntagmatic-paradigmatic system lies in its ambiguity. For example, the two pairs *significant-important* and *tuatara-reptile*, can each be considered paradigmatic associations, yet the type of paradigmatic knowledge is different. The two words in the first pair, *significant-important*, are related since *important* is a defining synonym of *significant*. In the second pair, *reptile* is not a defining synonym of *tuatara*; Rather, *reptile* is a hypernym of *tuatara* since *tuatara* is but one of a group of *reptiles*. Fitzpatrick (2006) set out to address these two issues. She collected word associations for 60 prompt words from 40 native speakers of English and from 40 non-native speakers of English. She categorized her participants' associations into *form-based*, *meaning-based* and *position-based* associations using Nation's (2001) word knowledge taxonomy as a foundation. She also included a fourth category, *erratic* associations, which included false cognates and instances where there was no decipherable link. Within each of these four categories, associations were further sub-categorized, creating 17 individual classifications. This categorization system was used to examine the types of associations produced by the 80 participants. The results revealed a more detailed picture of the kinds of associations made by NSs and NNSs. Comparing the NS speakers' associations to the NNS revealed that while both groups tended to produce *meaning-based* associations, NSs had a tendency to supply defining synonyms (e.g., *generally - overall*) than NNSs. On the other hand, the NNSs were more likely to give what Fitzpatrick (2006) referred to as conceptual associations (e.g., *visual - color*).

The research carried out by Fitzpatrick (2006) helps to shed light on the nature of associations produced by NS and NNS, yet it is not without its limitations. One issue is the degree to which the stimulus words were known by the participants before taking the word association test. In the study, the stimulus words were taken from the Academic Word List (Coxhead, 2000), and participants were selected on the basis that they had "experience of working with academic English, at undergraduate and/or postgraduate level" (p. 128). The participants' receptive vocabulary level was measured using the

Eurocentres Vocabulary Size Test (Meara & Jones, 1988), a receptive yes/no test. Based on this information, it is not possible to know the extent that the participants knew the stimulus words since it is not mentioned if the target words were tested using the Eurocentres Vocabulary Size Test or not. A well-known word likely has more associations compared to a word which is unknown. Zareva and Wolter (2012) address this issue by including a familiarity scale in their study, adopting methods from previous research (e.g., Dale, 1965; Paribakht & Wesche, 1993). The scale had four levels of increasing familiarity:

1. no knowledge (*I have not seen this word before.*)
2. form knowledge (*I have seen this word before but I don't remember what it means*)
3. knowledge of meaning I (*I think this word means [ ].*)
4. knowledge of meaning II (*I know that this word means [ ].*)

Levels one and two do not require demonstration of word knowledge, while levels three and four require evidence of word knowledge (i.e., knowledge of meaning). While the authors note that very little is known of the relationship between familiarity level and overall organization of the mental lexicon, they focus only on the familiar words in their study. This issue of level of familiarity is one that can be examined empirically (e.g., Zareva & Wolter, 2012). A bigger issue to be addressed is with regards to the classification system itself. Meara (1996) mentions that trait models such as the one in Fitzpatrick (2006) may become impracticable since more traits are continually added or subtracted. This seems to be the case considering what started off as a 3-factor classification system (e.g., Kent & Rosanoff, 1910) has become a multi-dimensional construct (e.g., Fitzpatrick, 2006). This is not to say the system does not serve a purpose; rather the fact that traits can be added and subtracted means the system may not become a comprehensive model of vocabulary knowledge.

2.5.3 A semantic web of interconnected words

Word associations provide a means of assessing a larger portion of a learner's lexicon than tests which examine single words (e.g., Meara, 1996, 2009). A small amount of second-language research exists which views word associations as a highly organized, interconnected semantic web with words having links between them. Researchers investigating word associations as lexical networks believe that a highly sophisticated network is representative of advanced proficiency, meaning as language proficiency increases, so do learners' semantic networks (Meara, 2009; Wolter, 2002). Meara (2009) investigated word association data produced by 10 advanced NNS of Spanish. In his study, rather than asking for responses to prompts, the participants were given two words and told to create a chain of associations linking them. If NNS' L2 lexicon is less connected, then they should produce longer association chains than in their L1. The results however showed the opposite; The learners produced shorter chains in L2 Spanish than in their English L1. Meara concludes that the data is not promising, but emphasizes that viewing vocabulary knowledge as a structure instead of individual words can lead to new perspectives on depth of vocabulary knowledge. Most notable is with regards to the relationship between breadth and depth of knowledge. The traditional view sees breadth and depth of knowledge as a dichotomy, whereby each word a learner knows is known to differing degrees. Under this assumption, depth develops in individual words with limited effect on the other words. On the other hand, viewing depth of vocabulary as a connected web means that an increase in the number of words known (i.e., an increase of nodes in the network) changes the structure of the newly acquired word *and* pre-existing words. Viewed this way, a semantic network accommodates the newly learned aspect into its structure, changing the structure itself. Similarly, when an additional aspect of knowledge about an already known word is acquired, it too becomes part of the lexical network, and the network restructures accordingly.

Looking at the type of associations that are produced provides evidence of one aspect of the structure of the network. What they do not provide evidence for, however, is the

strength of the links between the words in the network. For example, both response-prompt pairs *dog-cat* and *dog-cougar* have *dog* paired with (a kind of) cat. However, it could be argued that *dog-cat* seems somewhat stronger a relationship than *dog-cougar*. Meara (2009a) investigated the strength of associative relationships through *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, reports the strength of the association on a scale from one (weak association) to four (strong association). In this way, it becomes possible for the prompt-response pairs *dog-cat* and *dog-cougar* to have different strengths.

The addition of a measure of association strength, as in *V_links*, creates new issues to be addressed. One issue noted by Meara (2009) is that non-native speakers often select word pairs which are not selected by native speakers, perhaps to be expected given the different types of associations produced by NNS compared to NS (e.g., Riegel & Zivain, 1972). To circumvent this issue, a database of NS responses was compiled to be used with *V_links*. Each prompt-response pair produced by a test-taker becomes a valid pair if at least two NSs selected the same prompt-response pair. The issue here however relates back to the idea of using native-speaker norms, discussed earlier. Another issue with the inclusion of a test of association strength deals with the strength of associations reported by the learners. Meara (2009) found 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).

2.5.4 An interim summary of L2 word association research

In the end, it seems that lack of theoretical foundation has hindered research exploiting word association data (Wolter, 2002). For example, there is not a strong theoretical explanation for a learner to change their response patterns to be more native-like as their proficiency increases. Fitzpatrick (2006, 2007) attempts to address this lack of theory by creating a categorization framework which uses Nation's (2013) well-established

taxonomy of word knowledge as its core. However, this still leaves the subjectivity of the classification open for interpretation. Research attempting to investigate the structure of the network created by the word associations has also met with difficulties including the differences between NS and NNS responses. Limited work has been conducted into association strength, with *V_links* (Meara, 2009; Meara & Wolter, 2004) one of the only examples, and *V_links* is not without issues. As a result, researchers have not been able to agree on how word association patterns might be interpreted, and accordingly have not been able to show consistent, uncontentious findings (Fitzpatrick, 2007).

2.5.5 A novel approach to word association analysis: Latent Semantic Analysis

The research undertaken in this thesis uses an alternative method to measure the strength of associations produced in a word association test, and how this knowledge develops, through Latent Semantic Analysis. Latent Semantic Analysis (LSA) is a theory of meaning based on knowledge induction, 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 provides an computer-automated method for determining the strength of the relationship between two pieces of text (e.g., words, sentences, paragraphs, etc) by using a corpus of text to determine the association strength between each of the words in the corpus (further explanation of the method used in LSA is discussed in section 3.8.2). LSA provides a theory that says objects and experiences that are met near each other in time become associated. Of course, the human mind goes far beyond this temporal level of association and takes "the billions of local contiguity relations and fits them together into an overall map, a semantic space that represents how each object, event, or word is related to each other" (Landauer, 1998, p. 162). It somehow takes all of the dimensions which comprise the semantic map and reduces it to a more manageable size. LSA takes account of this by capturing the deeper semantic and associative structures in a way that simple co-occurrence data cannot, through a process of dimension reduction (Gunther, Dudschig, & Kaup, 2016). The resulting semantic

space is used to determine the strength of association between two words by examining their proximity in the semantic space. Words with a stronger association will tend to be closer to each other in the space, and words with a weaker association tend to be further apart (Deerwester, et al., 1990). For example, in Table 5, with the prompt *consume* and four possible meanings, the four options represent different meaning senses of *consume* since all of them can mean *consume* depending on the context. In other words, *consume* refers to using, eating or drinking, spending, and destroying. From LSA's point of view, the correct option will have the strongest association with the word *consume*, while the others have weaker associative strengths. LSA will choose the option with the highest value as the correct option. In the semantic space then, the correct answer will be closer to the word *consume* than the other options, which are further away. That is, each of these meaning senses, while all referring to *consume*, do so to differing strengths. This is the theory which motivates LSA.

In simple terms, LSA computes a (similarity) score using arbitrary lengths of text (e.g., a word) against a corpus. The similarity score represents the semantic strength that the word has in the corpus. Words with higher scores tend to have higher strength in the corpus than words with lower scores. For example, using a corpus consisting of a book about war, the word "battle" would likely have a higher similarity value than the word "carpet". This is because the corpus (the book about war) is not likely to include discussion of carpets. Going one step further, LSA is flexible in that it can compare arbitrary lengths of text. To that extent it becomes possible to compare for example "battle" against "carpet", "tank", and "gun" at the same time. This comparison would likely give a higher similarity score than the score computed between "battle" and "carpet" since the additional words "tank" and "gun" have a stronger relationship to war than "carpet" by itself. In this way, it becomes possible to determine the semantic similarity that words have in a corpus.

The workings of LSA can also be understood with an example. Imagine a language learner whose parents delivered milk in a lorry to the neighborhood when the learner was young. Everyday, the learner traveled around the neighborhood with their parents,

interacting with their neighbors as the milk was delivered. Sometimes, the learner would see the neighborhood's children and would play with them while their parents delivered the milk. These childhood experiences help to shape the learner's understanding of what a lorry is; namely, a truck used to deliver milk around the neighborhood. Accordingly, words such as "lorry", "milk" and "neighborhood" would have a stronger relationship to each other for this learner than words such as "war" and "death", since the child encountered a lorry in an environment with "milk" and "neighborhood". In other words, the input which the learner was exposed to shapes their understanding of the world around them, including their understanding of the relationships between words in their mental lexicon. This is analogous to the way that LSA works.

Theory is not the only motivating aspect of LSA; it has also shown merit in practice. Landauer and Dumais (1997) for example investigated the degree to which LSA was able to correctly choose the correct answer from a set of multiple-choice items taken from the Test of English as a Foreign Language, provided by the Education Testing Service (ETS; Landauer & Dumais, 1997). The authors used LSA to compute the strength of the relationship between the stem word and each of the multiple choice options. The option with the strongest relationship to the stem word was selected as the answer. The percentage correct using LSA was then compared to the average percentage correct from a sample of nonnative English speakers who took the test. The results revealed that the percentage correct from LSA and for the NNS were almost identical (64.4% and 64.5%, respectively). In other words, LSA scored as high as the test-takers did on the multiple choice test, a percentage high enough for admission to some U.S. universities.

LSA has also been used to assess development of cohesion over time. Crossley, Salsbury, McCarthy, & McNamara (2008) used LSA to examine the extent to which learners' writing develops cohesion as a function of time spent studying English. They hypothesized development in L2 vocabulary knowledge is in part motivated by the strength of a learner's semantic network and as development occurs, stronger associations and interconnections between words are created and consolidated. To test this hypothesis

they analyzed spontaneous speech data from six L2 English learners who were enrolled in an intensive English program at a U.S. university. The data was recorded fortnightly and was collected over a period of one year. The cohesiveness of the learners' speech was determined using LSA, with the assumption that as L2 vocabulary knowledge develops, learners exploit the strengths of semantic networks to produce stronger associations and interconnections in their speech. In order to determine the degree of cohesiveness, the authors used LSA to compare adjacent utterances in the students' speech data. These student utterances were analyzed using the computational tool Coh-Metrix, which measures cohesion at differing levels of discourse (Graesser et al., 2004). LSA values were derived from the college-level TASA corpus (see http://lsa.colorado.edu/spaces.html for detailed information regarding the composition of the TASA corpus). Due to the fact that their study was based on spoken utterances and not written text, LSA paragraph to paragraph values were analyzed. This was because any sentence punctuation for the spoken utterances would be artificial, and second because many of the learners' utterances were short and the authors felt that these short utterances would not provide proper lexical coverage.

To determine the extent that cohesion developed, the authors conducted a repeated measures Analysis of Variance (ANOVA) using the LSA paragraph to paragraph results. The results of the ANOVA revealed that the LSA values increased over time, meaning that the learners' speech became more cohesive as they gained experience with the language. In their conclusion, the authors note that LSA can be used as a model to approximate the development of semantic relations in second language learners. The authors note that future research in L2 lexical development would benefit from using different semantic spaces, ones designed from L2 input. This space would consider L2 reading texts for example, allowing for further investigations of L2 lexical networks and lexical development. Finally, the authors note that learning vocabulary through techniques which include interaction with language is a more valuable approach to lexical development than rote memorization because word meaning are individualistic and thus based on a person's experience. This means that "the more experience one

has with a word within context, the better the chance that connections between that word and other words will develop" (p. 140).

2.6 Summary

Research throughout the years has shown extensive reading to facilitate language development, including vocabulary learning. Upon closer inspection however, a number of issues arise warranting further research. These issues relate to lack of attention given to additional activities, especially output activities, unbalanced time on task between groups, and differing initial learning potential. Some of these issues may stem from the idea that ER should be a solitary activity, separate from any activity which is planned in a syllabus. However, activities which promote learner interaction offer important opportunities for language development, and the research in this thesis takes account of these interactions and examines the extent to which it facilitates learning. Other issues in ER research stem from the limits imposed by using tests which assess individual words. Word association data seems to be a promising area for examining how the links between words develop as a function of ER, whereby learners are presented with opportunities for repeated exposure to words in meaningful and diverse contexts. This thesis utilizes Latent Semantic Analysis to further investigate the degree to which ER facilitates development of lexical networks in an L2.

2.7 Research questions

In order to address the issues summarized in the previous section, the following broad research questions will be investigated:

1. How do different approaches to ER affect L2 vocabulary development, specifically an ER-only approach and an ER-plus approach?
2. How do different types of interaction following ER affect L2 vocabulary development, specifically spoken interaction and written interaction?
3. What additional aspects of language do learners focus on during the interactions?

3 Methodology

3.1 Introduction

There were a number of gaps mentioned in the previous chapter, gaps which were identified while reviewing the L2 vocabulary learning research which utilized extensive reading. These gaps included:

- unbalanced time on task between conditions,
- differences in vocabulary learning potential between the groups compared,
- the lack of attention to supplementary activities, and
- limitations in measuring of learning.

The current study attempts to ameliorate these issues by:

- ensuring all groups spend the same amount of time on task,
- counterbalancing for participant proficiency to equalize learning potential,
- measuring the extent of vocabulary learning in supplementary post-reading activities, and
- implementing a novel approach to measuring vocabulary knowledge.

This chapter details the methodology selected to address these gaps. The next section summarizes these issues and how they were addressed in the current study. Section 3.2 explains the rationale for the research design implemented in the current study. Subsequently, section 3.3 describes the research site and the participants for the study. Sections 3.4 through 3.6 describe the materials used in the study. The methods used to collect and analyze the data are explained in sections 3.7 and 3.8, respectively. After detailing the procedures in section 3.9, section 3.10 explains the pilot study conducted to assess the effectiveness of the research design. The chapter then digresses, presenting interim results which resulted in an additional phase of data collection. The justification for this second collection phase, as well as the research design modifications are detailed in section 3.11. Finally, section 3.12 summarizes the chapter.

3.2 Research design rationale

Section 2.3.2 in the literature review chapter explained that vocabulary knowledge is multidimensional, with at least 18 aspects of knowledge that can be learned about a word (Nation, 2013a). In order to investigate this knowledge, the current study uses word association data to measure the lexicon of L2 learners, i.e., the connected web of semantic links in the participants' minds. Word associations have the potential to test a large amount of this lexicon, and can result in large amounts of data. This data set can be measured quantitatively to examine certain patterns, to ultimately suggest (or deny) causal relationships between the treatment and learning. At the same time, this quantitative data set cannot explain the reasons behind the word association patterns. One way to shed light on the logic behind these word association patterns is to ask the producer of the associations (Fitzpatrick, 2007). To capture both the large amount of language data generated through word associations, and to gather data concerning the motivations for the associations produced, as well as other qualitative data, the current study adopts an explanatory mixed methods design (Creswell, 2012). In this way, the qualitative data collected helps to explain patterns found in the quantitative data.

3.3 Background and Participants

The study took place in the English Language Institute at Victoria University of Wellington. The English Language Institute offers an English Proficiency Program, an English for Academic Purposes course for students of an intermediate English level or above (i.e., IELTS 4.0). The program is a full-time, 12-week intensive course which leads to a certificate of proficiency in English. The course develops reading, writing, speaking and listening, preparing students for participation in a New Zealand university academic community, at both undergraduate and graduate levels. In the program students are placed into classes based on their proficiency, which is determined by a placement test administered at the beginning of the course. The placement test contains five components: A receptive vocabulary size test, a C-test, a dictation test, a

writing test, and a questionnaire. The scores from the vocabulary size test, the C-test and dictation test are summed to give an overall score which is used in determining which level a student should be placed. In addition to these five components, the Vocabulary Levels Test (Nation, 1983; Schmitt et al., 2001) is administered in the first week of the term to help determine the kinds of words students need assistance with, as well as the degree of work needed (Nation, 2013a). The maximum class size is 16 students.

Four intact classes (n = 48) volunteered to participate in the study; 28 were female and 20 were male. The students came from a variety of countries, including China (30), Japan (7), Columbia (3), Iran (1), Myanmar (2), Saudi Arabia (1), Thailand (1), and Vietnam (1) (see Table 6). The participants were on average 22 years old (SD = 3.6 years). The students were intermediate to upper intermediate proficiency, according to their English Proficiency Program placement test scores. Intermediate students are those learners who score between 100 and 130 points out of a possible 260 on the placement test (English Proficiency Program coordinator, personal communication, December 8, 2015). The participants' average score on the placement test was 126.5 (SD = 9.1).

The reason why intermediate-level classes were chosen for the study is because the teachers of these learners were advocates of ER. Two of Day & Bamford's (2002) principles focus attention on the teacher in implementing a successful ER program. The first principle centers on the teacher's role in orienting and maintaining learner progress, guiding students to help them receive maximum benefit from the program. The second principle states that the teacher should become a role-model for their students. This involves active membership, with the teacher engaging in ER with their class. These two principles require a motivated teacher, an ER advocate who believes in the power of reading. The teachers who participated in the study were believers.

Country	Students
China	30
Colombia	3
Iran	1
Iraq	1
Japan	7
Myanmar	2
Russia	1
Saudi Arabia	1
Thailand	1
Vietnam	1

Table 6: Participants country of origin

3.4 Graded Readers

All of the participants read the same five graded readers over the course of the study. Graded readers (GRs) are stories composed with simplified language, in part by controlling for the amount of unique words used to write them. The GRs were chosen based on criteria that coincides with Day & Bamford's (2002) principles. First, it was important to maximize the likelihood that the GRs would maintain reader interest so that they would enjoy reading. To increase the chances that the GRs would be interesting for the students, only those GRs which were Language Learner Literature Award winners were selected. The Language Learner Literature Award is given by the Extensive Reading foundation (http://erfoundation.org/) to books for their overall outstanding quality and likely enduring appeal. These books have interested a variety of learners, evidenced in their award-winning ability, and so they were likely to maintain reader interest.

The next factor that was taken into account when choosing the GRs was their difficulty level, determined by coverage rate or the amount of words which are known

in the books. Research has shown that in order for adequate comprehension of a piece of text, at least 95% of the words in the text should be known (Hu & Nation, 2000; Schmitt, et al., 2011). In addition to providing adequate comprehension, this percentage creates conditions for incidental vocabulary learning to occur since enough words will likely be known in a word's surrounding context to facilitate contextual learning (Nation, 2013a). In order to determine the appropriate GR level, it was necessary to find out how many words the participants knew. The scores from the Vocabulary Levels Test (Schmitt, et al., 2001) given during the first week of the trimester were used for this purpose, however at the time that the GRs were purchased, the participants had not yet taken the Vocabulary Levels Test since the trimester had not begun. Instead of the participants in the current study, three year's worth of previous EPP cohorts' VLT scores were used. The intermediate-level students in these previous cohorts scored 71% on average in the second 1,000 most frequent word band. This amounts to a vocabulary size of around 710 words for this band, since each word in the test accounts for 100 words. Read (1988) found that the scores on the VLT were such that knowing lower-frequency words tended to imply knowing higher-frequency words, and if this is true then it was likely that the students knew at least 710 words in the first 1,000 word band as well. The graded readers were thus chosen to be well within this level.

Table 7 displays information about the GRs selected for the study, in the order that they were read. The right-most column of the table shows that the number of words increase as the reading progressed. This was a purposeful decision to increase the ecological validity of the study. Beglar, Hunt, and Kite (2011) found that when given the freedom to choose GRs, their students tended to choose books at increasing word levels. This authenticity was incorporated into the current study by increasing the amount of words in each book as the participants progressed through the reading. The first two GRs were chosen to have the same amount of headwords to accustom the participants to the reading treatment.

Title	Publisher	English	Pages	Headwords
Jojo's Story	CUP	British	46	800
Dead Cold	CUP	American	48	800
Billy Elliot	Penguin	British	49	1200
Land of My Childhood: Stories from South Asia	OUP	British	72	1400
A Kiss before Dying	Macmillan	American	86	1600

Table 7: Graded readers used in the main study

3.5 Target words

Eighty target words were measured over the course of this study. Sixty of the target words occurred in the five GRs, while twenty did not. The target words were selected based on a number of factors. The first factor taken was frequency of occurrence of the words in the GRs. Frequency of occurrence is a major determinant of word learning (e.g., Horst, Cobb, & Meara, 1998; Nation, 2013a) and so should be a factor considered when assessing vocabulary knowledge through ER. Due to the importance of frequency of occurrence on vocabulary learning, four different frequency bands were created based on word *type*. 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. In this way, words at different frequency bands, and also different levels of familiarity, could be measured.

Another measure of frequency of occurrence was factored into target word selection. This was the frequency of occurrence of the words in the English language, according to the British National Corpus. As seen in column three of Table 7, the majority of the GRs were written using British English. As a result, it was deemed more appropriate to use a corpus of British English than another corpus (e.g., the Contemporary Corpus of American English). Nation's (2004) British National Corpus frequency lists were used

to determine which frequency band of the target words. Scholars tend to agree that high-frequency words are those in the first two or three 1,000-words lists, and for the current study a word was considered high-frequency if it occurred in the first 1,000-word frequency band. A word was classified as mid-frequency if it occurred within the fourth to the eighth frequency bands. Finally, a word was labeled low frequency if it occurred in the ninth frequency band or beyond (Nation, 2013a).

After the list of potential target words was refined using the frequency information, the length of the word, in characters, was taken into consideration. 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). Due to the controlled language used in the GRs, it was not possible to limit 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 deemed appropriate for selection. In this way, extremely short or extremely long words, which may draw special attention from the participants while reading, were excluded.

The next factor used to select target words was the range of occurrence, or the number of GRs that a word occurred in. Some research suggests that evenly spaced encounters with a word increase the likelihood that the word will be remembered (Nation, 2013a). As a result, the twenty high-frequency words occurred in all five of the GRs. The mid-frequency words occurred in approximately two GRs, although the restricted nature of the language used in GRs made it impossible to find enough words which met this criteria meaning some words occurred in only one GR. The low-frequency words occurred in only one GR.

There were four more factors taken into account when the target words were selected. These four factors were retrieved from the Medical Research Council psycholinguistic database (http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm), a database containing over 150,000 words with up to 26 psycholinguistic attributes for each entry. The first factor was word familiarity, or the degree that a word is "seen, heard, or used every day" (Gilhooly & Logie, 1980). Lower values mean lower

familiarity, and higher values mean higher familiarity. The more familiar a word is, the less difficult it tends to be (Leroy & Kauchak, 2014). Potential target words were those which had a familiarity rating between 300 and 600.

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). Any word with a meaningfulness rating between 300 and 600 was eligible to be a target word.

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). Any word with a concreteness value between 300 and 600 was a valid candidate to become a target word.

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), and only those words with an imageability value between 300 and 600 were deemed appropriate to become a target word.

The controlled language in the GRs made it impossible to select words which had exactly the same value for these four psycholinguistic traits. To that end, 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, and started with as narrow as range as possible, expanding only until enough potential words were available to have 20 words in each of the frequency bands. Some of the resulting words did not have information available in the Medical Research Council database (e.g., concreteness values for the low-frequency words), and where possible this was avoided, but it was not possible to avoid it completely. The blank spaces in the *low-frequency* column in Table 8 exist because no values were available for these words in the MRC database.

In addition to the sixty target words described above, twenty additional words were

selected which did not occur in the GRs (henceforth "off-words"). The purpose of the off-words was to determine the extent that outside learning may have occurred during the reading treatment. As mentioned in section 3.3, participants in the study were all enrolled in an intensive academic English program, and were exposed to English on a daily basis. To that extent, 20 words were chosen to be similar in frequency of occurrence in the British National Corpus as the low-frequency word band. Priority was given to words for which complete information was available in the Medical Research Council psycholinguistic database. This explains why in Table 8, which gives the average values (and standard deviations) for the factors described above, the off-words have higher values for the four psycholinguistic factors.

	high-frequency	mid-frequency	low-frequency	off-words
Frequency_in_GRs	59 (35)	14 (7)	3 (2)	0 (0)
Range_in_GRs	5 (0)	1 (0)	1 (1)	0 (0)
Length_in_characters	4 (1)	6 (1)	6 (1)	6 (1)
BNC_frequency_band	1 (0)	5 (1)	11 (3)	11 (3)
Familiarity	576 (35)	240 (7)		377 (0)
Concreteness	388 (63)	272 (288)		460 (140)
Imageability	445 (69)	273 (287)		443 (148)
Meaningfulness	462 (29)	179 (210)		331 (120)

Table 8: Mean (SD) values of target word characteristics by frequency band

3.6 Supplementary ER activities

The participants were assigned to one of two groups. The first group was an *ER-plus* group, in which participants completed the Say-it activity (Macalister, 2014), a post-reading small-group discussion task after finishing each GR. During the 15-minute task, the students formed triads and took turns choosing prompts from a three-by-three grid (nine prompts total) for another group member to discuss. Macalister (2014) explains

that these discussion prompts can be designed to serve a variety of functions, including

- recalling information from a story,
- making inferences based off of a story, and
- drawing on personal experience (p. 29).

Since the current study investigates vocabulary development, it was desirable to provide opportunities for the students to use the target words in the GRs. If the discussion prompts are designed so that the students discuss the characters and events which occurred in the GRs, it is likely they will use the vocabulary in the book to explain the characters and events which took place in the books. On the other hand, if the prompts are designed for learners to infer meaning, or to draw on their own experiences (the second and third functions listed above), students can utilize other vocabulary which was not necessarily in the GRs, excluding the target words from their production. To that extent, the Say-it activity prompts were designed solely to recall information from the story.

Each prompt places a learner in the role of a character and asks them to describe a certain event from the story. This makes the Say-it activity non-personalized since the learners are becoming the character and not discussing their own life (Ellis, 2003). For example, the following prompt occurred in the first Say-it activity: *You are Doctor Nicky. Talk about the Children's house and your work there.* One feature of this type of prompt is that the number of answers which are possible are limited; Doctor Nicky and her work at the Children's House comprises only so many things. This *closed* feature of the task, i.e., having a limited amount of possible answers, has been found to push learners to overcome linguistic difficulties to make themselves understood (Ellis, 2003; Long, 1989).

Designing the Say-it activity prompts was a balancing act between communicative value and lexical focus. It was desirable to design the prompts so that they would promote rich dialogue, engaging the learners in meaningful conversation. However, it was also desirable to provide learners opportunities to use the target words. In the

end, communicative value took priority since the Say-it activity was implemented as a task, requiring learners to use language, with an emphasis on meaning, to complete an objective (Bygate, Skehan, & Swain, 2001). The prompts were thus designed around main events or themes which ran through the GRs.

The Say-it activity was completed in two modes: a spoken mode and a written mode. This was for purposes of comparison, to examine the extent to which each mode facilitates L2 vocabulary development. There is a lack of research investigating the effects of task mode on language development (Joe, 2006). In one of the few studies comparing the two modes, Brown, Sagers, & Laporte (1999) collected and analyzed spoken and written dialogue journals that their L2 English university students produced. The dialogue journals were conversations between a student and their teacher. The students wrote in their journal four days a week over a four-month period. Each week the teacher collected the journals, responded to students' entries, and then returned the journals to the students. The results from the study revealed that a greater number of words were possibly acquired in the spoken mode but that the percentage of words possibly acquired between the two modes was not different. This is because more words were produced orally due to the slower process inherent in writing. Another finding in their study was that the mode in which learners encountered a word was likely to influence their subsequent production of that word. Specifically, they found that it was easier for learners to go from a spoken mode to written production than from a written mode to spoken production.

One confounding issue with Brown, Sagers, & Laporte's (1999) study was lack of assessment of initial lexical knowledge. This means that any gains in vocabulary were speculative at best (hence the term *possibly acquired* words, above). Not much is known about the relationship between mode of communication and vocabulary development. Thus, to determine the extent of task mode on vocabulary learning, the Say-it activity was completed orally by half of the ER-plus group and in writing by the other half of the ER-plus group. The participants in the spoken mode completed the tasks orally, and the written group completed the Say-it activity using Google Docs, a synchronous platform

for computer-mediated communication. Using Google Docs, all members of a group can access the same document in real time. This means that when one group member types something in the document from their computer, all members can see the typing as it is happening from their respective computers. In this way a dialogue can develop whereby the Say-it activity prompts are discussed through text-based communication.

Another feature of Google Docs is that it automatically saves document versions, and these versions can be viewed separately. This means that the dialogue is saved at different points, and by comparing these points, it becomes possible to examine the differences in each version. In particular this was used to locate the editing of a misspelled word, or other focus-on-form episodes (see section 2.4.1 for an explanation of these episodes). To make the search for differences easier, Google Docs color codes changes to each version saved by comparing the new version with the most-recently saved version, and highlights according to which learner produced the test. It thus becomes possible to track the group members' contribution to the tasks.

To familiarize the ER-plus group with the nature of the Say-it activity, a two-minute video was created and used as a pre-task activity (Willis, 1996). The video consisted of three L1 English speakers demonstrating the Say-it activity, using the supplementary materials found in Macalister (2014). Observing others performing a task in this way can help reduce the cognitive load on the learners and increase performance (Ellis, 2003; Skehan, 1996; Willis, 1996). Both spoken and written modes were recorded, and the video was shown to all participants in the ER-plus group before the first Say-it activity.

The second group in the study was an ER-only group. This group was included as a comparison group, reading the same five GRs as the participants in the ER-plus group. One of the issues mentioned in the previous chapter is the lack of control for time on task (see section 2.4.2). To control for this, while the ER-plus group was completing the Say-it activities, the ER-only group, read a chapter from a short story book. Seventeen short story books were selected for this purpose, and all were borrowed from the Language Learning Center at the university where the research took place. Due to the small number of short stories available at the Language Learning Center,

it was not possible to control for the factors described in section 3.4 when selecting these books, however only those books were chosen which were between 800 and 1600 headwords, within the range of headwords making up the GRs. The reason short-story books were used, as opposed to another graded reader for example, is because the short story books were comprised of stories (one per chapter) that could be read in a short amount of time (i.e., 15 minutes). The ER-only group read the short story books only when the ER-plus group completed the Say-it activity, approximately once per 10 days. It was felt that by reading a novel-type GR, the students in the ER-only group would not be able to follow the story, continually having to remember where in the story they read the previous week. On the other hand, reading a short story would allow for 15 minutes of reading of an entire story meaning the learner would not have to recall what they had previously read, since they would be starting a new story each time they read the short story books.

3.7 Data collection

The following instruments were utilized to collect data providing answers to the research questions posed at the end of the previous chapter. Each section describes the test used, and the justification for it.

3.7.1 Self-report test

Since a fairly large number of target words were selected (n = 80), it was desirable to have a test format which could assess these words in an efficient manner, so that the participants would not become fatigued from the test taking too long. It was also desirable to determine the initial state of target word knowledge so that development could be tracked. As discussed in section 2.3.2, the Vocabulary Knowledge Scale (Paribakht & Wesche, 1996) was created for measuring the initial stages of knowledge. However, the test is not without it skeptics and so caution was taken before deciding whether to use it or not. Bruton (2009) carried out a critical review of the Vocabulary Knowledge

Scale to determine its appropriateness as a lexical measurement tool. He found that since the test included measures of both receptive and productive knowledge, and the "redundant and dispensible" (p. 295) numerical assessment involved in assigning scores for each category was arbitrary, the knowledge scale may not be representative of the learner's actual level of understanding. Taking heed of this critique, the Vocabulary Knowledge Scale was not used, but rather a three-level self-report test was created using other research as a framework (e.g., Dale, 1965). The resulting test assessed each target word at the following three levels of understanding:

- no knowledge,
- form knowledge, and
- meaning knowledge

These levels were thought to be less controversial than the level scheme in the Vocabulary Knowledge Scale. They also represent three basic levels of vocabulary knowledge, and as such were deemed a good measure of initial knowledge.

In the test, each target word was presented one-by-one to each participant, and below the word the participant was to select their level of familiarity with the word from the following options:

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

One critique of self-report tests such as this is that they do not require demonstration of knowledge (Meara, 2009). As a result, there is no way to know for sure whether the learner is making a meaningful assessment of their own knowledge (Brantmeir, 2004). However, the research in this area is mixed, and some scholars have found that the nature of self-report tests are dependent on linguistic ability. Blanche (1988) for example, conducted a meta-survey of eighteen studies incorporating self-report measures and found that more proficient L2 learners tended to underestimate their linguistic ability, while lower proficiency students overestimated their knowledge (e.g., Russell et al.,

1978). Laufer & Yano (2001) found similar trends in their study. Their participants, 106 students studying English in China, Israel, and Japan, self-reported knowledge of twenty target words contained in a piece of text. Their results revealed all students tended to overestimate their knowledge, not just the lower level learners as in Russell et al., (1978). Laufer and Yano (2001) also found a relationship between the degree that a learner overestimated their knowledge and ethnicity; Israeli students overestimated to a greater degree than the Chinese learners, and the Japanese learners were the most modest. On the other hand, some research has shown self-report tests to be more promising. Paribakht and Wesche (1996) for example, found a strong correlation between reported knowledge and demonstrated knowledge, and Horst (2000) found that her test-takers accurately provided translation equivalents 80% of the time for words reported as known. In short, the aforementioned research suggests that self-report measures can be a reliable measurement tool, and coupled with its simple design allowing for measuring a large number of words in a small period of time, it was deemed appropriate for the current study. The self-report test was given to the participants during the pretest and during the post-test. The order of the target words was randomized for each student, and all students reported knowledge of all 80 target words.

3.7.2 Word association test

The target words which a participant self-reported as known at either form or meaning levels became prompts for the word association test. If a learner reported never having seen a word before, they would be unable to give a meaningful association, and for this reason, the target words which learners were unfamiliar with were omitted. The word association test was given twice, once during the pretest and again during the post-test.

As mentioned in section 2.5.3, word association tests are able to investigate a large portion of vocabulary (Meara, 2009) and it was desirable to take advantage of this feature. Scholars seem to be in disagreement about the superiority of single-response versus multi-response association tests. Some scholars speculate a single-response test allows for a more accurate measure since responses given quickly to a prompt are

said to be more indicative of the structure of a learner's lexicon (Clark, 1970). In addition, some scholars note that a single-response format eliminates the possibility of chaining responses, whereby each subsequent response produced is an associate of the previous response, not the prompt word (e.g., Fitzpatrick et al, 2013). Other research, in contrast, has shown that a multiple response test is a valid format for assessing word association knowledge (e.g., Kruse et al, 1987; Randall, 1980). To that extent, the word association test in the current study adopted a multi-response format, with learners providing up to five responses for each target word. In this way, it was possible to gain a more comprehensive picture of the state of learners' lexicons than if a single-response format was used.

Another issue which was confronted was whether to apply a restriction on the type of response (e.g., synonomous responses only), or to adopt a free association format which permits all response types. Since a wide variety of responses was desirable, a free association task was used. It should be noted that free association tasks are more likely to generate prompt-response pairs which are not semantically-related, but this does not mean that there is less value in them since they allow for a wider variety of associations to be produced (Zortea et al., 2014).

A third issue tackled regarding the format of the word association test was whether to use a recognition or recall format. In a recognition format, a test-taker chooses associations from a list of possible associates. Read's (1993) Word Associates Test is one example of a recognition-format word association test. In his test, the test-taker chooses words which have either a paradigmatic or syntagmatic relationship to the prompt. This format relies on the associates to "trigger knowledge in the test-takers mind of particular semantic aspects of the stimulus words" (Read, 2000, p. 186). An example test item is illustrated in Table 9.

Sudden	beautiful	quick	surprising	thirsty	change	doctor	noise	school

Table 9: Example of Read's (1993) Word Associates Test

In contrast, a recall format requires the test-taker to produce words which they themselves associate with a prompt instead of choosing from a selection of words. One example of a recall format word association test is Lex30 (Meara & Fitzpatrick, 2000). The test requires the test-taker to produce responses to a list of prompts taken from Nation's (1984) most-frequent word band.

In sum, the word association task employed in the current study was a multi-response, free association task in which participants provided responses for each target word that they were familiar with to some degree. The participants produced up to the first five words that they thought of when they thought of the target word, and in this way the test was a recall association test, assessing productive associational knowledge.

3.7.3 Focus-on-form questions

Section 2.4.1 discussed output tasks and how they allow for form-focused episodes to occur, facilitating language learning. To investigate the extent that the episodes which occurred during the Say-it activity facilitated lexical development, participants were assessed on the linguistic knowledge arising during the episodes. There is no way to know with certainty which linguistic items will manifest themselves in these episodes, meaning it is impossible to pretest them (Swain, 2001). However, the fact that a question was raised about a linguistic item, or an error was made, indicates learner difficulty with an item, meaning further consolidation of that item may be necessary (Ellis et al., 2001; Swain, 2001). In other words, the occurrence of the episode itself can be considered a type of pretest since it demonstrates lack of knowledge of a linguistic item (Loewen, 2005). A participant was assessed on those linguistic items for which they displayed a lack of knowledge (Loewen, 2005; Swain & Lapkin, 1998). In this way, the questions were individualized for each student. In addition, the questions created using episodes which focused on a target word were given to all of the participants since pretest data was available for these words.

The format of the questions was developed depending on the nature of the episode, adopting Loewen's (2005) question format. This resulted in three types of questions.

The first question type, *suppliance*, was used for meaning-based and spelling-based episodes. For this item type, a student was required to define a word (in the case of meaning), or produce the spelling of a word. The second item type, *correction*, dealt with grammar-based episodes. For this question type, a student was asked to fix a grammatical problem in a sentence, one that was taken from the form-focused episode. The third question type, *pronunciation*, assessed the pronunciation of a word. These items were assessed at the beginning of the post-test interview, by having students read a sentence with the linguistic item embedded in the sentence. Immediately followed the sentence was the word being assessed and the test-taker read this word as well. Table 10 summarizes these question types.

Question type	Explanation	Target language
Suppliance	A student produces linguistic information	meaning spelling
Correction	Ss fix a linguistic problem	grammar
Pronunciation	Ss read a sentence with the TW embedded, and again decontextualized	pronunciation

Table 10: Form-focused episode question format (adapted from Loewen, 2005)

3.7.4 Post-test interview data

In order to gain a qualitative understanding of the ER treatment, the participants completed a one-on-one interview with the researcher (Henriksen, 2008). The interview also provided an opportunity to assess pronunciation-centered focus-on-form episodes (Loewen, 2005). Five areas were addressed during the interviews. The first set of questions probed the aspects learners found (un)appealing about the treatment, and whether or not they felt the treatment was conducive to learning (Foster & Ohta, 2005). The following questions were asked:

- Did you enjoy the reading? Why?
- Did you have a (least) favorite book? Why?
- Do you enjoying reading in your L1/L2? Why?

The second area addressed the appropriateness of the GRs. The questions asked were:

- Were any of the books too difficult? Which ones? Why were they difficult?
- Where did you read when you took the books home?
- Have you read any of the books before?
- Have you seen any movies made based on the books?

The third area focused on the reading process that the participants went through, with the purpose being to ask about individual reading habits. Questions asked included:

- Did you finish reading all of the books? If not, why?
- Was it difficult to read at the pace on the reading schedule?
- When you read, did you take notes on the books, or just read?

The fourth area addressed in the interview was the Say-it activity. The purpose of this section was to get a view of the task from the participant's perspective. The following questions were asked:

- Have you ever done an activity like the Say-it activity before?
- Describe what happened during the tasks.
- Were there times where you stopped talking about the books to help each other with language?
- What were you doing while another group member was discussing a prompt?

The fifth area was included to get an idea of the word association patterns. The questions in this area included:

- Why did you produce [association they produced] for the target word?
- On the pretest you provided more associations for this word that on the post-test. Was there a reason for this?

3.7.5 Reading logs

Each week, participants handed in a reading log which detailed the amount of additional reading they did that week. These logs were used to determine the extent that participants read additional material, to investigate if this added time on task may have contributed to learning. While students were not told to write information regarding the GRs, many often did. In this way, the reading logs also provided evidence that the participants were in fact reading (Mason, 2004). The logs included information regarding the book title, date read, and the amount read that day in pages (Bamford & Day, 2004; Day & Bamford, 2002; Smith, 2006). Section 8.3 in the Appendix depicts an example of the reading log.

3.7.6 Data collection procedure

The data from the self-report test, word association test, and the focus-on-form questions (with the exception of pronunciation-based questions) was collected electronically through the online software Qualtrics (https://www.qualtrics.com). The pronunciation-based question data was collected during the post-test interview. The reading logs were collected weekly.

3.8 Data analysis

All of the analyses were performed using a combination of two pieces of software. The first was Emacs, an extensible open-source text editor freely available from https://www.gnu.org/software/emacs/download.html. Emacs was used to code and classify the data. The second piece of software was the R software environment (R Core Team, 2016). R was used for all statistical analyses and for data visualization in this thesis.

Like Emacs, R is open-source and freely available at https://cran.r-project.org/.

3.8.1 Self-report data analysis

The self-report data was used in the following ways. The first purpose that the self-report data served was to address the issue of learning potential mentioned in the previous chapter. If a group has more potential to learn some of the target words, then it becomes difficult to determine if the cause of learning was due to the treatment, or due to increased potential for learning. The self-report data was also used to address the first and second research questions. The first research question asks about the extent to which the three ER conditions (ER-only, spoken, written) affect vocabulary development. The second research question looks specifically at the effects of the Say-it activity on lexical development through the form-focused episodes which occurred in the spoken mode versus the written mode. How the data was analyzed to address these questions is detailed below.

Determining initial knowledge Section 2.4.3 discussed how some research investigating the relationship between ER and vocabulary learning did not consider the learning potential of their participants regarding the initial state of knowledge of the target words (e.g., Smith, 2006). This can be problematic since it becomes impossible to know whether learning occurred as a result of the ER treatment, or as the result of greater learning potential since a person with lower initial knowledge has more that can be learned, and thus greater learning potential, than another person who has a greater amount of initial knowledge (and so less potential). Thus, the self-report data from the pretest was used to determine the participants' initial state of knowledge of the target words. As mentioned in section 3.7.1, the self-report test was designed to measured the target words at three levels: no knowledge, form knowledge, and meaning knowledge. Table 11 shows each self-report option in the wording the participants were exposed to in the tests (column one) with its respective coded value used for analysis (column two).

Self-report response on the test	Coded value
I have never seen this word	none
I have seen this word, but don't know what it means	form
I know the meaning of this word	meaning

Table 11: Self-report data coding scheme

In order to examine whether the three ER groups had similar levels of learning potential, the total words reported in each of the three categories were compared statistically using mixed effects modeling. To do so, the self-report data was re-coded as shown in Table 11 to be easier to deal with in R. Next, the target words were split into their respective frequency bands, and self-reported knowledge was totaled for each of the three levels. This resulted in four sets of values for each student, each set comprising one of the four frequency bands. Furthermore, each of the four sets consisted of three values which were the total words reported at each level of the self-report test. These values were analyzed statistically with mixed effects modeling, using the R package *lme4* (Bates, Maechler, Bolker, & Walker, 2015).

Analyzing vocabulary development through self-reported knowledge The method described in the previous section for analyzing the self-report data allows for an understanding of the number of target words known, or the vocabulary *size* of the participants for the target words. However, the research in this thesis focuses on depth of vocabulary knowledge, and as such, an additional approach was used to analyze the self-report data, to better capture depth of knowledge. In section 2.3.2 it was mentioned that depth of vocabulary knowledge can be seen on a developmental continuum, from complete lack of knowledge to mastery (e.g., Melka, 1997; Wesche & Paribakht, 1996). In this way, word knowledge travels along the continuum starting from a certain point and ending at another point, creating what will be referred to as a learning trajectory. Perhaps the most obvious trajectory starts at no knowledge, and ends at some knowledge, whether it be knowledge of form, knowledge of meaning, or

any other aspect of word knowledge. This trajectory represents a change in vocabulary size since a new word is added to the lexicon which did not previously exist. Another trajectory starts at a location on the continuum, other than complete lack of knowledge and moves to a position on the continuum which adds more knowledge to the word. For example, this can occur when a student learns a new meaning for a word which they already knew to some extent. This trajectory, instead of adding a new word to the lexicon, increases the degree or quality of the word knowledge. This trajectory represents a change in depth of vocabulary knowledge.

It would be incorrect to think that these learning trajectories can only travel from less knowledge to more knowledge; learning vocabulary is an incremental process, one that is not necessarily linear (e.g., Nation, 2013a). For example, it is possible that a learner has seen a word before, and after so much time, does not remember ever having seen the word before. In this case, the learning trajectory has gone from some knowledge to no knowledge. Since the number of words in the learner's lexicon has decreased, this type of trajectory is another example of change in breadth of knowledge. Similarly a learning trajectory can occur whereby a learner reports knowing the meaning of a word, and after so much time forgets the meaning but retains form knowledge. This trajectory represents a change from some (or more) knowledge to less knowledge. Since no new words are added or subtracted from the lexicon, but rather the quality of knowledge has changed, this is another example of a change in depth of vocabulary knowledge. Since the direction of change in these two types differs than the types in the previous paragraph, the types in the previous paragraph will be referred to as *forward* change, and the types in this paragraph will be referred to as *backward* change to differentiate.

The self-report data allows for one more type of trajectory to be measured. Unlike the previous two types, the third type is characterized by a *lack* of change. That is, this type of trajectory occurs when a learner reports the same level of knowledge on the pretest and post-test. In this way, a word does not change, and stays at either the unknown, form, or meaning level. This will be referred to as *no shift*.

These three types of learning trajectories, dubbed *forward*, *backward*, and *no shift*,

create nine possible learning trajectories. These nine trajectories are summarized in Table 12.

Self-report (pre)	Self-report (post)	Knowledge shift	Shift type	Development type	Trajectory code
none	form	none -> some		breadth	FB
none	meaning	none -> some	forward	breadth	FB
form	meaning	some -> more		depth	FD
form	none	some -> none		breadth	BB
meaning	none	some -> none	backward	breadth	BB
meaning	form	more -> less		depth	BD
none	none	no shift		none	NN
form	form	no shift	none	none	NN
meaning	meaning	no shift		none	NN

Table 12: Self-report learning trajectories

As mentioned in section 3.7.1, participants self-reported their knowledge of the 80 target words on the pretest and again on the post-test. Using these two reported values, each word for each participant was classified according to its learning trajectory and labeled using the trajectory codes in column six of Table 12. Each code consists of two letters derived from columns four and five in the same table. The first letter of the code represents the type of shift which occured, i.e., forward (F), backward (B), or none (N). The second letter represents the type of development, either breadth (B), depth (D), or none (N). These trajectory codes enabled investigation of both the amount and type of development which occurred.

This thesis investigates development of depth of vocabulary knowledge, and as seen in Table 12, depth of development can be measured using two trajectories: FD and BD. These two trajectories are the main focus in the analysis, however the other trajectories will be included to provide a more detailed picture of the extent of learning. To determine the degree that each group developed, these trajectories were examined

in the following way. Each type of learning trajectory was totaled for each participant after separating the words into their appropriate frequency bands. This resulted in four sets of totals for each student, one for each frequency band. Each set consisted of the total amount of the trajectories listed in Table 12.

It was hypothesized that both spoken and written groups would report a greater increase in depth of knowledge than the ER-only group. Even though all three groups read the same five GRs, the spoken and written groups were provided with opportunities to discuss the stories. As mentioned in section 2.4.1, this discussion provides opportunities to address gaps in learner knowledge through focusing on form (Swain & Lapkin, 1998). These opportunities can lead to increases in lexical knowledge (Swain & Lapkin, 1998). Accordingly, any increases in knowledge should be most pronounced with target words which occurring during form-focused episodes.

3.8.2 Word association analysis

Section 2.5 discussed issues in word association research, one of which is a lack of theoretical foundation for association response patterns. For example, there is no reason why an L2 learner's response patterns should become more native-like as they increase in proficiency (Wolter, 2002). Another issue mentioned was a lack of a means to determine adequately the strength of the association between two words. Meara and Wolter (2004) set out to address this in *V_links* by adding a self-reported strength measure. However, they note that there are still issues and further work needs to be carried out. To address these gaps, the current study uses Latent Semantic Analysis (LSA) to determine the strength of associations between words. LSA is based on the distributional hypothesis which claims a correlation exists between distributional similarity and meaning similarity (Sahlgren, 2008). That is, words which occur in similar contexts are more similar than words which do not occur in similar contexts (see section 2.5.5). By implementing LSA, the current study is unique because it determines the strength of word association prompt-response pairs using the input the learners were exposed to. This is in contrast to previous research which used self-reported measures of strength

(Meara & Wolter, 2004). This method for determining association strength provides a more theoretically-founded, objective measure than self-reported values. The semantic space created using LSA will always be the same, as long as the input used to create the space remains the same (LSA and semantic spaces are described in the following section). This overcomes the issue raised in Meara (2009), that some test-takers rate associations as strong which others tend to rate lower.

The word association data was used to address the first and second research questions. The hypothesis for research question one is that the spoken and written groups will increase their strength of associations to a greater extent than the ER-only group. This is due to the opportunity to consolidate their lexical knowledge and address gaps in their knowledge during form-focused episodes. How LSA was used to analyze the data is explained next.

Preparing the word association data for LSA The word association data was prepared for analysis in the following manner. Each response that the participants produced was first corrected for spelling if it was incorrectly spelled but was immediately recognizable as the intended word (e.g., *pletend* instead of *pretend*) (Fitzpatrick, et al., 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). It was decided to use the British English spelling because the majority of the GRs were written in British English, and since LSA creates a frequency of occurrence matrix, described in the next section, it was important that the word forms matched those in the GRs. The third step in preparing data 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 (see section 2.3.1 for a discussion of word lemmas). Changing the

responses to their respective lemmas reduces statistical noise in the semantic space created using LSA by merging semantically similar word forms (Lifchitz et al, 2009; Kantrowitz, Mohit, & Mittal, 2000). 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). The list is provided by the University of Lancaster under the Creative Commons Attribution-Share Alike 2.0 UK: England & Wales License. The lemmatization process was computed automatically using R. A script was written which takes a response, locates it in the lemma list, and returns the lemma of the word. This script was used for all of the responses.

Creating the semantic space Before the semantic space could be created, electronic copies of the GRs were retrieved with permission from the publishers, and were then lemmatized using the same R script and lemma list mentioned in the previous section. Next, all words in the GRs with an American English spelling and a British English spelling were changed to their British English equivalents in the same way as explained in the previous section for the word association responses. After these two preparatory steps were completed, the semantic space was created using the R package *LSA* (Wild, 2014). LSA determines the relationships between words in text by creating a word-by-document matrix. In the current study, each row in this matrix represents one lemma occurring in the GRs, and each column represented one paragraph of text from the GRs. In this lemma-by-paragraph matrix, each cell contains the frequency of occurrence of a lemma in a document. In other words, LSA initially creates a word-by-document frequency of occurrence matrix which shows the number of times that each lemma occurs in each document. Although any length of text can be used for the documents in each column, previous research has preferred paragraphs since they tend to consist of one self-contained idea (Lifchitz, Jhean-Larose, & Denhiere, 2009).

Next, each lemma and paragraph (i.e., each row and column) in the matrix are assigned two weights to indicate their importance in the semantic space. The first weight is a global weight applied to the paragraphs in the matrix. Specifically, it is the

inverse paragraph frequency where every cell is one plus the logarithm of the number of paragraphs divided by the number of documents where the term appears. This weight ranges from zero, when the lemma is present in all of the paragraphs with the same frequency, to one, when the lemma is present in only one paragraph. This weight was assigned using the *gw_idf* function in R. The second weight in the lemma-by-paragraph matrix was assigned to each lemma. The weighting assigned was the logarithm of the lemma's frequency in each paragraph. This weighting was assigned using the *lw_logtf* function in R. The weighting scheme described is the most common method used and emphasizes words which are unique in a paragraph, and de-emphasizes common words (Landauer & Dumais, 1997).

After the weighting scheme was applied to the lemma-by-paragraph matrix, the next step in the creation of the semantic space was to identify stopwords. Stopwords tend to be removed from the matrix since they occur with extremely high frequency and carry little meaning, for example determiners (Wild, 2015). Lifchitz, Jhean-Larose, & Denhiere (2009) posit that stopwords should be specific to a given corpus. They note that a good stopword candidate must have a low global weighting value and created an approach to dynamically determine stopwords. They considered the 150-200 lowest-ranking globally-weighted words as possible stopword candidates. They filtered through these words manually to determine which to consider stopwords, and then omitted them. This approach was adopted in the current study. To do so, the 150-200 words with the lowest global weighting in the lemma-by-paragraph matrix were filtered manually. This resulted in a stopword list of 14 words: *a*, *and*, *are*, *at*, *be*, *but*, *do*, *for*, *it*, *no*, *not*, *so*, *the*, and *to*. This stopword list comprised 0.49% of the total lemmas in the matrix, compared to Lifchitz, et al., (2009) whose stopword list was 3% of their total corpus. These stopwords were omitted from the GRs, and a new lemma-by-paragraph matrix was created using the same process just described.

The final step in creating the semantic space applies singular value decomposition, a two-phase factor analysis, to the weighted lemma-by-paragraph matrix. In the first phase of singular value decomposition, the weighted lemma-by-paragraph matrix is de-

constructed into three component matrices. The first component matrix are the values from the weighted matrix for the lemmas. The second component matrix represents the values for the paragraphs in the original lemma-by-paragraph matrix. The third component matrix consists of the singular values of the space, a matrix of non-negative numbers on the diagonal and zeroes in all other cells. With these three component matrices singular value enters its second phase, a reconstruction of the original lemma-by-paragraph matrix. The key point in this reconstruction process is the ability to adjust the representation of the lemmas and paragraphs in the space through dimension reduction (Landauer et al, 2007). This reduction removes variability or noise in the matrix, allowing Latent Semantic Analysis to capture the underlying semantic structure in the matrix. As a result, lemmas similar in meaning are "near" each other in the semantic space even if they never co-occur in a document, and documents similar in meaning are near each other even if they have no types in common (Berry, et al., 1995). This reconstructed semantic space is the foundation for the semantic structures LSA exploits. While there is no agreed-on number of dimensions which best captures the semantic structure, some research suggests that between 100 and 300 dimensions provide the best results (Jessup & Martin, 2001; Lizza & Sartoretto, 2001). For the current study, 300 dimension were used.

Computing cosine similarity The semantic space described in the previous section was used to determine the similarity of the participants' associations they produced for the target words. LSA can only measure similarity of words which are contained in the semantic space. This means that the 20 off-words, and any associations produced for them, could not be measured and so were omitted from the analysis. To determine the degree of similarity for the remaining 60 target words, the cosine similarity measure was used. Cosine similarity is a common measure for determining the similarity of items in semantic space (Landauer et al., 2007). The cosine measure can range from -1 to 1, with larger values indicating a greater degree of similarity, and lower values indicating a smaller degree of similarity.

The word association test used in the current study accepted multiple responses for each target word. One question this raises in terms of analysis is how to best handle these multiple responses. One way could be to compute the cosine similarity between a target word and each response given by a participant. For example, suppose the prompt *apple* produced as responses *fruit*, *red* and *dexterity*. In this instance three cosine similarity values would be calculated: one for *apple-fruit*, one for *apple-red*, and one for *apple-dexterity*. One issue here stems from the fact that not all participants provided the same number of associations for each word. In other words, the values become difficult to interpret since there are different amounts depending on the student. One way around this would be to take the average value of the set of associations produced for each word, by each participant. However, calculating the average value for the set of associations gives their *average* strength, it does not give their *combined* strength. Rather than taking the average of the set of associations for each word, the cosine similarity between the target word and the entire set of associations was calculated. This resulted in one cosine value for each target word, and represented association strength between the target word and the set of associations, according to their relationship in the GRs.

It should be noted that computing the cosine similarity using a set of associations does not necessarily inflate scores. For example, in the *apple* example in the previous paragraph, computing the cosine for the prompt-response set *apple-fruit-red-dexterity* may result in a lower similarity value than only *apple-fruit-red*; Most people would agree that *apple-fruit-red* is a stronger set than with the addition of *dexterity*.

To determine the degree of change in the semantic knowledge of a word, the pretest value was subtracted from the post-test value. This means that a positive difference (e.g., a value of .05 on the pretest and .1 on the post-test equals a increase of .05) refers to an increase in semantic knowledge from the pretest to the post-test. Similarly, a negative difference refers to a decrease in semantic knowledge. One issue which arises is with regards to missing values in the data, since not all of the words were reported as known at either the beginning or end of the treatment. A number of the responses

on either the pretest or the postest were not available as a student did not provide any associations, or the associations provided did not occur in the GRs. These were handled in the following way. If a there was no cosine score on the pretest, but there was on the post-test, the absolute value of the post-test cosine value was used. This is because there was a change in the amount of semantic knowledge equal to the post-test value. Conversely, if a cosine value was present on the pretest, but not on the post-test, the negative absolute cosine value was used. This is because there was some degree of measurable semantic knowledge on the pretest, and it was not present in the post-test. In addition, if no cosine value existed for the pretest and post-test, the value was coded as "NA". Note that the costring function, used to derive the cosine value between each prompt and set of responses, omits any words which are not in the LSA semantic space before calculating the cosine value.

As mentioned earlier, it was hypothesized that the ER plus groups would make greater increases in strength of association. As it relates to LSA, the hypothesis is that the spoken and written groups will achieve higher cosine values on the post-test than the ER-only group, and this will be most pronounced in those target words which learners addressed during form-focused episodes in the Say-it activities. That is, the learners will produce associations with stronger associations to the contexts in the GRs as a result of the reading. In addition, since the spoken and written groups were provided with the opportunity to discuss these contexts, they will have a more refined set of associations.

3.8.3 The Say-it activity analysis

Four triads completed the Say-it activities in both the spoken and written group after reading each GR. This resulted in a total of 40 Say-it activities. However, three participants in the spoken group were absent for the first Say-it activity. In addition, technical difficulties with one of the audio recorders during the second Say-it activities resulted in loss of data from one of the spoken groups. Accordingly, data from 38 Say-it activities was available for analysis. This equals 570 minutes, or nine hours-worth of

transcription data.

Transcribing the Say-it activity discussions The Say-it activities were transcribed using the following conventions. This was done in order locate the form-focused episodes, which will from now be referred to as language-related episodes or LREs (Swain & Lapkin, 1999). Each turn in the dialogue was annotated with metadata about the discussion, and the first bits of information included the GR number that was being discussed during that turn, along with an abbreviation of the GR title. Next, the mode of the Say-it activity was noted (i.e., either spoken or written), followed by the Say-it activity discussion prompt being discussed during the learner's utterance. Any discussion which was not about a Say-it activity prompt, for example a participant's weekend plans, was marked as *XO*. Third, the turn number of the triad was noted, in case it would need to be referenced later. Next was the participant initial whose turn it was, along with their individual turn number for that Say-it activity. The next piece of information noted was the time in the recording that the turn began, in case this turn were to be referenced. The last bit of information was the LRE number, where appropriate. Table 13 shows an example of this annotation. The table depicts the meta-data for an utterance from the activity for the first GR (gr1), Jojo's Story (Jojo) in the spoken group (oral task, *ot*), discussing prompt A1. This utterance was the triad's first turn (t1) and student J's first turn (J1). The utterance began at 4:20 in the recording and was part of an LRE.

gr1_Jojo_ot_A1_t1_J1_(4:20)_[LRE]:

Table 13: An example of the transcription conventions

In addition to the previously described metadata, the language produced by the students was also marked up, to provide a clearer idea of the nature of the conversation. This included annotations for interruptions, pausing, quiet speech, and other phenomenon. These conventions are shown in Table 14

Symbol	Meaning
-	an interruption
(.)	a short pause
(..)	a longer pause
~	quietly said
[	talk at same time
[mispronounce]	preceding word was mispronounced

Table 14: Transcription conventions for the Say-It activity dialogue

Categorizing the language-related episodes After the Say-it activities were transcribed and annotated, the LREs were located and categorized. An LRE was operationalized as the point in a group's dialogue when attention shifted from the discussion of a prompt to the language being produced, and ended at the point when focus returned to the Say-it activity discussion. This shift in attention was the result of a learner producing a linguistic error, asking about something that was said by another member, or asking about how to say something in English (Loewen, 2005). Each turn involved in an LRE was annotated using the "LRE" label seen in Table 13. The LREs were subsequently categorized according to the framework shown in Table 15, adapted from previous research (Ellis, et al, 1999; Loewen, 2005).

Feature	Feature type	Categories	Code
Instigation	type	*Reactive*: Correction of linguistic item	R
		Student-initiated: Query raised by Student	Si
Linguistic target	focus	*Grammar*	G
		Meaning	V
		Pronunciation	P
		Spelling	Sp

Continued on next page

Feature	Feature type	Categories	Code
Apparent reason for instigation	source	*Code*: Inaccurate use of linguistic item with no apparent miscommunication	C
		Message: Problem understanding meaning	M
Length	length	*Simple*: Only one response move	S
		Complex: More than one response move	Co
Explicitness of feedback	direct	*Indirect*: Implicit (e.g., recast)	I
		Direct: Explicit (e.g., metalingual explanation)	D
		No Feedback: No feedback given	xID
Complexity + Directness	emphasis	*Light*: Indirect and simple	L
		Heavy: Direct, complex, or both	H
Response timing	time	*Immediate*	Im
		Deferred	De
		Both: partly corrected, then fully later	ID
Type of feedback provided by student	response	*Provide*: S gives information about a language form either by use of a recast or an inform	Pr
		Elicit: Another S attempts to draw out from S a language form or information about a language form	El
		No Response: S has difficulty, no one provides feedback	xRe
Student response to feedback	uptake	*Uptake*: S produces a response	U
		No uptake: S does not produce a response	xU
		No opportunity: S does not have a chance to respond	X
Quality of student response	Successful uptake	*Successful uptake*: S incorporates linguistic information into production	Su
		Unsuccessful uptake: S does not incorporate linguistic information into production	xSu

Continued on next page

Feature	Feature type	Categories	Code
		Not Applicable: Irrelevant	xx

Table 15: Language-related episode coding system, adapted from Loewen (2005)

To explain this framework an example will be used, shown in Table 16. This is an excerpt from a triad in spoken group which occurred during the first activity, after reading the graded reader *Jojo's Story*, a story about a young boy's experience with war. As seen in the table, the LRE is initiated when student S1 reads a Say-it activity discussion prompt to S2 containing the word *lorry*, which S2 is unfamiliar with. In turn 2, S2 asks for clarification of *lorry*. S3 provides a synonym in turn 3. In the final turn, S2 acknowledges this assistance and then returns to discussion of the prompt.

Turn	Student	Dialogue
1	S1	You are Jojo. You hear a lorry. What are you thinking?
2	S2	Mmm, lorry?
3	S3	Lorry uh a big car.
4	S2	Oh oh, ok. Firstly, I also really scared and, because I afraid that the man come...

Table 16: An excerpt from a form-focused episode which occurred in the spoken group (LRE 1)

Referring again to Table 15, this LRE is *reactive* since S2 is responding to something that was said. Next, the focus of the LRE is on the *meaning* of *lorry* and began due to a lack of understanding in the *message* of what was said. There was only one response turn making the LRE *simple*, and the feedback was *direct*. The response was *immediate* and the feedback was *provided* to the S2. There was *uptake* in turn four when S2 acknowledged the assistance from S3, however there was no *successful uptake*

because S2 did not produce *lorry* in their subsequent speech. Each LRE was categorized in this way, using the framework in Table 15. To ensure that LRE classification was accurate, three inter-raters, all studying at the post-graduate level at Victoria University of Wellington, independently rated the same 10% of the LREs. The inter-rater reliability statistic used was Maxwell's RE, and the resulting value was 0.718. After rating, the raters met with the researcher to resolve all inconsistencies, and the researcher classified the remaining LREs.

Analyzing the language-related episodes As discussed in section 3.7.3, questions were created based on the nature of LREs during the activities. These questions were given to the participants in the triads who initiated a form-focused episode due to a lack of linguistic knowledge. Three question types were created as a result of the nature of the LREs: suppliance, correction, and pronunciation (Table 10 explains these three types). Each of these questions was scored as either *correct*, *partially correct*, *other correct*, or *incorrect* following Loewen (2005). Specifying four types of answers increases the sensitivity of the test, allowing for a more precised measurement of the extent of learning (Dobao, 2014). An answer was marked *correct* if it accurately matched the linguistic item being tested. This signifies that the learner produced the item correctly which they had difficulty with during the LRE. An answer was marked *partially correct* if it was an improvement, but still not entirely accurate. This tended to occur with the pronunciation-based items, since there were two chances for a participant to pronounce the word correctly. If the learner mispronounced a word one of the two times, it was scored as partially correct. An answer was *other correct* if the targeted linguistic item in the LRE was corrected, but in a way that differed from the information present in the LRE. This could occur when a student provided another word which also fit a context, instead of the word which occurred during the LRE. Finally, an answer was *incorrect* if it was the wrong item, or if no response was given. Table 17 summarizes this scoring rubric.

In order to determine the extent that the LREs were conducive to language learning,

Score	Explanation
correct	Provides correct item from the LRE trigger
partially correct	Improves on the error, but not completely
other correct	Provides accurate information, but not the targeted item
incorrect	Provides incorrect or no item

Table 17: Language-related episode scoring rubric, adapted from Loewen (2005)

the four-point system in Table 17 was simplified following Loewen (2005). Specifically, *correct*, *partially correct*, and *other correct* scores all include aspects of correctness and so were all considered *correct*. The *incorrect* answers remained *incorrect*. Using this dichotomy, a mixed effects model was built using the *glmer* function in R (Bates, Maechler, Bolker, & Walker, 2015) with the *family = binomial* option specified. The dependent variable in the model was score, and had two levels: *correct* and *incorrect*. The independent variables were group (ER-only, spoken, written). In addition, person was specified as a random effect in the model.

3.8.4 Post-test interview analysis

A total of 41 out of 48 possible interviews were conducted due to participant absenteeism. The data in the interviews was used to gain a deeper understanding of the quantitative data. This included participant perspectives on the testing and also about the extensive reading. The interviews were used to collect information about the reading habits of the participants, as well as to ask about the Say-it activities (for those who were in the spoken or written groups). In addition, the interview data was used to investigate the reasons for the word association patterns found in the test. It was not possible to ask all participants the same questions, and it was often the case that

the questions asked during an interview was in part determined by the answers given by the participant. As a result data from the interviews will be discussed only as it becomes relevent.

3.8.5 Reading log analysis

The reading logs were used to determine the extent that the three ER groups read in addition to the reading assigned in the study. The participants handed in their reading logs on a weekly basis. After the last log was collected, the total amount of pages read by each student was calculated. Then, the average amount of pages was computed by taking the total amount of pages each student read and dividing it by the number of logs that they submitted. This resulted in the average amount of pages read per week. The reason that the average amount of pages per week was used, as opposed to the total amount, is because not all of the students handed in reading logs every week. Computing the average amount of pages read was thought to be a way around this. As discussed in the literature review, scholars agree that ER leads to gains in proficiency (e.g., Elley & Mangubhai, 1981). As a result, it was possible that the lexical development which occurred could relate to the fact a group was reading more than another.

An ANOVA was computed to determine whether the three ER groups engaged in additional reading to the same degree. The independent variable in the model was ER group and had three levels (ER-only, spoken, written). The dependent variable was average pages per week.

3.8.6 Determining proficiency of the three ER groups

In order to determine whether the three ER groups had similar proficiency, an ANOVA was computed using the participants' combined English Proficiency Program placement test scores. The purpose of this analysis was to rule out the possibility that proficiency was responsible for changes in vocabulary development, and a stronger case could made

for the treatment causing the development. The independent variable in the ANOVA was ER group, and had three levels (ER-only, spoken, and written). The dependent variable in the ANOVA was the combined placement test scores. These combined scores is what the English Language Institute uses to determine which class a learner should be placed into. It was thus deemed appropriate to use the combined score as a measure of proficiency.

3.9 Procedure

Ethics approval was obtained from the Victoria University of Wellington Human Ethics Committee before the study began. Participants were asked to volunteer for the study following a short introduction about the researcher. Due to the nature of incidental learning, the participants were not told that incidental vocabulary acquisition was being investigated. Instead, they were told that reading and language learning was being looked at. A consent form along with an information sheet explaining the study were distributed to each student. All students' questions regarding the study were addressed before they signed the consent forms.

A few days after signing the consent forms but before the reading commenced, participants took a pretest during normal class time. The format of the test was explained to the participants before the test began and examples of the different test sections were shown to the participants and were worked through as a class. After this introduction, and ensuring no students had any questions, they began the test. The test took approximately 40 minutes to complete. Figure 1 depicts the design of the pretest.

The participants began reading a few days after the pretest. The GRs were delivered to the students' classrooms the morning of the first day of reading each book. The reading took place for 15 minutes in class everyday. Students were given a reading schedule which had a number of chapters to read each day, amounting to approximately one chapter each weekday, and three chapters over the weekend. All of the books

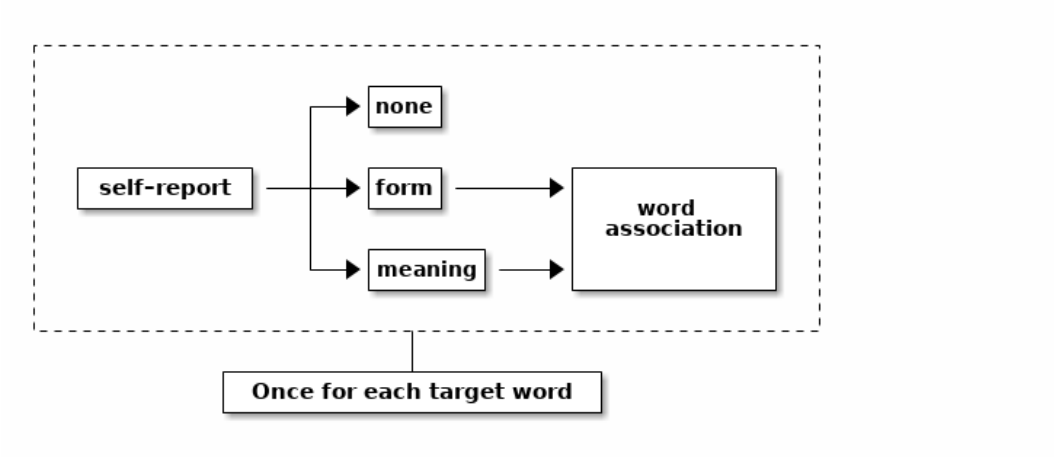

Figure 1: Pretest design

were completed in about 7-9 days, a pace that scholars note allows for repetition and reinforcement of new input (Nation & Wang, 1999).

The day after finishing each GR, those in the ER-only group read a chapter from the short-story book that they chose, while those in the ER-plus group completed the Say-it activity. Both groups completed their respective activity during normal class, for 15 minutes. Before the first Say-it activity, participants watched the video of three native English speakers completing the task in both spoken and written modes. Participants were then assigned to a triad. The triads were created using participant proficiency and gender, and the triads did not change during the study. Using the EPP placement tests as a guide, triads were composed of one higher-proficiency student, one average proficiency student, and one lower-proficiency student. In addition, each group had at least one female and one male student. Triads were randomly assigned to either the spoken or the written mode, and remained in that mode for the duration of the study.

After the fifth and final Say-it activity, all participants took the post-test, and subsequently had a one-on-one interview with the researcher. All of the interviews took place within one week of finishing the post-test. Figure 2 illustrates the design of the post-test.

The data collected in the study from the self-report test, word association test, and

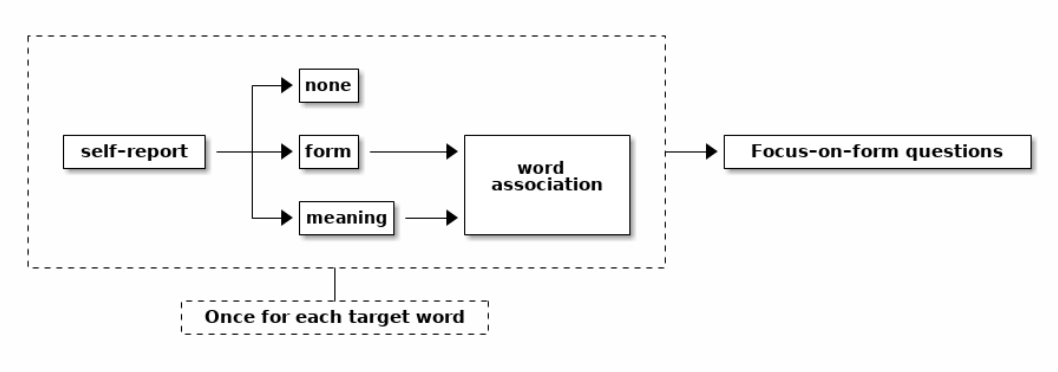

Figure 2: Post-test design

the focus-on-form questions (with the exception of pronunciation-based questions) was done so electronically through the online software Qualtrics (https://www.qualtrics.com). The pronunciation-based question data was collected during the interview. The reading logs were collected weekly. Figure 3 illustrates the research design of the main study for phase one. Phase two will be explained in section 3.11.

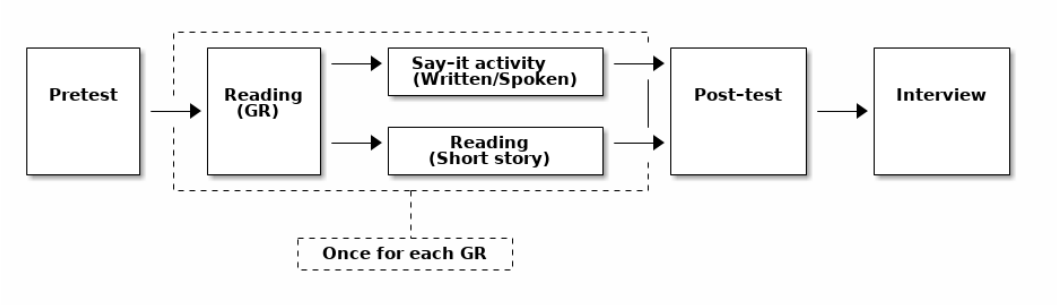

Figure 3: Research design of the main study for phase one

3.10 Pilot study

A pilot study was conducted to trial the materials and procedures to be used in the main study. Twenty-three English Proficiency Program students participated in the study. These students were not part of the main study. The 23 students were from a similar proficiency level as the students in the main study. Two graded readers were

used to pilot the Say-it activity: *Billy Elliot* and *A Kiss Before Dying*. Similar to Joe (2006), the pilot study was limited to two GRs due to the degree of intrusion on both the teachers' and learners' time. Figure 4 illustrates the design of the pilot study.

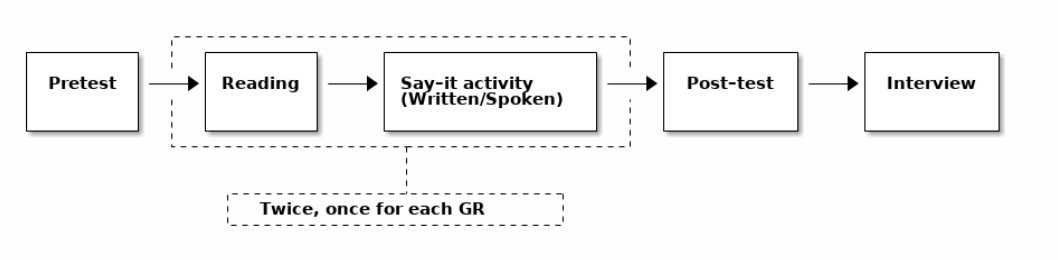

Figure 4: Design of the pilot study

3.10.1 Procedure for the pilot study

The participants sat the pretest, and the following week began reading. The reading took place Monday to Friday in a room reserved specifically for the reading. Each day, the students met in this room after their English class and read for 15 minutes with the researcher. After finishing each GR, the participants were split into triads, and randomly assigned to either the spoken group or the written group. An ER-only group was not included in the piloting since the purpose of the pilot was to trial the materials and procedures, not to compare groups.

The researcher explained the Say-it activity to the students, and after ensuring all participants understood what was expected, each spoken group triad was asked to sit together at a table in the room, and subsequently received a piece of paper with the Say-it activity discussion prompts (the three-by-three grid). An audio recorder was also placed on each table to record learner dialogue during the activity. The spoken groups were then told to begin, and they discussed the prompts on the paper for 15 minutes. An assistant was asked to help for the pilot study, to monitor the students in the spoken group while they completed the activity. The assistant was an Honors student studying Linguistics at the same university where the research took place. She

timed the students, collected the audio recorders and prompts at the end of the activity, and kept them in a safe place until they were retrieved by the researcher.

After the spoken group began the Say-it activity, the written group was taken to a computer room reserved for the pilot study. Upon arriving in the computer room, the students were asked to sit at a computer and to log in with their university login and password. The Say-it activity procedure was then explained to them, explaining that they would be doing the same activity as the spoken group, but that instead of talking they would be typing. In order to use Google Docs, a Google email address is necessary, and those students who did not have an Google email address were asked to create one. Each student was assigned to a triad, and they then logged into their communal Google document using a link provided to them via email. At the top of each triad's document was a digital version of the Say-it activity discussion prompt grid, the same prompts that the spoken group received. Once all of the learners were logged in to their respective Google Documents, they were told to begin the 15-minute task.

Approximately one week after the second Say-it activity was completed, all participants took the post-test. Over the following three weeks, participants were contacted and asked if they would volunteer for an interview.

3.10.2 Piloting lessons learned

A few changes were made to the research design for the main study as a result of piloting. These changes came about from issues which arose during the pilot study. One issue which arose was that the instructions for the word association section of the test confused some of the participants. The original instructions read *Type up to 5 words that are related to [target word]*, and some participants were unsure what was meant by "related". The instructions were modified for the main study to become: "Type (up to) the first 5 words that come into your mind when you think of [target word]".

Another issue which arose was that the participants all read to slightly varying

speeds. This meant that some of the students finished the GRs before other students, some even days before others. In order to address this issue, the design of the main study was modified so that learners were assigned a certain amount of chapters each day, of which 15 minutes of in-class time was allowed for this reading. In this way, the students would know how much to read each day, and all participants would be reading at the same pace.

The third issue which arose in the pilot study regarded the Say-it activity. Some of the participants mentioned that some of the discussion prompts, especially those which included a minor character or minor event, were difficult to remember. For the main study, prompts were redesigned to omit these minor characters and events. A related issue was with character names themselves. Some participants mentioned that it was difficult for them to remember which character was which since the names were hard to remember. To avoid confusion in the main study, the prompts were rewritten to be more descriptive. For example, one of the prompts relating to the GR *Billy Elliot* was initially *You are Tony. Talk about why you went to jail.* For the main study, it was changed to become: "You are Billy's brother, Tony. Talk about why you went to jail".

Another issue with the Say-it activity occurred in the written mode. After the first pilot study Say-it activity, one of the participants said that they were uncomfortable editing their group members' text, and even if they noticed a mistake, they were hesitant to change it. To combat this, during the introduction to the Say-it activity, when the learners were sitting at their computers immediately before the first Say-it activity, participants were told that it was fine to interrupt their group member's typing if there was an issue, or if they wanted to ask a question or say something.

The final issue arising in the pilot study involved the target words. Section 3.5 describes the criteria applied to target word selection for the main study. In the pilot study, only the first three criteria were used, namely:

- frequency of occurrence in the GRs
- frequency of occurrence in the British National Corpus
- word length in characters

Some of the participants in the pilot study mentioned having difficulty producing word associations for some of the words, despite knowing the meaning of the word. This was because the word was sometimes hard to define, difficult to imagine, or fairly abstract. To control for these psycholinguistic aspects, a more detailed list of selection criteria was establishing using the psycholinguistic features from the Medical Research Council database (section 3.5).

3.11 Interim results and modifications

As mentioned in the introduction to this chapter (section 3.1), a decision was made to include a second phase of data collection, and this decision arose as a result of an interim analysis of the data. Of the 48 participants in the study, 13 made up the spoken group, and 12 comprised the written group. It was feared that these numbers were low and as a result may not give reliable results, especially when applying statistical tests. To mitigate this, it was decided to conduct a second collection phase. This provided an opportunity to determine if any modifications to the design, which hitherto limited the design, should be implemented. To that extent, an interim analysis was carried out before the second collection phase which examined the design of the study and also looked at the data to ensure that there were no issues. This analysis revealed a few areas which were to be addressed for phase two of collection. The first issue that was addressed regarded the target words. The data revealed that the off-words, the 20 low-frequency words not occurring in the GRs, were not behaving as they were intended. Originally, they were designed to determine the amount of learning outside of the study (refer to section 3.5). However, upon investigation that many of the students had misinterpreted them as other words. The off-word *sward*, which refers to an area of grass, is one example of this. Many students reported knowing the meaning of *sward* on either the pretest or the post-test, yet their word association responses suggested otherwise. The participants often produced associations such as *weapon* or *blade*, suggesting that they misinterpreted *sward* for the word *sword*. These off-words

were not adequately serving the purpose for which they were included, and so were deemed unreliable and removed from the study. It is acknowledged that they may not have been well-chosen.

Another issue discovered during the interim analysis related to the Say-it activity. Comparing the spoken and written groups' dialogues revealed a large difference in the amount of turns each group had the opportunity to take. The total number of turns, derived from adding both the spoken and written groups' transcribed dialogues was 2494. The spoken group had 1802 of these turns (72%), while the written group had only 692 (28%) turns. That the written group produced fewer turns than the spoken group is most likely due to the nature of the medium of communication. It takes longer to type a sentence than it does to say it. To determine if this was so, the average words produced per turn for both groups was computed by summing the number of words in each turn divided by the amount of turns each group had. In the spoken group, the average turn length was 11.4 words (SD = 19.7), and in the written group the average turn length was 11.1 words (SD = 12.1). These averages are very close and support the idea that it takes longer to type than to say; The students, by typing the same length of sentence as those in the spoken group said, are taking more time to type, and less time to communicate with their group members. With these two issues confronting the validity of the research, the written mode was deemed unfit to be implemented in the design created for the current study, and as a result it was not included in the second phase of collection.

The last issue discovered was with the LSA values. As mentioned in section 2.3.2, incidental vocabulary learning is slow and incremental. By nature then, it was not expected to find large changes in development. The interim analysis confirmed this, with little movement in the groups. The most promising words were those in the mid-frequency word band, since they showed the most movement. Since the off-words had been removed from the test, an opportunity to include an additional measure of productive knowledge without fear of learner fatigue setting in presented itself. As a result, a C-test was included in the second phase of data collection.

The methods and procedures in phase two were identical to those in phase one, with the only differences being the participants and the addition of the C-test.

3.11.1 Phase two: Participants

The participants in phase two consisted of two intact English Proficiency Program classes, similar to those classes in phase one. In order to ask for volunteers, the teachers from two classes were approached and asked if they would like to participate in study. One teacher gave permission to enter the class and ask the students. The other teacher, however, said they had too much going on in their class to include anything else. To that extent, the teacher from a third class, a slightly higher proficiency group, was asked to participate and they said they would be willing. Since this had slightly higher proficiency test scores than the students in phase one, they were assigned to the ER-only condition, to not jeopardize the ER-plus groups' data.

combine dat_p2 with dat12 Twenty-seven students participated in phase two, 10 whom were female and the remaining 17 whom were male. From these 27 participants, 14 formed the ER-only group, and 13 students formed the spoken group. The ER-only had an average score on the placement test of 150.2 (SD = 6.9) points, approximately 20 points higher than the phase one participants. The spoken group, on the other hand, had a similar score to the phase one participants at 137.2 (SD = 8.1). The proficiency scores will be described in more detail in the following chapter.

3.11.2 Phase two: Data collection

The C-test questions were created to assess productive knowledge of the mid-frequency target words. The format of the test asked participants to supply a missing word in a sentence, a format similar to Laufer and Nation's (1995) Productive Vocabulary Levels Test. For example, for the sentence "I work too much and want a vaca_____", students were to supply *vacation*, although "*tion*" was also acceptable. The first few letters of each answer were provided to eliminate other possible answers (Laufer &

Nation, 1995). The blank line following the first few letters of each item was always the same length and did not give any clues as to the length of the answer. There were 20 questions in this test, one for each mid-frequency word.

3.11.3 Phase two: Data analysis

The C-test questions were formatted in a similar manner to the word association data (See Section 3.8.2), with the exception that they were not lemmatized. Each question was scored as either correct or incorrect. Mistakes made by participants were scored as correct only if the learner's intention was immediately clear and the answer was correct. Table 18 gives examples of accepted answers. Participant scores were summed and then compared across the two groups using mixed effects modeling to determine the extent to which development occurred.

Correct answer	Acceptable answers
blouse	blose; blosse; blous
closet	closent
goat	ghote
papa	papee; pappy
toxic	toxin; toxical
campus	campos; campous
gloves	glovies
lorry	lory

Table 18: Accepted answers for the C-test

Phase two C-test: Piloting The C-test questions were piloted using fifteen post-graduate students. These students were all studying at the Victoria University of Wellington, either at the Master or PhD level. Some of students were L1 speakers of English, while others were L2 speakers. The test was piloted to ensure that the

wording of the questions was comprehensible. No student reported that the questions were difficult to understand.

3.11.4 Phase two: Procedure

Figure 5 shows the design of phase two, which was identical to phase one with three notable differences which were mentioned in Section 3.11 and are summarized here. The first difference was that the Say-it activity was completed in the spoken mode only; the written mode was omitted. The second difference between phases one and two is that the 20 off-words used in phase one were omitted from phase two. The third difference found in the two phases was the addition of the C-test in phase two, which was absent in phase one.

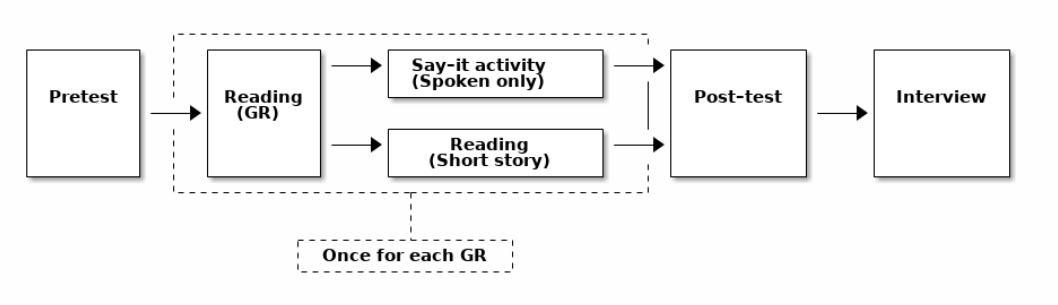

Figure 5: Research design of the main study for phase two

3.12 Chapter summary

This chapter described how the issues found in the literature review in the previous chapter were addressed. The extent that this mixed-methods design was able to capture the type of development it was designed for will be discussed in the conclusion of this thesis. The next chapter presents the findings of the study, and will be presented according to each research question posed in section 2.7.

4 Results

4.1 Introduction

This chapter presents the findings of the current study, to answer the research questions posed in section 2.7. Note that any development which occurs may depend on the initial state of a person's knowledge; if the three ER groups in this thesis started their respective interventions at different levels of proficiency for example, any changes which took place may be due to this initial difference. To that extent, the chapter begins by presenting the results of the three groups' initial knowledge state, before the intervention. Results are presented for the three groups in terms of

- English proficiency levels, determined by the English Proficiency Program's proficiency test (section 4.2.1),
- amount of additional reading, determined by the learners' reading logs (section 4.2.2),
- reported knowledge of the target words on the pretest (section 4.2.3), and
- semantic knowledge of the target words on the pretest (section 4.2.4).

The spoken and written groups were provided opportunities to discuss the material they read in small groups during the Say-it activities, which provided opportunities for language-related episodes to occur. The episodes have been shown to facilitate vocabulary development, and to examine the extent to which this was true in the current study, the results illuminating the nature of these episodes are presented in section 4.3.

After establishing the initial status of the three groups, as well as the nature of the language-related episodes which took place during the Say-it activities, results are presented showing the extent that development occurred in the 60 target words for each of the three ER groups. Section 4.4 focuses on the reported development of the 60 target words, and section 4.4.3 zooms in on the self-reported knowledge of those target words which arose in a language-related episode. Building on these results, section

4.5 presents the results detailing the extent of semantic knowledge development which occurred for the 60 target words, while section 4.5.1 focuses on the semantic knowledge development for those target words which triggered a language-related episode. Next, section 4.6 presents results showing the extent that the groups answered the questions created from the language items which triggered a language-related episode, followed by section 4.7 which presents the C-test results for phase two of the study. Finally, section 4.8 concludes the chapter by summarizing the key findings presented, and subsequently shifting attention to the following chapter which discusses these key findings in order to provide an explanation for the development which occurred.

4.2 Determining the comparability of the ER groups at the beginning of the study

4.2.1 To what extent were the groups' English proficiency similar?

To determine the extent that the three groups had similar proficiency, an ANOVA was computed using the combined proficiency test score as the dependent variable, and ER group (ER-only, spoken, written) as the independent variable. Student data from phases one and two were included in this model. Table 19 presents the results of the ANOVA, revealing a significant effect of group. Multiple comparisons of means using Tukey contrasts revealed that the ER-only group had a significantly higher proficiency than the written group ($t = 2.827$, $p = 0.02$). No significant differences were found between the ER-only and spoken groups ($t = 1.779$, $p = 0.18$), or the between the written and spoken groups ($t = 1.369$, $p = 0.36$). A Shapiro-Wilks test was computed using the residuals from the ANOVA model, showing a normal distribution ($W = 0.97$, $p = 0.15$). These results suggest that all of the data from phases one and two should not be combined since proficiency would become a confounding variable.

The previous paragraph revealed that the data from phases one and two should not be combined due to group proficiency differences. To determine the homogeneity of the groups in phase one, an ANOVA was computed with combined proficiency score as

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2.00	1225.04	612.52	4.39	.02 *
Residuals	68.00	9482.54	139.45		

Table 19: ANOVA results comparing ER group (ER-only, spoken, written) by combined proficiency score for phases one and two combined

the dependent variable and ER group (ER-only, spoken, written) as the independent variable. Table 20 presents the proficiency test results for this data, revealing no significant differences between the three groups ($F = 1.389$, $p = 0.26$). A Shapiro-Wilks test using the residuals from the ANOVA confirmed the data was normally-distributed ($W = 0.98$, $p = 0.43$). In short, the three groups in phase one had similar proficiency levels, confirming the counterbalancing was successful.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2.00	223.95	111.97	1.39	0.26
Residuals	43.00	3465.53	80.59		

Table 20: ANOVA results comparing ER group (ER-only, spoken, written) by combined proficiency score for phase one

The proficiency test results have revealed that all of the data from phases one and two should not be combined due to proficiency differences between the ER-only and written groups. It was mentioned in section 3.11 that the Say-it activity in the written mode was deemed unfit and so was omitted from the second phase of data collection. To determine if the ER-only group and spoken group had similar proficiencies, the data from phases one and two were combined, omitting the written group's data, and an ANOVA was computed with combined proficiency test score as the dependent variable, and ER group as the independent variable (ER-only, spoken). Table 21 presents these results, confirming that the two groups had similar proficiency ($F = 2.771$, $p = 0.1$). A Shapiro-Wilks test was computed on the ANOVA residuals and revealed a normal distribution ($W = 0.96231$, $p = 0.07$).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	1	441.23	441.23	2.77	0.1015
Residuals	57	9075.62	159.22		

Table 21: ANOVA results comparing ER group (ER-only, spoken) combined proficiency score for phase two

In short, these results revealed that the written group had a significantly lower proficiency than the ER-only group when combining the data from phase one and phase two. When omitting the written group from the data, the proficiency of the groups become similar. In light of these findings, the results for phase one will be analyzed separately, and to reinforce the findings, the data from phase two will be combined with phase one and analyzed separately, omitting written group's data.

4.2.2 To what extent did the three groups read additional material during the intervention?

For phase one, reading log results from 32 participants were available: 16 students from the ER-only group, nine students from the spoken group, and seven students from the written group. To determine the degree to which the three groups read in addition to the five graded readers used in the intervention, an ANOVA was computed with pages read per week as the dependent variable, and group as the independent variable. The results revealed no significant effect of group ($F = 1.444$, $p = 0.25$), suggesting that the three groups read to a similar degree. However, a Shapiro-Wilks normality test was computed using the residuals of the ANOVA model revealing a non-normal distribution ($W = 0.6587$, $p < 0.001$). To confirm the results of the ANOVA, a test which assumes a normal distribution, a Kruskal-Wallis test was additionally computed. The results of the Kruskal-Wallis confirmed no significant difference in the amount of reading that the three groups did outside of class ($\chi^2 = 3.19$, $p = 0.2$).

A similar analysis was conducted using the phase one and phase two data, omitting the written group for reasons mentioned in section 4.2.1. Reading logs results from

50 participants were available: 30 in the ER-only group and 20 in the spoken group. An ANOVA was computed with pages read as the dependent variable, and group as the independent variable. The results revealed no significant effect of group, meaning the three groups read outside of class to similar degrees ($F = 0.089$, $p = 0.77$). A Shapiro-Wilks test of normality on the residuals of the ANOVA revealed a non-normal distribution ($W = 0.59978$, $p < 0.001$), and so to confirm the results of the ANOVA, a Kruskal-Wallis test was computed. The results confirmed no significant difference in the amount of reading that the three groups did outside of class ($\chi^2 = 1.5941$, $p = 0.21$).

In sum, the results of the reading log analysis revealed that the groups in phase one and phase two engaged in additional, extracurricular reading to similar degrees. This result provides evidence that can be used to rule out any significant effects that this extracurricular reading may have had on any lexical development.

4.2.3 Did the ER groups self-report target word knowledge to similar degrees before the intervention?

The self-reported knowledge each group gave for the 60 target words were compared across groups and target word frequency bands to determine if there were any differences in initial knowledge. To do so, a mixed effects model was fitted to the data with number of target words as the dependent variable, and ER group (ER-only, spoken, written) and frequency band (high, mid, low) as independent variables. In addition, participant was specified as a random variable in the model. Since the dependent variable represented counts (of words), the model was fitted assuming a Poisson distribution using the *glmer* function in R and the "family=poisson" option.

	Estimate	Std. Error	CI95lower	CI95upper	z value	Pr(>|z|)	
(Intercept)	-1.06	0.35	-1.75	-0.36	-2.99	0.003	*
groupSpoken	0.69	0.49	-0.26	1.64	1.42	0.157	
groupWritten	1.06	0.46	0.16	1.95	2.31	0.021	*
bandmid	3.12	0.36	2.42	3.83	8.65	0.000	*
bandlow	3.82	0.36	3.12	4.52	10.68	0.000	*

knowledgeform	-0.29	0.54	-1.35	0.77	-0.53	0.594	
knowledgemeaning	4.02	0.36	3.32	4.72	11.27	0.000	*
groupSpoken:bandmid	-1.10	0.51	-2.09	-0.11	-2.18	0.029	*
groupWritten:bandmid	-1.00	0.47	-1.93	-0.08	-2.12	0.034	*
groupSpoken:bandlow	-0.76	0.49	-1.73	0.21	-1.53	0.125	
groupWritten:bandlow	-0.93	0.46	-1.84	-0.02	-2.01	0.045	*
groupSpoken:knowledgeform	1.27	0.67	-0.04	2.58	1.90	0.057	
groupWritten:knowledgeform	-0.81	0.79	-2.36	0.74	-1.03	0.305	
groupSpoken:knowledgemeaning	-0.79	0.49	-1.76	0.17	-1.61	0.107	
groupWritten:knowledgemeaning	-1.09	0.46	-2.00	-0.19	-2.36	0.018	*
bandmid:knowledgeform	-0.74	0.56	-1.84	0.35	-1.33	0.184	
bandlow:knowledgeform	-1.47	0.56	-2.56	-0.37	-2.63	0.008	*
bandmid:knowledgemeaning	-3.86	0.37	-4.59	-3.14	-10.42	0.000	*
bandlow:knowledgemeaning	-6.42	0.40	-7.21	-5.64	-16.04	0.000	*
groupSpoken:bandmid:knowledgeform	-0.13	0.70	-1.50	1.25	-0.18	0.856	
groupWritten:bandmid:knowledgeform	0.76	0.83	-0.86	2.38	0.92	0.357	
groupSpoken:bandlow:knowledgeform	-0.86	0.70	-2.23	0.51	-1.23	0.218	
groupWritten:bandlow:knowledgeform	-0.73	0.88	-2.45	1.00	-0.83	0.409	
groupSpoken:bandmid:knowledgemeaning	1.17	0.53	0.14	2.20	2.23	0.026	*
groupWritten:bandmid:knowledgemeaning	1.00	0.49	0.03	1.96	2.01	0.044	*
groupSpoken:bandlow:knowledgemeaning	0.83	0.58	-0.31	1.96	1.42	0.154	
groupWritten:bandlow:knowledgemeaning	0.96	0.56	-0.14	2.05	1.72	0.086	

Table 22: Mixed effects model showing a three-way interaction on the self-report pretest, phase one

The results of the mixed effects model, shown in Table 22 reveal a significant three-way interaction effect between group, frequency band and self-reported knowledge. To determine where the differences were, multiple comparisons of means using Tukey contrasts were conducted. The contrasts were specified to compare each of the groups on the same frequency band (i.e., high, mid, and low) and self-reported knowledge (none, form, meaning). This way, the ER-only group's results for the high-frequency words reported at the meaning level for example could be compared to the spoken and written groups' results for the high-frequency words reported at the meaning level. Effect sizes can be problematic for mixed effects modeling (Nakagawa & Schielzeth, 2013), and so were not computed.

The post-hoc multiple comparisons revealed four significant three-way interactions.

The comparisons revealed that the ER-only group reported significantly fewer high-frequency words at the form level compared to the spoken group (z = -4.287, p = 0.01). To put it another way, the spoken group reported more high-frequency words at the form level than the ER-only group. Similarly, the ER-only group reported significantly fewer mid-frequency words at the form level compared to the spoken group (z = -4.303, p = 0.01). This means that the spoken group reported more mid-frequency words at the form level than the ER-only group. The third interaction showed that the ER-only group reported significantly more low-frequency words at the form level than did the written group. Coupled with the fourth interaction showing that the spoken group also reported significantly more low-frequency words at the form level than the written group (z = 4.602, $p < 0.01$), it can be said that the written group reported significantly fewer low-frequency words at the form level compared to the other two groups.

To reinforce the findings from phase one, the analysis was computed again, this time incorporating the phase two data, and omitting the written group's data. The dependent variable was word count, and the independent variables were group (ER-only, spoken), frequency band (high, mid, low), and self-reported knowledge (none, form, meaning). The results of the analysis are presented in Table 23. The table shows no significant three-way interactions, in contrast to the phase one analysis. This suggests that neither group reported more words at a certain level in a certain band. Put another way, the ER-only group and the spoken group both reported similar amounts of words at each combination of frequency band and self-report level.

	Estimate	Std. Error	CI95lower	CI95upper	z value	Pr(>|z|)	
(Intercept)	-1.41	0.33	-2.07	-0.76	-4.24	0.000	*
groupSpoken	0.64	0.44	-0.22	1.50	1.45	0.146	
bandmid	3.30	0.34	2.64	3.97	9.73	0.000	*
bandlow	4.15	0.34	3.49	4.81	12.34	0.000	*
knowledgeform	0.00	0.47	-0.92	0.92	0.00	1.000	
knowledgemeaning	4.38	0.34	3.73	5.04	13.07	0.000	*
groupSpoken:bandmid	-0.72	0.45	-1.61	0.17	-1.59	0.112	
groupSpoken:bandlow	-0.63	0.45	-1.50	0.25	-1.40	0.161	
groupSpoken:knowledgeform	0.81	0.59	-0.34	1.96	1.39	0.166	
groupSpoken:knowledgemeaning	-0.69	0.44	-1.57	0.18	-1.56	0.119	

bandmid:knowledgeform	-0.68	0.48	-1.63	0.27	-1.41	0.160	
bandlow:knowledgeform	-1.57	0.48	-2.52	-0.63	-3.26	0.001	*
bandmid:knowledgemeaning	-3.97	0.35	-4.65	-3.29	-11.49	0.000	*
bandlow:knowledgemeaning	-6.76	0.36	-7.47	-6.04	-18.52	0.000	*
groupSpoken:bandmid:knowledgeform	-0.39	0.61	-1.58	0.80	-0.64	0.525	
groupSpoken:bandlow:knowledgeform	-0.81	0.61	-2.00	0.37	-1.34	0.180	
groupSpoken:bandmid:knowledgemeaning	0.68	0.46	-0.23	1.59	1.47	0.141	
groupSpoken:bandlow:knowledgemeaning	0.46	0.50	-0.53	1.45	0.92	0.359	

Table 23: Mixed effects model showing a three-way interaction on the self-report pretest, phase two

To summarize, at the beginning of the first phase the spoken group reported more high-frequency and mid-frequency words at the form level compared to the ER-only group. In addition, the written group reported fewer low-frequency words at the form level compared to both the ER-only and spoken groups. When omitting the written group from the data, and combining phase one and two, the results revealed that the ER-only group and spoken group both reported similar degrees of knowledge of the 60 target words.

4.2.4 Were the three groups similar in degree of semantic knowledge of the target words before the intervention?

This section presents the results of the initial state of semantic knowledge for the ER groups. A mixed-effects model was fitted to the data with the cosine similarity value as the dependent variable, and group (ER-only, spoken, written), and target word frequency band (high, mid, low), and as the independent variables. The results of the best-fitting model are shown in Table 24.

The model reveals no significant main effect of group, meaning the three ER groups had similar degrees of semantic knowledge of the target words on the pretest. The table does show a significant effect of frequency band, and to determine where the differences were multiple comparisons of means using Tukey contrasts were computed.

The results, shown in Table 25 reveal that the mid-frequency words had a signifi-

	Estimate	Std..Error	CI95lower	CI95upper	t.value	p.value	
(Intercept)	0.159	0.02	0.13	0.19	10.02	0.000	*
groupSpoken	-0.004	0.01	-0.02	0.01	-0.54	0.589	
groupWritten	0.005	0.01	-0.01	0.02	0.69	0.491	
bandmid	-0.084	0.02	-0.13	-0.04	-3.77	0.000	*
bandlow	-0.101	0.02	-0.15	-0.05	-4.25	0.000	*

Table 24: Results of the mixed effects model for semantic knowledge data in phase one

cantly lower degree of semantic knowledge than the high-frequency words (z = -3.77, $p < 0.01$). The table also reveals that the the low-frequency words have a significantly lower degree of semantic knowledge than the high-frequency words (z = -4.25, $p < 0.01$). There was no difference between the mid-frequency and low-frequency words (z = -0.731, $p = 0.75$). In short, these results suggest that the three ER groups in phase one had similar degrees of initial semantic knowledge. This knowledge was greater in the high-frequency words compared to the mid-frequency and low-frequency words.

	Estimate	Std. Error	CI95lower	CI95upper	z value	Pr(>|z|)	
mid - high	-0.08	0.02	-0.14	-0.03	-3.77	0.000	*
low - high	-0.10	0.02	-0.16	-0.05	-4.25	0.000	*
low - mid	-0.02	0.02	-0.07	0.04	-0.73	0.745	

Table 25: Multiple comparisons of means (Tukey Contrasts) between frequency bands in terms of similartiy in semantic knowledge

Before presenting the results for phase two, it should be noted that a Shapiro-Wilks test computed on the model residuals in Table 24 revealed a non-normal distribution ($W = 0.98$, $p < 0.01$). To ensure that the results presented were robust, permutation testing was adopted using the R package *predictmeans* (Luo, Ganesh, & Collard, 2014). A total of 1,000 permutations were run on the data, and these results, presented in Table 26 confirm that the groups scored similarly, and also the difference in frequency bands.

	Df	Sum Sq	Mean Sq	F value	p value	
group	2	0.0104	0.0052	0.6948	0.517	
band	2	0.1680	0.0840	11.1745	0.001	*

Table 26: Permutation testing for initial semantic knowledge in phase one (1,000 simulations)

To reinforce the results above, the analysis was rerun after combining the data from phases one and two, and omitting the written group. The results, shown in Table 27 reveal the same pattern as the phase one data. No significant difference was found between the ER groups (t = 0.30, p = 0.76). The table shows a significant effect of frequency band, and so multiple comparisons of means using Tukey contrasts were computed to determine where the differences were. The results revealed that the mid-frequency band had significantly lower semantic knowledge than the high-frequency word band (z = -4.498, $p < 0.01$). The results also revealed that the low-frequency words had significantly lower semantic knowledge than the high-frequency words (z = -4.791, $p < 0.01$). There was no difference between the mid-frequency and low-frequency band (z = -0.535, p = 0.85).

	Estimate	Std..Error	CI95lower	CI95upper	t.value	p.value	
(Intercept)	0.167	0.01	0.14	0.20	11.37	0.000	*
groupSpoken	0.002	0.01	-0.01	0.01	0.30	0.760	
bandmid	-0.093	0.02	-0.13	-0.05	-4.50	0.000	*
bandlow	-0.104	0.02	-0.15	-0.06	-4.79	0.000	*

Table 27: Results of the mixed effects model for semantic knowledge data in phase two

Similar to phase one, the mixed effects model used in phase two was not normally distributed as determined by a Shapiro Wilks test (W = 0.95, $p < 0.01$). To ensure the phase two results were robust, permutation testing was computed on the model. These results, displayed in Table 28, confirm that the ER groups were similar in semantic knowledge, and that there was a difference in frequency band.

	Df	Sum Sq	Mean Sq	F value	p value	
group	1	0.0009	0.0009	0.1046	0.744	
band	2	0.2545	0.1273	14.7890	0.001	*

Table 28: Permutation testing for initial semantic knowledge in phase two (1,000 simulations)

4.2.5 Summary of the initial state of knowledge

A number of important results have been revealed at the onset of the reading treatment. The proficiency test results revealed that the phase one data should be analyzed and presented separately from the phase two data since combining the data from both phases, the written group becomes significantly lower in proficiency than the ER-only group. In the remaining sections in this chapter, the phase one data will be presented first, and the trends uncovered will be reinforced by omitting the written group's data and adding the phase two data.

Another important finding is that all of the groups engaged in extracurricular reading to similar degrees, as determined by the reading logs. These logs were implemented to examine the degree to which the time on task for extracurricular reading was similar between the groups, and since the results show no differences, it is unlikely that any significant lexical development could be the result of the extra exposure to English from this extracurricular reading.

A third finding was that the spoken group reported more high-frequency and mid-frequency words at the form level than the ER-only group. This means that the spoken group had more potential for words to develop from form knowledge on the pretest to meaning knowledge on the post-test, or from form knowledge on the pretest to no knowledge on the post-test, compared to the ER-only group. However, combining the data from phases one and two and omitting data from the written group, the ER-only and spoken groups both reported similar degrees of knowledge.

The last finding was that the groups in both phases began the reading intervention

with similar levels of semantic knowledge. This knowledge was greatest in the high-frequency words and similar between the mid-frequency and low-frequency words.

4.3 The nature of the language-related episodes occurring during the Say-it activities

After the initial proficiency testing and after the students sat the pretest, the reading treatment began. As mentioned in section 3.6, the spoken and written groups participated in five Say-it activities, one after finishing each graded reader. These meaning-focused post-reading discussions provided opportunities for the learners to discuss the stories they read, however also allowed for language-focused learning to occur, through language-related episodes. This section presents results detailing the nature of the episodes which took place during these discussions.

Combining the data from phase one and phase two, a total of 769 language-related episodes were located during the Say-it activities. Of these 769 LREs, 455 episodes (59%) were lexical and 314 episodes (41%) were grammatical. An LRE was considered a lexical LRE if it centered on word meaning, spelling, or pronunciation. An LRE was considered grammatical if it centered on other linguistic phenomenon (e.g., word tense). Refer to Table 39 for more information.

The nature of each LRE was determined using the categorization framework shown in Table 15. Applying this framework to each of the LREs revealed a total of 98 unique patterns, of which 31 patterns (32%) occurred in the grammatical LREs, while 89 patterns (91%) occurred in the lexical LREs. Table 29 depicts the three most-common patterns, along with the number of times each pattern occurred and the percentage of the total LREs which had the pattern. The most-common LRE pattern occurred 364 times, a frequency almost eight times greater than the second and third most-common LRE patterns, both of which occurred 46 times. Table 30 illustrates the most-common pattern through an LRE which arose in the spoken group.

During the Say-it activity discussion in which this LRE occurred, L begins assigning

Rank	LRE pattern	Frequency	% of total LREs
1	R-C-S-I-L-Im-xRe-xU-Su	364	47%
2	R-C-S-I-L-De-xRe-xU-Su	46	6%
3	R-C-S-I-L-Im-Pr-U-Su	46	6%

Table 29: The three most-common language-related episode types

L	I ask to you. S2, S2. Ok S2, now you are Bud Corliss[mispronounced], Corliss[corrected]. Talk about your plans to kill Dorothy.

Table 30: The most common LRE pattern

a discussion prompt to their group member when they mispronounce one of the character's surnames. They repeat the surname, this time pronouncing it correctly, and continue assigning the discussion prompt. Referring to Table 29, this LRE is reactive (R) because L reacted to something that she said, instead of for example asking her group for the correct pronunciation of the word. There was no apparent miscommunication in the message that was being conveyed, meaning the LRE was code-related (C). Since the feedback (the correctly-pronounced word) happened only once, or in one turn, the LRE was simple (S). The feedback that was given, in the form of the correctly-pronounced word, was indirect since it was not a meta-linguistic explanation but the word itself (I). In other words, this was a light LRE (L), and the feedback was immediately given (Im). The response, i.e., the correctly pronounced word, was not given to the student, nor did another student elicit the correct pronunciation from L, meaning there was no response (xRe). Since no response was given, there was no opportunity to acknowledge the response, meaning no opportunity for uptake to occur (xU). However, L incorporated the correct pronunciation into her speech meaning that there was successful uptake (Su).

4.3.1 What was the nature of the lexical language-related episodes?

The 455 lexical LREs which occurred during the Say-it activities consisted of pronunciation-focused, meaning-focused, and spelling-focused episodes. Table 31 presents the results for each of these LRE types, including the total amount of each type of episode, the number of unique patterns, the most common pattern, and how many times the pattern occurred along with the percentage of the total LREs for each type. Looking at the first row, it can be seen that there were 253 pronunciation LREs, and 55 unique patterns. The most common pattern is identical to the overall most common pattern, and this pattern occurred 87 times, or in 34% of the 253 total pronunciation-focused episodes.

One thing seen in the table is the large amount of pronunciation-focused LREs compared to meaning-focused and spelling-focused episodes. Interestingly, there were less than half as many meaning-focused LREs as there were pronunciation-focused LREs however the amount of unique patterns were similar. Also interesting is that there were less than half as many spelling-focused LREs as there were meaning-focused LREs, yet there were five times fewer unique spelling-focused patterns (11) than there were unique meaning-focused patterns (53).

Lexical LRE type	Total LREs	Unique patterns	Most common pattern	Pattern frequency (% of Total LREs)
Pronunciation	253	55	R-C-S-I-L-Im-xRe-xU-Su	87 (34%)
Meaning	138	53	R-C-S-I-L-Im-xRe-xU-Su	36 (26%)
Spelling	64	11	R-C-S-I-L-De-xRe-xU-Su	23 (36%)

Table 31: The total lexical LRE types and most common pattern

Looking at the fourth column in Table 31, it can be seen that the most common LRE patterns are identical in the pronunciation-focused and meaning-focused episodes. The most common spelling-focused pattern, interestingly, differed from the pronunciation-focused and meaning-focused episodes only in the timing of the feedback. While the

former two episode types had immediate feedback, the timing of the most-common spelling-based LRE pattern was delayed (De). Table 32 illustrates an example of this type of LRE pattern, taken from the written group, to highlight the delayed nature of the episode.

Student	Turn	Time	Dialogue
J	5	9:04	I just can't believe that my best friend was dead,and **I look around and didin't find anyone** around there.
J	8	9:09	I look around and **did-i-n't** find anyone

Table 32: The most common spelling-based lexical LRE pattern (LRE 280)

The episode is triggered while J is discussing (typing) a response to a discussion prompt in her fifth turn of the discussion. As seen in the emphasized portion, she misspells the word *didn't* as *didin't*. The conversation continues for five minutes, at which point J deletes the unneeded letter *i* which she typed, correcting her mistake.

4.3.2 What was the nature of the grammatical language-related episodes?

As shown in section 4.3, just under half of the LREs were grammatical LREs. The "grammatical" label given to these LREs can be broken down further, focusing on the specific aspect of grammar which was focused on. This section exemplifies these aspects that the learners focused on in the language-related episodes occurring during the Say-it activities.

One type of grammatical LRE focused on location. Table 33 illustrates this type of LRE, an example taken from the spoken group during the second Say-it activity. S1 is recalling an event which occurred in the story. They begin explaining the story and then realize they have used the word **here**, mistakingly telling the story as if their group was there. They then correct the word to **there**, and finish explaining the event.

Another type of grammatical LRE focused on punctuation. Table 34 illustrates an example of this. During the discussion, S1 forms the question "Scott was dead because

Student	Dialogue
S1	Ok. This is S1 speaking. Now I'm Flick, ok? And uh when I I go there for skiing, uh, I I want to go to somewhere that there are less people **here, there.** So I can think about this this case. And uh when I go to the lift, oh there's a man uh behind me and have a gun with with that, with him so I'm very scared at that time. I think I can be ok. Ok?

Table 33: A grammatical LRE focusing on location/proximity (LRE 12)

of the gun?". In their next turn, they realize they did not mean to formulate this as a question, and correct their language as a statement by deleting the question mark and inserting a period. This type of grammatical LRE only occurred in the written group.

Student	Time	Dialogue
S1	9:21	the first time,in our action.Scott was dead because of gun **?**
S1	9:22	the first time,in our action.Scott was dead because of gun **.**

Table 34: A grammatical LRE focusing on punctuation (LRE 51)

In addition to the two previous types of grammatical LREs, another kind involved issues with tense. An example of this is shown in Table 35. During this LRE, occurring in the spoken group during the third Say-it activity, S1 is recalling an event which took place in the third graded reader. Their attention shifts from recalling the event to the tense used to explain what happened in the story; initially they use present tense, but think that the past tense may be more appropriate and so corrects themselves. S2, thinking that S1 is done explaining the event, directs the group's attention to who should be discussing the next prompt (in this case S3), and then begins reading the next discussion prompt.

Another type of grammatical LRE which occurred focused on person. This is illustrated in Table 36. This LRE occurred during the fifth Say-it activity by S1, a member of the spoken group. During their recalling of an event in the story, their focus shifts

Student	Dialogue
S1	Yeah and uh haha yeah and uh **I go to jail. I went to the jail.**
	I go to the prison?
S2	Ok You (..)
S3	Ok ask me.
S2	You are Billy...

Table 35: A grammatical LRE focusing on tense (LRE 80)

from the event they are recalling to the choice between **I** or **we**, and after voicing each, decide to stay with **I**. In the next turn, S2 jumps in to provide a word, and S1 corrects S2, saying that it wasn't simply money, but the father's money.

Student	Dialogue
S1	I **I, we, I love her**, but I don't love her but I love his, her, hers
S2	money
S1	NO, father's money
S2	Oh yeah, father's money

Table 36: A grammatical LRE focusing on person (LRE 144)

Table 36 also illustrates another type of grammatical LRE; gender. During S1's recalling of the relationship between two characters in the graded reader, they use the male possessive **his**, then realize that **his** should be **hers**. After S1 corrects themselves, the conversation continues.

Some of the grammatical LREs which occurred focused on the use of determiners. This is exemplified in Table 37, an LRE which occurred in the spoken group during the first Say-it activity. S1 begins explaining the reasons why the main character, Jojo, left the children's house. During their explanation, the use the definite article **the**, and then they realize that the indefinite article is perhaps more appropriate, so they change to the indefinite article **a**, and then continue with their explanation. In other words, S1's attention shifted from the explanation of the main character's reasons for leaving

the children's house, to deciding whether to use a definite or indefinite article.

Student	Dialogue
S1	And uh, secondly I don't want to uh living with poor people because if i live with poor people I will always have to take care, taken care of by doctors or nurse, I will feel **I'm I'm the poor man, a poor man**, but you, I need to grow up and to be stronger than today so I must be, and I am brave so I need freedom, I need to meet new friends, although maybe in the uh war war time a lot of people get dead, but this is the destiny that I can control the most.

Table 37: A grammatical LRE focusing on person (LRE 208)

Another type of grammatical LRE dealt with certainty. An example of this is shown in Table 38, taken from an LRE occurring in the first Say-it activity, in the spoken group. S1 is recalling how the main character felt when they heard a lorry coming to their village carrying soldiers. While S1 is recalling what the main character felt, they first mention that all of the children in the village are sad and crying. As they start to say this, S1 realizes that **all** might be too strong, and so decides to correct themselves and say **most**.

Student	Dialogue
S1	I'm always sad because I have children that they are homeless, they lose their family and some of them are hurt, uh um, and um **and all, most of them are sad and crying.** Here is a sad place I think.

Table 38: A grammatical LRE focusing on person (LRE 226)

To conclude, just over 40% of the total LREs were grammatical, and upon closer examination, these grammatical LREs comprised a number of specific grammatical

phenomenon. The grammatical LREs thus provided assistance for a wide-variety of grammatical phenomenon. Table 39 summarizes this range of grammatical features.

Grammatical focus	Example
location	here > there
punctuation	” > ’?’
tense	go > went
person	I > we
gender	his > hers
determiner	the > a
certainty	all > most

Table 39: Types of grammatical LREs, with examples

4.3.3 The ’untestable’ nature of language-related episodes

As mentioned in section 3.8.3, a student was tested on a linguistic item if they were the person to display lack of knowledge during an LRE. However, not all of the LREs could be assessed. Table 40 depicts one example of an LRE which could not be tested. In the first turn of the LRE, S1 says that they thought one of the stories in the fourth graded reader was boring. When S2 asks them why, S1 begins explaining why they thought it was boring, and comes upon a gap in their knowledge; they are unsure of a certain word in English. As S1 pauses to think of the word, S2 takes the opportunity to voice their opinion of the story. S1 then takes the floor in the next turn and voices that they are thinking of a **specific word**, with no assistance from their group members. In the final turn, S3 moves the conversation along and there is no more mention of the **specific word**. In this example, even though there was a shift in S1’s attention from the meaning that they were communicating to the language being produced, the **specific word** did not manifest itself, and so could therefore not be tested.

Another type of untestable LRE occurred when a student was responding to a discussion prompt, and became confused between two phrases, both of which were

Student	Dialogue
S1	But I think it's so boring for me
S2	Why
S1	I think maybe **there was some uh-**
S2	I think this story was touching
S1	-uh **some specific word**
S3	No no no why is she saying this

Table 40: An untestable LRE (121)

grammatically correct. Table 41 illustrates this type of LRE. During the Say-it activity, S1 is recalling the main character's encounter with a girl. They stop midway through their sentence to compare two alternatives: *at the* train station and *in the* train station. They end up deciding on *at the* train station, however both phrases are grammatically correct. Even though S1's attention shifted from recalling an experience in a story, to the language being produced to explain the story, there is no linguistic error to assess.

Student	Dialogue
S1	I met this girl **at the uh, in the, at the** train station.

Table 41: An untestable LRE (99)

These two examples demonstrate that an LRE can include a shift in focus from meaning to the language being produced without an explicit language item manifesting itself (Table 40) or without a linguistic error occurring (Table 41). As a result, not all of the LREs were tested; In the first phase of collection, 191 LREs were not tested, meaning that 271 (59%) of the LREs were tested. For phase two, only the lexical LREs were tested, with the exception of two grammatical LREs which focused on two of the target words. This amounted to 168 lexical LREs tested in the second phase, or 55% of the total LREs in phase two excluding any untestable LREs.

4.3.4 Section summary

This section has presented results showing the nature of the language-related episodes which took place during the Say-it activities. Of the 769 total episodes during the discussions, 455 (59%) focused on lexis and 314 (41%) focused on grammar. There were a large amount of patterns found in the episodes, and the most common pattern occurred 364 times (47%). In terms of these patterns, the only difference between the spoken and written groups' patterns was in the spelling-focused episodes, where the written group experienced delayed feedback more often than the spoken group. This is an interesting finding, yet perhaps not surprising given the permanent nature of written discussion, versus the ephemeral nature of speech. Another interesting finding is the large amount of untestable episodes which took place. The untestable nature does not necessarily weaken the episode's facilitative effects on language development; the episodes can still contain processes such as generating hypotheses about language and assessing alternatives which have been found to be conducive to language learning (Swain & Lapkin, 1995).

4.4 How much self-reported development occurred in the ER groups?

4.4.1 Self-report results: Raw scores

This section presents the self-report raw scores for the ER-only group and spoken group on their pretest and post-test. The data was compiled from phases one and two, omitting the 20 off-words for a total of 60 target words. Table 42 presents the results for the ER-only group. As seen in the table, the first row reveals that for the target word *alsatian*, two people reported form knowledge for the word on the pretest (column 2) and the post-test (column 3). Column four shows that zero people reported meaning knowledge of *alsatian* on the pretest, while column five shows that one person reported meaning knowledge on the post-test. The last column shows that *alsatian* occurred

once in the GRs.

word	form-pre	form-post	meaning-pre	meaning-post	GR freq.
alsatian	2	2	0	1	1
arsenic	5	4	1	3	6
audition	10	10	16	20	28
babu	7	6	1	4	2
back	0	0	37	37	127
blouse	10	7	7	14	8
boxing	3	3	32	33	27
campus	2	0	35	37	13
carapace	5	6	1	1	5
closet	11	11	15	18	7
clucking	5	7	2	4	1
copper	7	11	6	13	15
cupboard	5	9	22	23	7
dead	1	2	35	34	57
down	0	0	37	36	137
envious	10	5	3	7	1
felicity	5	8	3	4	2
fell	5	2	29	33	31
flick	8	12	4	4	19
gelatin	4	4	3	4	4
give	0	0	37	37	50
gloves	2	1	28	33	18
goat	4	4	21	22	7
handsome	2	0	33	37	28
hard	0	0	37	36	40
hibiscus	6	7	4	4	1
hopper	12	16	4	7	1
jasmine	6	4	10	13	3
jeep	7	3	27	30	7
kite	4	5	16	24	16
long	0	0	37	37	59
lorry	5	8	5	12	8
name	0	0	37	37	50
near	0	0	37	36	41
need	0	0	37	37	38
papa	5	7	20	21	13
pharmacy	4	2	27	31	15
picket	13	12	10	10	8
prawns	6	5	7	13	4

quiet	0	0	37	37	36
raffle	6	5	2	5	4
read	0	0	37	37	45
right	0	0	37	37	53
round	3	5	32	30	35
rupees	6	5	2	3	5
sahib	3	6	1	0	5
sans	6	4	1	1	1
saris	4	6	1	2	2
shaft	10	13	2	5	17
stop	0	0	37	37	53
stupid	0	0	36	36	34
tell	0	0	36	36	147
thing	0	0	36	37	37
trailer	9	7	13	21	9
turban	8	14	4	7	2
vacation	3	1	32	36	11
vats	6	14	3	4	1
well	0	0	37	37	68
white	0	0	37	37	43
whoosh	6	8	0	3	1

Table 42: Raw scores for the ER-only group's self-report test (n = 37), with target word frequency information

Table 43 illustrates the results for the spoken group. As seen in the table, the first row shows information for the target word *alsatian*. The table depicts that on the pretest, one student reported having form knowledge of the word (in column two). On the post-test, no-one reported form knowledge. Columns four and five show that no-one reported knowing the meaning of *alsatian* on the pretest or the post-test. Column six shows that *alsatian* occurred once in the five GRs, and column seven shows it did not occur in a Say-it activity prompt. Finally, column eight shows that *alsatian* did not occur in an LRE.

word	form-pre	form-post	meaning-pre	meaning-post	GR freq.	Say-it prompt freq.	LRE freq.
alsatian	1	0	0	0	1	0	0
arsenic	4	6	2	5	6	0	0
audition	12	11	5	12	28	0	0
babu	3	8	1	4	2	0	0

back	1	1	25	25	127	0	0
blouse	8	6	6	13	8	0	0
boxing	4	3	22	23	27	0	8
campus	2	0	22	26	13	0	0
carapace	2	4	0	0	5	0	0
closet	10	7	10	15	7	0	0
clucking	5	4	2	3	1	0	0
copper	7	10	7	10	15	0	0
cupboard	7	5	13	18	7	0	0
dead	2	1	22	23	57	1	0
down	1	1	24	25	137	0	0
envious	6	13	5	4	1	0	0
felicity	6	5	3	4	2	0	0
fell	2	1	21	24	31	0	0
flick	1	10	2	2	19	5	5
gelatin	3	2	3	6	4	0	0
give	1	1	25	25	50	0	1
gloves	5	3	16	21	18	0	0
goat	8	5	13	18	7	0	0
handsome	2	1	24	25	28	0	0
hard	1	1	25	25	40	0	0
hibiscus	5	4	0	1	1	0	0
hopper	9	8	2	4	1	0	0
jasmine	7	7	3	10	3	0	0
jeep	4	3	19	22	7	0	0
kite	7	4	13	18	16	2	3
long	1	0	25	25	59	0	0
lorry	6	8	5	15	8	1	3
name	1	0	25	25	50	0	0
near	1	1	24	24	41	0	1
need	1	1	25	25	38	0	3
papa	6	6	11	14	13	0	0
pharmacy	7	4	13	21	15	0	0
picket	8	10	3	9	8	0	0
prawns	5	6	1	4	4	0	0
quiet	1	2	24	24	36	0	1
raffle	1	7	0	0	4	0	0
read	1	1	25	25	45	0	1
right	1	1	25	25	53	0	0
round	5	3	19	21	35	0	0
rupees	3	2	0	2	5	0	0

sahib	1	3	0	2	5	0	0
sans	4	8	1	3	1	0	0
saris	3	8	1	2	2	0	0
shaft	8	11	2	4	17	0	3
stop	1	1	25	25	53	0	0
stupid	2	1	24	25	34	0	0
tell	1	1	23	24	147	1	0
thing	1	1	25	25	37	1	0
trailer	9	8	8	14	9	0	0
turban	3	9	2	4	2	0	0
vacation	2	2	24	24	11	0	0
vats	6	9	3	5	1	0	0
well	1	0	25	25	68	0	0
white	1	1	25	25	43	0	0
whoosh	7	7	1	2	1	0	0

Table 43: Raw scores for the spoken group's self-report test (n = 26), with target word frequency information

Tables 42 and 43 show the degree that knowledge differed for each target word from the pretest to the post-test for the ER-only and spoken groups, respectively. To that extent it can help answer questions such as "How many people reported knowing the meaning of *lorry*?". However, it does not detail *how* the words developed. For example, half-way down table 43, in the row detailing information for the target word *lorry*, it can be seen that six people reported knowing the form on the pretest, and eight reported knowing the form on the post-test. Was this increase the result of two people reporting no knowledge of *lorry* on the pretest and then reporting form knowledge on the post-test? Or, was this increase in form knowledge the result of two people reporting meaning knowledge on the pretest and then reporting form knowledge on the post-test? Tables 42 and 43 do not provide answers for these types of questions, however the next section helps to clarify these questions.

4.4.2 Self-report results: Development

This section reports the results for the self-reported development in lexical knowledge by the three ER groups first for phase one, and then for phases one and two omitting the written group's data. The results in this section incorporate the learning trajectories for each student, for each word. Section 3.8.1 describes these trajectories in detail, but to summarize, using the three self-report levels (i.e., none, form, and meaning), it was possible to assign to each word one of five learning trajectories based on the reported levels a student gave on the pretest and the post-test:

- no change in reported knowledge (NN),
- backward-breadth (BB),
- forward-breadth (FB),
- backward depth (BD), and
- forward-depth (FD).

To determine the extent that the ER groups in phase one reported development to similar degrees, a mixed-effects model was fitted to the data. The dependent variable was target word count, and the independent variables were ER group (ER-only, spoken, written), learning trajectory (NN, BB, FB, BD, FD), and frequency band (high, mid, low). In addition, person was specified as a random variable in the model. Also, since the dependent variable represented counts (of words), the model was fitted assuming a Poisson distribution. The results of the model are shown in Table 44.

	Estimate	Std. Error	CI95lower	CI95upper	z value	Pr(>|z|)	
(Intercept)	2.94	0.04	2.87	3.02	73.88	0.00	*
groupSpoken	-0.08	0.05	-0.19	0.02	-1.61	0.11	
groupWritten	0.02	0.05	-0.08	0.12	0.41	0.68	
trajectoryBB	-3.97	0.29	-4.54	-3.41	-13.78	0.00	*
trajectoryFB	-3.79	0.22	-4.22	-3.35	-17.15	0.00	*
trajectoryBD	-4.75	0.37	-5.47	-4.02	-12.84	0.00	*
trajectoryFD	-4.28	0.27	-4.81	-3.74	-15.67	0.00	*
bandmid	-0.34	0.05	-0.44	-0.24	-6.52	0.00	*
bandlow	-0.27	0.05	-0.37	-0.17	-5.29	0.00	*
groupSpoken:trajectoryBB	-0.46	0.23	-0.91	-0.00	-1.97	0.05	*

groupWritten:trajectoryBB	-0.74	0.26	-1.24	-0.24	-2.91	0.00	*
groupSpoken:trajectoryFB	0.34	0.13	0.08	0.60	2.55	0.01	*
groupWritten:trajectoryFB	0.10	0.14	-0.18	0.38	0.72	0.47	
groupSpoken:trajectoryBD	0.77	0.29	0.21	1.33	2.71	0.01	*
groupWritten:trajectoryBD	-0.76	0.46	-1.66	0.14	-1.65	0.10	
groupSpoken:trajectoryFD	0.93	0.22	0.49	1.36	4.19	0.00	*
groupWritten:trajectoryFD	0.21	0.26	-0.30	0.73	0.80	0.42	
trajectoryBB:bandmid	1.56	0.32	0.93	2.18	4.87	0.00	*
trajectoryFB:bandmid	2.34	0.23	1.90	2.79	10.28	0.00	*
trajectoryBD:bandmid	1.61	0.38	0.86	2.35	4.22	0.00	*
trajectoryFD:bandmid	1.74	0.27	1.21	2.26	6.48	0.00	*
trajectoryBB:bandlow	1.92	0.31	1.32	2.52	6.26	0.00	*
trajectoryFB:bandlow	2.26	0.23	1.81	2.70	9.90	0.00	*
trajectoryBD:bandlow	0.84	0.42	0.02	1.67	2.01	0.04	*
trajectoryFD:bandlow	0.47	0.32	-0.16	1.10	1.46	0.15	

Table 44: Mixed effects model results for self-reported develeopment in Phase one

Table 44 reveals two significant interactions. The first interaction is between ER group and learning trajectory. To determine where the differences were, multiple comparisons of means using Tukey contrasts were computed. The results are shown in Table 45.

contrast	rate.ratio	SE	CI95lower	CI95upper	df	z.ratio	p.value	
ERonly,NN - Spoken,NN	1.09	0.06	0.91	1.30		1.61	0.96	
ERonly,NN - Written,NN	0.98	0.05	0.82	1.17		-0.41	1.00	
Spoken,NN - Written,NN	0.90	0.05	0.74	1.10		-1.77	0.91	
ERonly,BB - Spoken,BB	1.72	0.39	0.80	3.70		2.39	0.52	
ERonly,BB - Written,BB	2.06	0.52	0.88	4.82		2.89	0.20	
Spoken,BB - Written,BB	1.20	0.36	0.44	3.29		0.61	1.00	
ERonly,FB - Spoken,FB	0.77	0.10	0.51	1.17		-2.08	0.75	
ERonly,FB - Written,FB	0.88	0.12	0.57	1.38		-0.93	1.00	
Spoken,FB - Written,FB	1.14	0.16	0.71	1.84		0.96	1.00	
ERonly,BD - Spoken,BD	0.50	0.14	0.19	1.30		-2.45	0.47	
ERonly,BD - Written,BD	2.09	0.95	0.44	9.81		1.61	0.96	
Spoken,BD - Written,BD	4.15	1.87	0.90	19.19		3.16	0.10	
ERonly,FD - Spoken,FD	0.43	0.09	0.21	0.89		-3.93	0.01	*
ERonly,FD - Written,FD	0.79	0.20	0.33	1.90		-0.90	1.00	
Spoken,FD - Written,FD	1.85	0.45	0.80	4.24		2.50	0.44	

Table 45: Results of multiple comparisons for interaction between ER group and learning trajectory, phase one

As seen in the table, the ER-only group reported significantly fewer words in the FD trajectory than the spoken group did in the same trajectory (z = -3.926, p = 0.01). This is the only statistically significant difference. This result suggests that the spoken group reported a greater amount of lexical development in depth of knowledge, however these results should be interpreted with caution considering the results in section 4.2.3 revealed that the spoken group had more potential for this kind of development, reporting a significantly greater amount of words at the form level before the reading intervention commenced.

Table 44 also reveals a significant interaction between learning trajectory and frequency band. To determine where the differences were, multiple comparisons of means using Tukey contrasts were computed and the results are shown in Table 46.

contrast	rate.ratio	SE	CI95lower	CI95upper	z.ratio	p.value	
NN,high - BB,high	79.38	22.64	30.18	208.78	15.34	0.00	*
NN,high - FB,high	38.05	8.05	18.57	77.97	17.20	0.00	*
NN,high - BD,high	114.73	40.07	35.10	375.05	13.58	0.00	*
NN,high - FD,high	49.38	11.84	21.90	111.31	16.27	0.00	*
BB,high - FB,high	0.48	0.17	0.15	1.58	-2.09	0.74	
BB,high - BD,high	1.45	0.65	0.32	6.61	0.82	1.00	
BB,high - FD,high	0.62	0.23	0.18	2.18	-1.29	0.99	
FB,high - BD,high	3.02	1.22	0.76	11.93	2.72	0.29	
FB,high - FD,high	1.30	0.41	0.44	3.79	0.82	1.00	
BD,high - FD,high	0.43	0.18	0.10	1.79	-2.00	0.80	
NN,mid - BB,mid	16.73	2.78	9.53	29.39	16.97	0.00	*
NN,mid - FB,mid	3.65	0.32	2.72	4.91	14.87	0.00	*
NN,mid - BD,mid	23.02	4.75	11.44	46.35	15.20	0.00	*

NN,mid - FD,mid	8.69	1.10	5.65	13.35	17.07	0.00	*
BB,mid - FB,mid	0.22	0.04	0.12	0.40	-8.52	0.00	*
BB,mid - BD,mid	1.38	0.36	0.57	3.31	1.23	1.00	
BB,mid - FD,mid	0.52	0.10	0.26	1.03	-3.26	0.07	
FB,mid - BD,mid	6.31	1.37	3.02	13.15	8.50	0.00	*
FB,mid - FD,mid	2.38	0.34	1.47	3.86	6.07	0.00	*
BD,mid - FD,mid	0.38	0.09	0.17	0.84	-4.14	0.00	*
NN,low - BB,low	11.61	1.62	7.23	18.62	17.58	0.00	*
NN,low - FB,low	3.98	0.35	2.96	5.35	15.87	0.00	*
NN,low - BD,low	49.36	13.40	19.65	123.94	14.36	0.00	*
NN,low - FD,low	30.90	6.75	14.73	64.80	15.71	0.00	*
BB,low - FB,low	0.34	0.05	0.20	0.58	-6.90	0.00	*
BB,low - BD,low	4.25	1.28	1.54	11.77	4.82	0.00	*
BB,low - FD,low	2.66	0.67	1.13	6.28	3.87	0.01	*
FB,low - BD,low	12.39	3.47	4.80	32.00	8.99	0.00	*
FB,low - FD,low	7.76	1.77	3.57	16.84	8.96	0.00	*
BD,low - FD,low	0.63	0.22	0.19	2.01	-1.36	0.99	

Table 46: Results of multiple comparisons for interaction between learning trajectory and frequency band, phase one

In the high-frequency band, there was a significantly greater amount of words in the NN trajectory compared to all other trajectories. In addition, there were no significantly different amounts of words in the remaining four trajectories. In the mid-frequency band, a slightly different picture emerged. Similar to the high-frequency band, there was a significantly greater amount of mid-frequency words in the NN trajectory compared to the other four trajectories. In addition, there were significantly fewer mid-frequency words reported in the BB trajectory than the FB trajectory (z

= -8.518, $p < 0.01$). Put another way, for those words which developed in breadth of knowledge, a significantly greater amount developed from no knowledge to some knowledge (i.e., the FB trajectory), not vice versa (i.e., the BB trajectory). Similarly, there were significantly fewer words reported in the BD trajectory than the FD trajectory (z = -4.141, $p < 0.01$). This means that for those words which developed in depth of knowledge, a significantly greater amount developed from some knowledge to more knowledge (i.e., the FD trajectory), rather than from more knowledge to less knowledge (i.e., the BD trajectory). Another significant finding seen in the table is the significantly greater amount of mid-frequency words which developed in the FB trajectory compared to the FD trajectory ($z = 6.065$, $p < 0.01$). This resonates with the idea that breadth of knowledge tends to develop to a greater extent than depth of knowledge (e.g., Laufer & Nation, 1999). Moving to the low-frequency word band, Table 45 reveals that there was a significantly greater amount of words reported in the NN trajectory compared to the other learning trajectories, a result similar to the high-frequency and mid-frequency word bands. Also similar to the high- and mid-frequency word bands, there was a significantly greater amount of words reported in the FB trajectory compared to the FD trajectory ($z = 8.959$, $p < 0.01$). Interestingly, there were significantly more low-frequency words reported at the BB compared to the BD trajectory ($z = 4.821$, $p < 0.01$), a result found only in the low-frequency word band. This may highlight the unstable nature of vocabulary development, a result which occurs after minimal exposure to a word. In short, the multiple comparisons revealed the incremental nature of vocabulary development, evidence by the majority of the words in each frequency band not developing (i.e., the NN trajectory). The results also resonate with the idea that breadth of knowledge develops to a greater extent than depth of knowledge, yet at the same time both breadth and depth can be unstable after a small number of exposures.

To determine if the trends in phase one persisted with a larger number of participants, the analysis was repeated combining the data from phase one with the data from phase two, and omitting the written group from the dataset. The results of the mixed effect model using this data are shown in Table 47.

	Estimate	Std. Error	CI95lower	CI95upper	z value	Pr(>|z|)	
(Intercept)	2.97	0.03	2.90	3.03	91.08	0.00	*
groupSpoken	-0.05	0.04	-0.12	0.02	-1.36	0.17	
trajectoryBB	-4.25	0.26	-4.76	-3.74	-16.38	0.00	*
trajectoryFB	-4.39	0.25	-4.88	-3.91	-17.65	0.00	*
trajectoryBD	-4.82	0.33	-5.47	-4.16	-14.43	0.00	*
trajectoryFD	-4.34	0.24	-4.82	-3.86	-17.76	0.00	*
bandmid	-0.34	0.04	-0.42	-0.25	-7.51	0.00	*
bandlow	-0.28	0.04	-0.37	-0.20	-6.46	0.00	*
groupSpoken:trajectoryBB	-0.18	0.17	-0.52	0.16	-1.04	0.30	
groupSpoken:trajectoryFB	0.32	0.10	0.11	0.52	3.06	0.00	*
groupSpoken:trajectoryBD	0.08	0.25	-0.40	0.57	0.33	0.74	
groupSpoken:trajectoryFD	0.43	0.16	0.12	0.74	2.72	0.01	*
trajectoryBB:bandmid	1.50	0.29	0.93	2.06	5.16	0.00	*
trajectoryFB:bandmid	2.83	0.26	2.33	3.33	11.05	0.00	*
trajectoryBD:bandmid	1.70	0.36	1.00	2.40	4.75	0.00	*
trajectoryFD:bandmid	2.07	0.25	1.58	2.57	8.20	0.00	*
trajectoryBB:bandlow	1.94	0.28	1.40	2.48	7.03	0.00	*
trajectoryFB:bandlow	2.80	0.26	2.30	3.30	10.93	0.00	*
trajectoryBD:bandlow	0.98	0.39	0.21	1.74	2.51	0.01	*
trajectoryFD:bandlow	1.15	0.28	0.60	1.69	4.14	0.00	*

Table 47: Mixed effects model results for self-reported develeopment in phase two

As seen in the table, there are two significant interactions, the same two interactions that were present in the phase one data. The first interaction was between ER group and learning trajectory. To determine where the differences were, multiple comparisons of means using Tukey contrasts were computed. The results, in Table 48 only depict those interactions which compare across group in each learning trajectory. The table shows two strong trends, the first of which suggests that the ER-only group reported fewer words in the FB trajectory than the spoken group, although this did not reach statistical significance (z = -2.755, p = 0.15). Compared to the phase one results (see Table 46), the phase two results show a stronger trend, suggesting that the spoken group developed their breadth of knowledge to a larger degree than the ER-only group.

The second trend, seen in Table 48, involved the FD trajectory. The trend suggests

that the ER-only group reported fewer words in the FD trajectory than the spoken group, although this did not reach statistical significance (z = -2.466, p = 0.29). Combined with the significant trend in the phase one results, where the trend was significant, the results suggest that the spoken group's depth of knowledge increased to a greater degree compared to the ER-only group.

contrast	rate.ratio	SE	CI95lower	CI95upper	z.ratio	p.value
ERonly,NN - Spoken,NN	1.05	0.04	0.93	1.19	1.36	0.94
ERonly,BB - Spoken,BB	1.26	0.21	0.74	2.16	1.37	0.94
ERonly,FB - Spoken,FB	0.77	0.07	0.57	1.04	-2.75	0.15
ERonly,BD - Spoken,BD	0.97	0.24	0.45	2.10	-0.13	1.00
ERonly,FD - Spoken,FD	0.69	0.10	0.42	1.11	-2.47	0.29

Table 48: Results of multiple comparisons for interaction between ER group and learning trajectory, phase two

The second interaction shown in Table 47 was between learning trajectory and frequency band. To determine where the differences were, multiple comparisons of means using Tukey contrasts were computed, the results of which are shown in Table 49.

contrast	rate.ratio	SE	df	z.ratio	p.value	
NN,high - BB,high	76.5416	19.3535		17.156	<.0001	*
NN,high - FB,high	69.1231	16.8858		17.340	<.0001	*
NN,high - BD,high	118.9169	37.8359		15.018	<.0001	*
NN,high - FD,high	61.7916	14.2904		17.831	<.0001	*
BB,high - FB,high	0.9031	0.3153		-0.292	1.0000	
BB,high - BD,high	1.5536	0.6281		1.090	0.9991	
BB,high - FD,high	0.8073	0.2746		-0.629	1.0000	
FB,high - BD,high	1.7204	0.6864		1.360	0.9905	
FB,high - FD,high	0.8939	0.2984		-0.336	1.0000	

BD,high - FD,high	0.5196	0.2033	-1.674	0.9412	
NN,mid - BB,mid	17.1579	2.5085	19.442	<.0001	*
NN,mid - FB,mid	4.0759	0.3168	18.078	<.0001	*
NN,mid - BD,mid	21.7869	3.5933	18.683	<.0001	*
NN,mid - FD,mid	7.7674	0.7937	20.061	<.0001	*
BB,mid - FB,mid	0.2376	0.0376	-9.081	<.0001	*
BB,mid - BD,mid	1.2698	0.2730	1.111	0.9988	
BB,mid - FD,mid	0.4527	0.0777	-4.618	0.0004	*
FB,mid - BD,mid	5.3453	0.9393	9.539	<.0001	*
FB,mid - FD,mid	1.9057	0.2264	5.427	<.0001	*
BD,mid - FD,mid	0.3565	0.0670	-5.491	<.0001	*
NN,low - BB,low	10.9649	1.2793	20.525	<.0001	*
NN,low - FB,low	4.2084	0.3230	18.725	<.0001	*
NN,low - BD,low	44.7179	10.1483	16.746	<.0001	*
NN,low - FD,low	19.6218	2.9982	19.481	<.0001	*
BB,low - FB,low	0.3838	0.0504	-7.290	<.0001	*
BB,low - BD,low	4.0783	1.0226	5.606	<.0001	*
BB,low - FD,low	1.7895	0.3334	3.123	0.1104	
FB,low - BD,low	10.6258	2.4952	10.064	<.0001	*
FB,low - FD,low	4.6625	0.7660	9.371	<.0001	*
BD,low - FD,low	0.4388	0.1182	-3.057	0.1321	

Results are averaged over the levels of: group

P value adjustment: tukey method for comparing a family of 15 estimates

Tests are performed on the log scale

Table 49: Results of multiple comparisons for interaction between learning trajectory and frequency band, phase two

The results of the multiple comparisons revealed an overall similar pattern to the phase one results. There were significantly more words reported in the NN trajectory than any of the other trajectories, in each frequency band. There were significantly fewer mid-frequency words reported in the BB trajectory compared to the mid-frequency words reported in the FB trajectory (z = -9.081, $p < 0.01$), a result found using the phase one data as well. Another result reinforced with the phase two data is the significantly fewer amount of mid-frequency words reported at the BD trajectory than mid-frequency words reported in the FD trajectory (z = -5.491, $p < 0.01$). The phase two data also reinforces that there were significantly fewer low-frequency words reported at the BB trajectory than the FB trajectory (z = -7.290, $p < 0.01$). Finally, the phase two results show a stronger trend in the low-frequency words in the depth trajectory compared to the phase one results. Specifically, there was a strong trend for fewer low-frequency words to be reported in the BD trajectory than the FD trajectory, although this did not reach statistical significance (z = -3.057, $p = 0.13$). These results suggest that the learning in the mid-frequency and low-frequency word bands, in terms of both breadth and depth, tended to be the result of increased knowledge rather than decreased knowledge. For breadth of knowledge this means that words tended to develop from no knowledge to some knowledge rather than from some knowledge to no knowledge. For depth of knowledge, this means that words tended to develop from some knowledge to more knowledge, rather than from more knowledge to some (less) knowledge.

In short, the results suggest that the spoken group reported more mid-frequency and low-frequency words developing in breadth and depth of knowledge compared to the other two groups. This is strengthened by the fact that the self-report results from the pretest revealed the groups in phase two to report similar levels of knowledge. In other words, even though the two groups started with similar levels of knowledge (see section 4.2.3), the spoken group reported a significantly greater increase in breadth and depth of knowledge than the ER-only group.

4.4.3 How much self-reported development occurred in the target words in LRE triggers?

The previous section reported the results for all of the 60 predetermined target words tracked in the study. Ten of these words arose in an LRE, as a result of a learner struggling with an aspect of the word's knowledge, e.g., spelling. This section presents the results for these ten target-word LRE triggers. There is a further distinction in the students in that some of the students in the spoken group have initiated these LREs, while others have not. The same is true for some of the students in the written group. To account for these students, two additional groups were created; the students in the spoken and written groups were further divided according to whether they were one of the initiators of the LREs. The students in the spoken group who displayed a lack of knowledge of the target words were grouped into the *spoken-lack* group, and those students in the written group who displayed a lack of knowledge of a target word were grouped into the *written-lack* group. The remaining students in the spoken and written groups did not display a lack of knowledge leading to an LRE and so were grouped as *spoken-no-lack* for the students in the spoken group, and *written-no-lack* for the students in the written group. The *ER-only* was the last group. Here is a summary of each of the groups:

1. Spoken group, lack of knowledge (Spoken-lack),
2. Written group, lack of knowledge (Written-lack).
3. Spoken group, no lack of word knowledge (Spoken-no-lack),
4. Written group, no lack of word knowledge (Written-no-lack),
5. ER-only group (ERonly),

If the LRE facilitated the students' knowledge of the target word, then the degree to which the *spoken-lack* and *written-lack* groups' knowledge developed could indicate the effectiveness of the LREs. In other words, the learners in the *spoken-lack* and *written-lack* groups should have reported an increase in reported knowledge if the LREs

facilitated development. To investigate these changes, a mixed effects model was fitted to the data with number of words as the dependent variable, and group, frequency band, and learning trajectory as independent variables. Table 50 presents the results.

	Estimate	Std. Error	CI95lower	CI95upper	z value	Pr(>|z|)	
(Intercept)	0.09	0.24	-0.39	0.57	0.37	0.71	
group2Spoken-no-lack	-0.67	0.72	-2.07	0.73	-0.94	0.35	
group2Spoken-lack	-0.68	0.35	-1.37	0.00	-1.95	0.05	
group2Written-lack	-0.09	1.03	-2.11	1.93	-0.09	0.93	
trajectoryBB	-0.81	0.46	-1.72	0.10	-1.75	0.08	
trajectoryFB	-0.57	0.24	-1.05	-0.10	-2.37	0.02	*
trajectoryBD	-1.03	1.01	-3.01	0.94	-1.03	0.30	
trajectoryFD	-0.70	0.40	-1.48	0.08	-1.75	0.08	
bandmid	0.94	0.27	0.41	1.47	3.48	0.00	*

Table 50: Mixed effects model results for self-reported development of the target words arising as LRE triggers, phase one

As seen in the table, there were no significant interactions between group and trajectory (since there were none included in the final model). This suggests that regardless of whether a student showed lack of knowledge of a word or not, they reported the same degree of learning trajectories. A note of caution is due however; interpretation of this model should be made cautiously considering the low number of students in the groups (the written-lack group only had three students). To determine if adding more students would change these phase one results, the self-report data from phase one and phase two were combined, and the written group was omitted. These phase two results are shown in Table 51. As can be seen in the table, a similar pattern emerges with the phase two data. Taken together with the phase one data, this suggests that the LREs may have had limited effect on self-reported knowledge.

	Estimate	Std. Error	CI95lower	CI95upper	z value	Pr(>|z|)	
(Intercept)	0.42	0.15	0.13	0.72	2.83	0.00	*
group2Spoken-no-lack	-0.76	0.71	-2.16	0.63	-1.07	0.28	
group2Spoken-lack	-0.71	0.34	-1.38	-0.04	-2.06	0.04	*
trajectoryBB	-0.86	0.39	-1.63	-0.09	-2.20	0.03	*
trajectoryFB	-0.65	0.20	-1.05	-0.25	-3.17	0.00	*
trajectoryBD	-1.01	0.42	-1.84	-0.19	-2.41	0.02	*
trajectoryFD	-0.69	0.32	-1.32	-0.07	-2.18	0.03	*
bandmid	0.59	0.18	0.24	0.94	3.32	0.00	*

Table 51: Mixed effects model results for self-reported development of the target words arising as LRE triggers, phase two

4.4.4 Section summary: self-reported development

Combining the results from phase one and phase two, this section has presented a number of findings related to the three ER groups' self-reported development. First, the results showed that the majority of words were reported to not change in knowledge. In addition, the high-frequency and mid-frequency words both had more forward development in breadth and depth than they did negative development. For the low-frequency band, there were more words reported to change in backward breadth of knowledge compared to forward breadth. This result may be highlighting the unstable nature of vocabulary learning with limited exposures to the word.

Regarding the ER groups, the results revealed that the spoken group reported more development in (forward) depth of knowledge compared to the ER-only group. A strong trend was also present showing the spoken group to report a greater amount of development in (forward) breadth of knowledge compared to the ER-only group.

Finally, the results looking at only the ten target words arising in LRE triggers suggested that the LREs may not have had a strong facilitative effect on the self-reported development knowledge of the students. Yet, this finding should be taken

cautiously considering the low number of target words which arose in LRE triggers. The next section incorporates the learning trajectories used in this section and presents the results showing the extent to which the three ER groups developed their semantic (associational) knowledge of the target words.

4.5 How much semantic knowledge development occurred in the ER groups?

This section presents the results for the word association data. Similar to the previous sections, the results are presented twice; first using the data from phase one and then again with the phase two data. The results show the extent that the groups' semantic knowledge developed for the 60 target words. To determine the extent that semantic knowledge developed, a mixed effects model was fitted to the data using the *lmer* function in the R package (*lme4*). The dependent variable in the model was cosine similarity value (see section 3.8.2 for an explanation of this statistic). This value represented the degree of change in the similarity value; a positive value indicates an increase in knowledge from the pretest to the post-test, while a negative value indicates a decrease in knowledge from the pretest to the post-test. The independent variables in the model were group (ER-only, spoken, written), learning trajectory (NN, BB, FB, BD, FD), and frequency band (high, mid, low). In addition, person and target word were specified as random variables in the model. The results of the best-fitting model for phase one are shown in Table 52.

	Estimate	Std..Error	CI95lower	CI95upper	t.value	p.value	
(Intercept)	0.007	0.01	-0.00	0.02	1.17	0.242	
groupSpoken	-0.004	0.01	-0.02	0.01	-0.48	0.632	
groupWritten	-0.017	0.01	-0.03	-0.00	-2.08	0.038	*
trajectoryBB	-0.209	0.03	-0.27	-0.15	-6.81	0.000	*
trajectoryFB	0.194	0.02	0.15	0.24	7.96	0.000	*
trajectoryBD	-0.019	0.04	-0.10	0.06	-0.47	0.639	
trajectoryFD	0.034	0.03	-0.02	0.09	1.31	0.191	
bandmid	0.001	0.01	-0.01	0.01	0.14	0.888	
bandlow	0.007	0.01	-0.02	0.03	0.52	0.604	

trajectoryBB:bandmid	0.147	0.04	0.07	0.22	3.91	0.000	*
trajectoryFB:bandmid	-0.138	0.03	-0.19	-0.09	-5.18	0.000	*
trajectoryBD:bandmid	0.021	0.05	-0.07	0.11	0.46	0.642	
trajectoryFD:bandmid	-0.006	0.03	-0.06	0.05	-0.20	0.845	
trajectoryBB:bandlow	0.124	0.04	0.05	0.20	3.22	0.001	*
trajectoryFB:bandlow	-0.143	0.03	-0.20	-0.09	-4.86	0.000	*
trajectoryBD:bandlow	0.028	0.05	-0.07	0.13	0.54	0.588	
trajectoryFD:bandlow	-0.054	0.04	-0.13	0.02	-1.41	0.157	

Table 52: Mixed effects model results for semantic knowledge data in phase one.

The results in the table reveal that there was no significant interaction involving the ER groups. This suggests that the groups' semantic knowledge developed similarly at each level of each independent variable. Table 52 shows a significant main effect of group, however multiple comparisons of means using Tukey contrasts revealed that while there was a trend showing the written group to have a lower degree of semantic knowledge compared to the ER-only group, this trend did not reach statistical significance (z = 2.076, p = 0.09). The table also reveals a significant interaction between learning trajectory and frequency band. To determine where these differences were, multiple comparisons of means were computed using Tukey contrasts. The results of these comparisons are shown in Table 53. The high-frequency words reported in the BB trajectory had significantly lower cosine values compared to the mid-frequency and low-frequency words reported in the same trajectory (z = -3.99, $p < 0.01$; z = -3.62, $p < 0.01$, respectively). This result makes sense given that 4.2.4 revealed that the high-frequency words had significantly higher semantic knowledge than the mid- and low-frequency words, and since the change in semantic knowledge for those words in the BB learning trajectory was derived from taking negative pretest value, it makes sense that the high-frequency words would be lower on the post-test in the BB trajectory.

The multiple comparisons in Table 53 also reveal a significant interaction in the forward-breadth (FB) trajectory. The high-frequency words in this trajectory had significantly greater semantic knowledge than the mid-frequency and low-frequency words

in the FB trajectory ($z = 5.27$, $p < 0.01$; $z = 5.15$, $p < 0.01$, respectively). In other words, the high-frequency words which were unknown on the pretest and known to some degree on the post-test developed a greater degree of semantic knowledge than the mid-frequency and low-frequency words unknown on the pretest and known to some degree on the post-test.

	Estimate	Std. Error	CI95lower	CI95upper	z value	Pr(>|z|)	
NN.high - NN.mid	-0.00	0.01	-0.02	0.02	-0.14	1.000	
NN.high - NN.low	-0.01	0.01	-0.05	0.03	-0.52	1.000	
NN.mid - NN.low	-0.01	0.01	-0.05	0.04	-0.43	1.000	
BB.high - BB.mid	-0.15	0.04	-0.25	-0.04	-3.99	0.001	*
BB.high - BB.low	-0.13	0.04	-0.24	-0.03	-3.62	0.004	*
BB.mid - BB.low	0.02	0.03	-0.07	0.10	0.56	1.000	
FB.high - FB.mid	0.14	0.03	0.06	0.21	5.27	0.000	*
FB.high - FB.low	0.14	0.03	0.06	0.21	5.15	0.000	*
FB.mid - FB.low	-0.00	0.01	-0.04	0.04	-0.06	1.000	
BD.high - BD.mid	-0.02	0.05	-0.15	0.11	-0.49	1.000	
BD.high - BD.low	-0.03	0.05	-0.18	0.11	-0.71	0.999	
BD.mid - BD.low	-0.01	0.03	-0.11	0.08	-0.37	1.000	
FD.high - FD.mid	0.00	0.03	-0.08	0.09	0.17	1.000	
FD.high - FD.low	0.05	0.04	-0.06	0.15	1.31	0.912	
FD.mid - FD.low	0.04	0.03	-0.04	0.12	1.49	0.820	

Table 53: Results of multiple comparisons for interaction of learning trajectory and target word frequency band, phase one

The results presented above were from the first phase of collection, revealing that the three groups' semantic knowledge developed to similar degrees over the course of the reading. To determine if these results prevailed after increasing the number of data points, data from phases one and two were combined, the data from the written group was omitted, and the analysis was redone. A mixed effects model was fitted to the data using the *lmer* function in R, with the cosine similarity value as the dependent variable, and ER group (ER-only, spoken), learning trajectory (NN, BB, FB, BD, FD), and frequency band (high, mid, low). The results of the best-fitting model are presented in Table 54.

	Estimate	Std..Error	CI95lower	CI95upper	t.value	p.value	
(Intercept)	0.001	0.01	-0.01	0.01	0.20	0.845	
groupSpoken	-0.011	0.01	-0.02	0.00	-1.88	0.060	
trajectoryBB	-0.194	0.03	-0.25	-0.13	-6.28	0.000	*
trajectoryFB	0.172	0.03	0.12	0.23	5.88	0.000	*
trajectoryBD	-0.017	0.04	-0.10	0.06	-0.42	0.677	
trajectoryFD	0.007	0.03	-0.05	0.06	0.23	0.820	
bandmid	0.012	0.01	-0.00	0.03	1.88	0.061	
bandlow	-0.004	0.01	-0.03	0.02	-0.30	0.765	
groupSpoken:trajectoryBB	0.024	0.03	-0.03	0.08	0.86	0.387	
groupSpoken:trajectoryFB	0.039	0.01	0.01	0.07	2.68	0.007	*
groupSpoken:trajectoryBD	0.043	0.03	-0.02	0.10	1.44	0.150	
groupSpoken:trajectoryFD	0.028	0.02	-0.01	0.07	1.50	0.134	
trajectoryBB:bandmid	0.123	0.04	0.05	0.19	3.39	0.001	*
trajectoryFB:bandmid	-0.122	0.03	-0.18	-0.06	-4.02	0.000	*
trajectoryBD:bandmid	-0.005	0.04	-0.09	0.08	-0.11	0.916	
trajectoryFD:bandmid	0.016	0.03	-0.04	0.07	0.55	0.582	
trajectoryBB:bandlow	0.146	0.04	0.08	0.22	4.08	0.000	*
trajectoryFB:bandlow	-0.126	0.03	-0.19	-0.06	-3.88	0.000	*
trajectoryBD:bandlow	0.011	0.05	-0.08	0.11	0.23	0.818	
trajectoryFD:bandlow	0.010	0.03	-0.06	0.08	0.29	0.775	

Table 54: Results of mixed effects model assessing semantic knowledge development of the 60 targer words, phase two

The table reveals a significant interaction involving ER group and learning trajectory, an interaction not present in the results shown above using the phase one data. Multiple comparisons of means using Tukey contrasts, shown in Table 55, revealed that the Spoken group had significantly greater semantic knowledge of the words reported at the FB learning trajectory compared to the ER-only group ($t = 1.98$, $p = 0.048$). Put another way, the results suggest that the students in the spoken group developed semantic knowledge of the words in the FB trajectory to a greater extent than the students in the ER-only group developed their knowledge of the words in the FB trajectory.

Table 54 also reveals a significant interaction between learning trajectory and frequency band, an interaction also found in the results in phase one. To determine where the differences were, multiple comparisons of means using Tukey contrasts were com-

	Estimate	Std. Error	t.value	p.value	
ERonly NN - Spoken NN	0.01	0.01	1.88	0.063	
ERonly BB - Spoken BB	-0.01	0.03	-0.46	0.648	
ERonly FB - Spoken FB	-0.03	0.01	-1.98	0.048	*
ERonly BD - Spoken BD	-0.03	0.03	-1.07	0.287	
ERonly FD - Spoken FD	-0.02	0.02	-0.92	0.359	

Table 55: Results of multiple comparisons examining interaction effects between ER group and learning trajectory, phase two.

puted. The results of the multiple comparisons are depicted in Table 56. Similar to the results from the phase one analysis, the mid-frequency and low-frequency words in the BB trajectory had significantly greater semantic knowledge than the high-frequency words in the BB trajectory ($t = 3.77$, $p < 0.01$; $t = 4.18$, $p < 0.01$, respectively). In addition, Table 56 reveals that the semantic knowledge of the high-frequency words in the FB trajectory developed to a greater extent compared to the mid- and low-frequency words in the FB trajectory ($t = 3.66$, $p < 0.01$; $t = 4.26$, p < 0.01, respectively). These results reinforce the results from phase one.

	Estimate	Std. Error	t.value	p.value	
NN high - NN mid	-0.01	0.01	-1.88	0.067	
NN high - NN low	0.00	0.01	0.30	0.766	
NN mid - NN low	0.02	0.01	1.27	0.206	
BB high - BB mid	-0.14	0.04	-3.77	0.000	*
BB high - BB low	-0.14	0.03	-4.18	0.000	*
BB mid - BB low	-0.01	0.03	-0.24	0.812	
FB high - FB mid	0.11	0.03	3.66	0.000	*
FB high - FB low	0.13	0.03	4.26	0.000	*
FB mid - FB low	0.02	0.01	1.39	0.164	
BD high - BD mid	-0.01	0.04	-0.17	0.861	

Continued on next page

	Estimate	Std. Error	t.value	p.value
BD high - BD low	-0.01	0.05	-0.16	0.874
BD mid - BD low	0.00	0.03	0.01	0.995
FD high - FD mid	-0.03	0.03	-0.99	0.321
FD high - FD low	-0.01	0.03	-0.19	0.852
FD mid - FD low	0.02	0.02	1.05	0.296

Table 56: Results of multiple comparisons examining interaction effects between learning trajectory and frequency band, phase two.

To summarize this section, semantic knowledge of the high-frequency words in the BB learning trajectory showed a greater decrease compared to the mid-frequency and low-frequency words. As mentioned in section 3.8.2, the cosine value for the words in the BB trajectory were arrived at by taking the negative value of the pretest cosine value. Coupled with the finding that the high-frequency words had the highest degree of semantic knowledge compared to the mid-frequency and low-frequency words, it is not surprising that the high-frequency words in the BB trajectory had the lowest semantic knowledge. In addition, the high-frequency words in the FB learning trajectory showed a greater increase in semantic knowledge compared to the mid-frequency and low-frequency words in the FB trajectory. This suggests that when a high-frequency word changes from no knowledge to some knowledge, it is often accompanied by a large increase in semantic knowledge.

Another finding, perhaps of greater importance, was the interaction between ER group and learning trajectory which revealed that the spoken group gained significantly more semantic knowledge than the ER-only group for those words reported in the FB trajectory. At the beginning of this chapter, the two groups were deemed similar in terms of English proficiency, amount of extracurricular reading, initial amount of words reported at each self-report level, and initial state of semantic knowledge of the

target words. That they were similar in each of these areas focuses attention to one important difference between the groups that has yet to be investigated; the effects of the language-learning episodes (LREs). As mentioned in section 4.4.3, if the LREs facilitated knowledge development of the target words, then this would be most easily visible in those target words which arose as LRE triggers. This is because the language production occurring during these episodes necessitates focus on both form knowledge and meaning knowledge, and this increased focus may facilitate learning (Ellis, 2003). To determine the extent to which this is true, the next section focuses on the change in semantic knowledge of the target words which arose as triggers in the language-related episodes.

4.5.1 How much semantic knowledge development occurred in target words in LRE triggers?

This section presents the results for the ten target words which arose in language-related episodes due to a student struggling with some aspect of their knowledge. At the beginning of section 4.4.3, five groups were established to assess the effectiveness of the episodes for phase one:

1. ER-only,
2. Spoken-no-lack,
3. Spoken-lack,
4. Written-no-lack, and
5. Written-lack.

If the episodes facilitated lexical development, then those students in the spoken-lack and written-lack groups should have a greater amount of semantic knowledge than those students who didn't display a lack of knowledge. To investigate this, a mixed effects model was fitted to the data. The dependent variable was cosine similarity value, and the independent variables were ER group (ER-only, spoken-no-lack, spoken-lack, written-no-lack, and written-lack), frequency band (high, mid, low), learning trajectory

(NN, BB, FB, BD, FD). In addition, person and target word were specified as random effects in the model. The results of the best-fitting model are presented in Table 57.

	Estimate	Std..Error	CI95lower	CI95upper	t.value	p.value	
(Intercept)	-0.003	0.01	-0.08	0.03	-0.27	0.788	
group2Spoken-no-lack	0.004	0.02	0.02	0.43	0.26	0.797	
group2Written-no-lack	-0.027	0.02	0.00	0.00	-1.57	0.116	
group2Spoken-lack	-0.184	0.07	-0.32	-0.02	-2.47	0.013	*
group2Written-lack	0.015	0.07	-0.16	0.25	0.21	0.836	
bandmid	0.005	0.02	-0.05	0.08	0.20	0.839	
trajectoryBB	-0.004	0.11	-0.45	-0.16	-0.03	0.972	
trajectoryFB	0.065	0.02	0.03	0.15	2.93	0.003	*
trajectoryBD	0.005	0.11	-0.27	0.14	0.05	0.962	
trajectoryFD	0.186	0.11	0.05	0.22	1.77	0.077	
group2Spoken-no-lack:bandmid	0.022	0.03	0.03	0.38	0.79	0.432	
group2Written-no-lack:bandmid	0.059	0.03	0.00	0.00	1.99	0.047	*
group2Spoken-lack:bandmid	0.244	0.08	0.00	0.00	2.89	0.004	*
bandmid:trajectoryBB	-0.233	0.12	0.00	0.00	-1.90	0.057	
bandmid:trajectoryBD	-0.021	0.12	0.00	0.00	-0.18	0.857	
bandmid:trajectoryFD	-0.115	0.11	0.00	0.00	-1.04	0.296	

Table 57: Mixed effects model results for target words in LRE triggers, phase one.

The table reveals a significant interaction between ER group and frequency band. Multiple comparisons of means using Tukey contrasts were computed to determine where the differences were. The comparisons of means revealed that the ER-only group had significantly greater semantic knowledge of the high-frequency words compared to the spoken-lack group ($t = 2.47$, $p = 0.01$). The comparisons also revealed that the spoken-no-lack group had significantly greater semantic knowledge than the spoken-lack group ($t = 2.51$, $p = 0.01$) for the high-frequency words. Finally, the comparisons revealed that the written-no-lack group had a significantly greater degree of semantic knowledge than the spoken-lack group ($t = 2.10$, $p = 0.36$). Taken together, these results suggest that the facilitative nature of the LREs may not extend to high-frequency words.

The phase one results revealed interactions between ER group and frequency band,

which when probed further revealed that the ER-only group and the spoken-no-lack groups had significantly greater semantic knowledge than the spoken-lack group for those words which arose as an LRE trigger. This suggests that the language-related episodes may have not facilitated lexical development. However, the results described in this section use a small number of participants. To determine if these trends remained with data from more participants, the data from phase one and two were combined, and the data for the written group was omitted and the analyses were recomputed. The mixed effects model was fitted to the data with cosine similarity as the dependent variable, and the independent variables were ER group (ER-only, spoken-no-lack, spoken-lack), frequency band (high, mid, low), and learning trajectory (NN, BB, FB, BD, FD). In addition, person and target word were specified as random effects in the model. The results of this model are presented in Table 58.

	Estimate	Std..Error	t.value	p.value	
(Intercept)	-0.005	0.01	-0.54	0.590	
group2Spoken-no-lack	-0.013	0.01	-0.97	0.331	
group2Spoken-lack	-0.086	0.06	-1.47	0.143	
bandmid	0.010	0.02	0.56	0.573	
trajectoryBB	-0.146	0.05	-2.76	0.006	*
trajectoryFB	0.091	0.02	4.33	0.000	*
trajectoryBD	-0.034	0.04	-0.91	0.361	
trajectoryFD	0.103	0.03	3.43	0.001	*
group2Spoken-no-lack:bandmid	0.042	0.02	1.85	0.065	
group2Spoken-lack:bandmid	0.199	0.07	2.97	0.003	*

Table 58: Mixed effects model results for target words in LRE triggers, phase two.

The results in the table reveal a significant interaction between ER group and frequency band. To determine where the differences were, multiple comparisons of means were computed using Tukey contrasts. The results of the comparisons are shown in Table 59, revealing two significant differences. First, the results reveal that the ER-only

group's semantic knowledge of the mid-frequency words was significantly lower than the spoken-lack group's semantic knowledge of the mid-frequency words (t = -3.50, p = 0.001). Second, the table shows that the Spoken-no-lack group also had significantly lower semantic knowledge of the mid-frequency words than the spoken-lack group had of the mid-frequency words (t = -2.49, p = 0.013). Put another way, the students in the Spoken-lack group, i.e., the group of learners who struggled with a target word and triggered a language-related episode, had significantly greater semantic knowledge than the other two groups. That the students in the Spoken-lack group had a significantly higher degree of semantic knowledge reinforces the language-learning benefits of the language-related episodes. Figure 6 illustrates these findings, showing the average cosine value for each of the groups. The numbers above each of the bars represents the N-size for that group.

	Estimate	Std. Error	t.value	p.value	
ERonly high - Spoken-no-lack high	0.01	0.01	0.97	0.331	
ERonly high - Spoken-lack high	0.09	0.06	1.47	0.143	
Spoken-no-lack high - Spoken-lack high	0.07	0.06	1.24	0.217	
ERonly mid - Spoken-no-lack mid	-0.03	0.02	-1.58	0.116	
ERonly mid - Spoken-lack mid	-0.11	0.03	-3.50	0.001	*
Spoken-no-lack mid - Spoken-lack mid	-0.08	0.03	-2.49	0.013	*

Table 59: Results of multiple comparisons for the interaction between group and frequency band, phase two data, for the target word LRE triggers.

4.5.2 Section summary: semantic knowledge development

This section has provided a large amount of findings regarding the extent that the groups' semantic knowledge developed. The key finding in this section is that those students in the spoken group who displayed lack of knowledge of a mid-frequency target

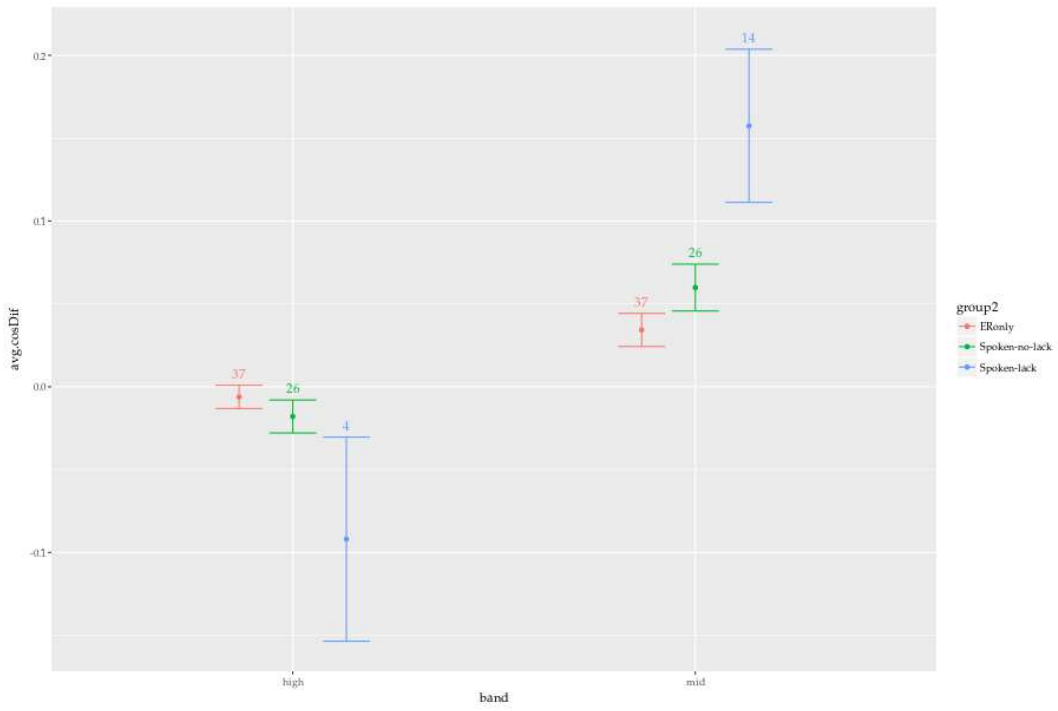

Figure 6: Average change in semantic knowledge of target words arising in LREs as triggers (error bars are +/- 1 standard error) phase two

word, triggering a language-related episode had significantly higher semantic knowledge than the ER-only group as well as those students in the spoken group who did not display lack of knowledge of the mid-frequency target words. That both spoken groups, i.e., the spoken-no-lack and spoken-lack groups, scored significantly higher in semantic knowledge than the ER-only group suggests that the Say-it activities were conducive to this kind of semantic knowledge development. Since the spoken-lack group had significantly greater semantic knowledge of the words which they displayed lack of knowledge of during the Say-it activities compared to the spoken-no-lack group further suggests that the language-related episodes provided another level of facilitation in addition to discussing the stories. What it is about the episodes that may have been conducive to development will be discussed in the next chapter.

4.6 To what extent did the groups answer the LRE-based questions correctly on the post-test?

The LRE-based questions assessed those words which a learner struggled with during the Say-it activities, resulting in an LRE. To determine the amount of learning which occurred for the lexical LRE-based questions, the data from phase one and phase two were combined, resulting in data from 37 students in the ER-only group, 29 students in the spoken group, and 10 in the written group. Similar to Loewen (2005), those LRE-based questions which were scored as *correct*, *partially-correct*, or *other correct* were all considered correct since they all had aspects of correctness. These answers were given a score of one. The incorrectly answered questions were given a score of 0. The four-point scoring system described in section 3.8.3 will be used in the next chapter to help paint a picture of possible explanations for the learning which took place. A total of 395 LREs were tested, with 306 being lexical LREs, and 89 being grammatical LREs. Within the lexical LREs, 184 were pronunciation-based, 42 were spelling-based, and 80 were meaning-based. See section 4.3.3 for an explanation of why all 769 LREs were not tested.

Table 60 presents descriptive statistics for the LRE-based questions. The table is divided according to ER group and LRE focus, and depicts the total amount of correct answers each group in each category, the total questions asked in each category, and the percentage correct. As seen in the table, the majority of the LREs were pronunciation questions, followed by grammar-based questions, meaning-based questions and finally spelling-based questions (see section 3.8.3 for an explanation of how the LREs were categorized). The ER-only group scored higher than both the spoken and written groups in the grammatical questions, followed by the spoken group and then the written group. The ER-only group scored 85% on the pronunciation-based questions, higher than the spoken group's score of 74%. Not surprisingly, the written group had the most spelling-focused LRE questions, answering 86% of the 42 correct, while the ER-only group answered 14. The spoken group had the greatest number of meaning-focused

questions (96), answering 69% of them correctly. These results shed light on the extent that learning occurred, at the same time however it is difficult to compare these scores across groups. It should be kept in mind that these results are a combination of different words. Additionally, these results combine LRE items of which the majority were answered by only one person, i.e., the person struggling with the item. However, these results will be returned to in the next chapter, when discussing the effectiveness of the LREs qualitatively, to provide a better understanding of the facilitating nature of the LREs.

group	focus	Correct	Total	Percent
ERonly	Grammar	53	70	76%
Spoken	Grammar	75	106	71%
Written	Grammar	23	37	62%
ERonly	Pronunciation	117	138	85%
Spoken	Pronunciation	202	273	74%
ERonly	Spelling	14	14	100%
Spoken	Spelling	1	3	33%
Written	Spelling	36	42	86%
ERonly	Meaning	35	69	51%
Spoken	Meaning	66	96	69%
Written	Meaning	1	1	100%

Table 60: Scores for the lexical LRE-based questions

4.7 How much learning occurred in the C-test?

The C-test was implemented in phase two, as a means of assessing an additional type of productive vocabulary knowledge, that is, productive knowledge of form-meaning. The mid-frequency words were utilized for this test since the majority of the target words which arose in the language-related episodes were mid-frequency words. A 20-

item test was created with a format similar to Laufer & Nations's (1999) Productive Vocabulary Levels Test. To determine the extent to which the two groups differed in their performance on the test, a mixed effects model was fitted to the data with score for each question (1 for correct and 0 for incorrect) as the dependent variable, and group (ER-only and spoken) and test (pretest and post-test) as the independent variables. In addition, person and target word were specified as random effects in the model. The results of the model revealed significant improvement from the pretest to the post-test ($z = 6.147$, $p < 0.0001$). The results also revealed that the spoken and ER-only groups did not score significantly different from each other ($z = 0.90527$). In short, the results suggest that the two groups performed similarly on the C-test, both groups improving to a similar degree from the pretest to the post-test.

It should be noted however, that these results include both those words which arose in language-related episodes as well as those that did not. To determine the extent that learning occurred in the target words arising in the episodes, the relevent C-test data was subset and the analysis was rerun. These results, shown in Table 61, revealed a significant effect of group and test. Multiple comparisons of means using Tukey contrasts revealed that the spoken-lack group, i.e., those students who struggled with a target word which triggered a language-related episode, score significantly higher than the ER-only group ($z = 2.773$, $p = 0.01342$) as well as the spoken-no-lack group ($z = 3.056$, $p = 0.00552$). In addition, Table 61 reveals that the post-test scores were significantly higher than the pretest scores ($z = 6.18$, $p < 0.001$). Taken together, these results suggest that all three groups increased from the pretest to the post-test, however those students who struggled with a target word which triggered a language-related episode scored significantly higher than both the ER-only group and those students in the spoken group who did not display lack of knowledge.

	Estimate	Std. Error	CI95lower	CI95upper	z value	Pr(>|z|)	
(Intercept)	-1.27	0.46	-2.16	-0.37	-2.77	0.006	*
group2Spoken-no-lack	-0.09	0.27	-0.61	0.43	-0.35	0.728	
group2Spoken-lack	1.89	0.68	0.55	3.22	2.77	0.006	*
testpost	0.93	0.15	0.64	1.23	6.18	0.000	*

Table 61: Mixed effects model results for the C-test, for the target words arising in language-related episodes in phase two

4.8 Chapter summary

This chapter presented results to help answer the three research questions motivating the current study (section 2.7 lists the research questions in full). The first research focused on the effects of two approaches to extensive reading and their effects on vocabulary development: a reading-only approach and an approach which supplemented reading with post-reading discussions. The results in this chapter revealed that both approaches facilitated lexical development with regards to the 60 target words tracked over the course of the study. At the same time however, the knowledge of a large portion of these target words remained at their initial state of reported knowledge, highlighting the incremental nature of lexical development. Similarly some of the target words increased in reported knowledge while others decreased in knowledge, reinforcing the fact that learning a word is not an all or nothing phenomenon, but a process (Nation, 2013a).

The second research question focused on the supplementary activities and the facilitating effects that these activities can have on lexical development. The results revealed that the activities provided opportunities for both the spoken and written groups to develop their vocabulary knowledge, although the written condition was deemed unsuitable for the current research design and as a result this condition was omitted. The development found spanned multiple aspects of word knowledge including knowledge of meaning, spelling and pronunciation and included increases in both breadth and depth

of knowledge. The results showed that the students in the spoken group increased their semantic knowledge of the mid-frequency target words to a greater degrees than the written group and the ER-only group. This increase was most easily seen in those words which were the focus of a language-related episode which confirms the hypothesis stated in section 3.8.2. Additionally, there were a large amount of language-related episodes which focused on words which were not part of the 60 target words. That aspects of knowledge of these additional words were also acquired emphasizes the importance of providing learners with the opportunity to interact with each other through meaning-focused output activities.

The third research question centered on the additional aspects of language which the learners focused on during their interactions while engaged in the Say-it activities. The results revealed that the learners explicitly focused on a variety of language aspects including verb tense, punctuation, and determiner usage (see Table 39 for a list of the phenomena). Coupled with the results for the second research question detailed in the previous paragraph, the results presented in this chapter reveal that the post-reading discussions provided an environment ripe with language learning opportunities. The next chapter expands on these findings in more detail, positing possible explanations for the facilitating nature of the post-reading discussions by focusing on the language-learning episodes which centered on mid-frequency target words.

5 Discussion

5.1 Introduction

The results in Chapter 4 revealed that a traditional reading-only approach to extensive reading and an approach to extensive reading which is supplemented with post-reading discussions both led to lexical development. However, a task-based approach which included supplementary activities provided additional opportunities for language development and consolidation to occur, and this was most clearly seen in the spoken group with the mid-frequency target words. This chapter details possible explanations for the spoken group's increased development by incorporating information including:

- the nature of the language-related episodes which centered on a mid-frequency target word,
- the target word itself, including frequency information and contexts in which it appeared in the graded readers, and
- pretest and post-test data of the learners who struggled with an aspect of lexical knowledge triggering a language-related episode.

Drawing from the issues mentioned in the literature review of this thesis (chapter 2), a number of possible confounding variables which may have affected lexical development are discussed in section 5.2. Section 5.2.1 looks at the effects that learner proficiency may have had on lexical development. Section 5.2.2 discusses the possibility that time on task was responsible for the spoken group's development. Section 5.2.3 discusses the role that initial knowledge played in the development found in chapter 4. Section 5.2.4 posits reasons for why the mid-frequency words showed the greatest gains in knowledge. After these confounding variable are discussed, section 5.3 reveals that for some words, extensive reading by itself was sufficient for lexical development to occur. Then, in section 5.4, the discussion shows how in many cases reading was not enough for word knowledge to develop, and that the addition of a supplementary activity was necessary. Finally, section 5.5 summaries the key ideas presented in this chapter and transitions

into discussion addressed in the final chapter of this thesis.

5.2 Confounding variables

This section discusses four confounding variables which may have had an effect on the students' lexical development: Learner proficiency, time on task, initial learner knowledge, and target word frequency of occurrence. These variables were discussed in chapter 2 as issues found in previous research which may have affected their results.

5.2.1 Learner proficiency

The results presented in chapter 4 revealed that the students in the spoken group gained greater lexical knowledge than the ER-only and written groups. This development occurred in the mid-frequency word band, and was the greatest for those students who engaged with a word initiating a language-related episode after struggling with some knowledge aspect of the word. One reason for this greater development could be due to language proficiency differences which is one of the issues discussed in section 2.4.3 of this thesis. If groups that have different initial proficiency levels are compared, any differences observed could in part be attributed to the difference in proficiency. The current research design controlled for this phenomenon by counter-balancing the treatment groups according to their proficiency. In phase one of data collection, counter-balancing was successful and the participants in each group had similar proficiency. In phase two of collection, counter-balancing was not possible and after combining the data from phases one and two, the results revealed proficiency differences between the ER-only group and the spoken group. It is because of these proficiency differences that the results were presented separately for phase one.

5.2.2 Time on task

Chapter 2 included discussion of time on task as a confounding variable when measuring language development. Beglar, Hunt, and Kite (2012) found a positive correlation

between time on task and learning, with amount learned increasing as a function of time. The research design employed in this thesis accounted for time on task in and out of class. In class, this was done by designing the study to ensure that the three ER groups received equal amounts of treatment time. This was important because the spoken and written groups engaged in the Say-it activity for 15 minutes after finishing each graded reader while the ER-only group did not. To equalize time on task, the ER-only group read a short story from a graded reader edited collection for 15 minutes while the other two groups engaged in the Say-it activity. Time on task was accounted for out of class by tracking extracurricular reading, that is, reading the participants did in addition to the five graded readers used in the study. As shown in section 4.2.2, the participants engaged in similar amounts of extracurricular reading. In other words, it was not likely that time on task was responsible for the spoken group's greater amount of lexical development of the mid-frequency target words.

5.2.3 Initial knowledge

Learner proficiency was counterbalanced in the current study so that the participants would begin the study at similar levels. This accounts for the issue discussed in section 2.4.3 about differing initial states of knowledge (e.g., Smith, 2006). Smith (2006) concluded that adding supplementary activities to reading did not lead to greater gains in proficiency (in terms of reading comprehension). However, he did not discuss the effect that initial proficiency may have played in facilitating development. This is problematic considering that the ER-only group in his study had lower initial test scores than the group that engaged in supplementary activities. This means that the ER-only group had greater learning potential than the reading plus supplementary group since there was more for the ER-only group to learn.

As mentioned earlier, the current study's research design counterbalanced for learner proficiency in phase one. For phase two it was not possible to counterbalance as described in chapter 2 and as a result the ER-group had a significantly higher proficiency level than the written group. This highlights the importance of controlling for differ-

ences in proficiency. Removing the written group, however, revealed that the ER-only group and the spoken group both reported self-reported knowledge to similar degrees In addition, both the ER-only and spoken groups reported similar levels of semantic knowledge at the beginning of the study. In other words, these two groups were comparable in terms of self-report knowledge and semantic knowledge, meaning proficiency is an unlikely candidate for facilitating the increased development seen in spoken group's semantic knowledge development.

5.2.4 Frequency of occurrence

In the current study, the mid-frequency words showed the greatest increase in semantic knowledge compared to the high-frequency and low-frequency words. One explanation for this could be due to the number of times they occur in the stories. As mentioned in section 2.3.3, research is unclear on the number of times needed for learning to occur however estimates upwards of ten exposures have been found to be sufficient for learning to begin to accelerate (Waring & Takaki, 2003). One of the participants comments on the repetition of vocabulary and the facilitating effects it has on language learning. J, the student who provided the target word *boxing* to her group member S, mentions in her interview that there was a lot of unknown vocabulary in the fourth graded reader. Then she added "...but the vocabulary appears again and again so when I read after final stories it is better". This repetition hints at another possible explanation for the increase in knowledge of the mid-frequency words. The extent that the mid-frequency words occurred in the readers may be an indication of their importance in the stories. Put simply, the mid-frequency words may relate to central themes in the graded readers. If this were the case, skipping the word may lead to significant loss of comprehension of the story, or in other words understanding the story necessitated an understanding of the mid-frequency words. One example of this is the mid-frequency word *boxing*. Occurring only in the third graded reader *Billy Elliot*, *boxing* is central to the story's plot. The story begins with Billy being forced to box by his father, a tradition passed down over the generations. However, Billy soon becomes intrigued by ballet and begins

secretly attending ballet lessons at the gym, meanwhile telling his father he is boxing. These two reasons, a large amount of repetition in combination with an importance to understanding the story plot are two possible phenomena to explain the increased development in the mid-frequency words.

5.3 Extensive reading alone provided occasion for lexical development

The previous section discussed four confounding variables which were mitigated by the research design employed in this thesis. With the effects of these variables minimized, the discussion turns now to the ER treatments themselves. The results showed that the students in the ER-only group developed their lexical knowledge of the target words, and this is also clear from the students' interview data. One participant in the ER-only group, student Y, mentioned during his interview that he learned a new meaning for the target word *boxing*. Table 62 depicts this part of his interview. Y discussed the relationship between reading and lexical development and how he believed that reading could help to learn vocabulary. He also said that he himself learned some vocabulary from the reading and when asked which words he learned, began talking about *Billy Elliot*. During his interview, Y mentions that he initially thought *boxing* meant a container for storing items. On the pretest, Y reported meaning knowledge of the word *boxing* and provided five word associations which reinforce the interview data, suggesting *boxing* refers to a container: *inside*, *wrapped*, *gift*, *surprise*, and *music*.

In his interview, Y continues discussing how he was confused when he initially came across *boxing* in *Billy Elliot* because it did not match the meaning of *boxing* he was familiar with. After being exposed to *boxing* in *Billy Elliot*, Y became aware that *boxing* was referring to something other than a container, relating to "punching" and "hitting a face". In other words Y's interview data excerpt in Table 62 shows that he was able to learn word knowledge through reading. On the post-test, Y again reported meaning knowledge of *boxing* and provided three word associations: *inside*, *outside*, and *prison*.

These three associations had even less of a relationship to the contexts in *Billy Elliot* than Y's pretest associations, representing a decrease in semantic knowledge. This decrease stems from the fact that Y's associations are not representative of the context that *boxing* occurs in *Billy Elliot*.

This example shows how lexical development occurred through extensive reading alone without supplementary activities. Y reported during his interview that he learned a new meaning for the word *boxing*, however knowledge of this additional meaning did not manifest itself in Y's word association responses. As mentioned in section 2.3.2, a learner's lexicon can be conceptualized as a semantic network with words linked to each other by the relationships between them (Henriksen, 1999). The network develops through repeated language exposure in which words are encountered in varying contexts. That the contexts are different aids the learner in differentiating meaning relations between words, and increasing precision of meaning (Haastrup & Henriksen, 2000). In Y's case, his semantic network has increased in density since he has learned a new meaning for the word *boxing*, increasing the number of semantic links between the words in his lexicon (Milton & Fitzpatrick, 2014). Adding this new meaning to a word form which was already known represents an increase in Y's depth of knowledge. Due to the fact that a learner's semantic network is continually updating and restructuring itself (Haastrup & Henriksen, 2000), Y's newly acquired knowledge of *boxing* as a sport may be at its initial stages meaning that it still cannot be utilized during a word association test.

Person	Dialogue
T	Did you have any problems reading [Billy Elliot]? I mean was it too difficult?
Y	you know, like uh boxing? At first I read uh boxing I think it's a box, I think what is this? Haha
T	Haha, like this [points to a box in the room]?
Y	Yeah, its a box, and like two people punch...yeah but so I first see the word, of course we think the box, but it's exactly different.

Continued on next page

Person	Dialogue
T	Yeah, so you thought it was box when [you
Y	[yeah. At first I thought box, reading I read half the book, I think I understand what is boxing. No I think read just a little; after the first boxing they say punch or hit uh his face yeah. I think it's easy to understand, but no like that kind of words to remind you it's hard to...
T	What do you mean?
Y	Like uh you know like if uh they the information like uh, the teacher say I should punch, hit his face...yeah it remind me, tell me what is the meaning. But if you put a single word for me, I will choose box.
T	So the context sort of helped you to understand?
Y	Yes.

Table 62: Interview excerpt with Y talking about the different meanings of the target word *boxing*.

5.4 The language-related episodes also provided occasion for lexical development

The discussion in the previous section focused on lexical development through extensive reading, but this was not the only activity which led to lexical development. The language-related episodes (LREs) also facilitated development of semantic knowledge, as illustrated below. The first example takes the form of an episode that occurred during the third Say-it activity, after finishing *Billy Elliot*. The focus of the episode was again on the meaning of the mid-frequency target word *boxing*. The word family *box* occurred a total of 44 times in *Billy Elliot*; 28 times as *boxing*, seven times as *box*, six times as *boxed*, and once as each of *boxer*, *boxers*, and *boxes*. In the British National Corpus, *boxing* occurs in the fourth 1,000-word frequency band meaning it is considered a mid-frequency word. Table 63 shows the initial context in which *boxing*

occurred. In *Billy Elliot*, the main character Billy is playing the piano one day when his father approaches him and slams the piano cover closed; Billy's father strongly believes that Billy should be a boxer and not waste his time practicing anything else.

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. "Listen. I boxed, my dad boxed, you box". That is my dad.

Table 63: Initial context of the target word *boxing* in Billy Elliot.

During the Say-it activity, students S and J are discussing how Billy felt after his first ballet lesson. S is recalling that Billy felt wonderful because he really enjoys ballet. This recollection is the first utterance in Table 64, which depicts the language-related episode focusing on the word *boxing*. 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 Chinese how to say *boxing* in English. J responds by providing S with *boxing* and S then reformulates her initial utterance to include the word *boxing*. S stops momentarily at the beginning of her second turn, perhaps to determine the best way to include the word *boxing* into her sentence. J takes that pause as an opportunity to repeat *boxing* twice quietly. This additional assistance from J suggests that J understood S's brief pause as an attempt to remember how to say *boxing*. S continues with her sentence, mispronouncing *boxing* and then repeating it again with its correct pronunciation.

On the pretest, S reported knowing the meaning of *boxing* and provided two associations: *tidy* and *mess*. Given these two associations, it seems that S associated *boxing* with a container used for storing items, similar to student Y in the previous section 5.3. Unlike student Y in the previous section, reading *Billy Elliot* was not sufficient for S to be able to produce *boxing* during the Say-it activity. This is evident in Table 64, which took place after reading *Billy Elliot*; if reading alone were sufficient for S, it is unlikely she would have asked J for assistance. On the post-test, S reported knowing the meaning of *boxing*, the same as she reported on the pretest. According to these re-

Student	Dialogue
S	Mm, I feel life is wonderful because I achieve, I have achieved my uh dream. And uh, I don't need to, uh, (.) [speaks Chinese asking how to say boxing in English].
J	˜boxing˜
S	Uh I don't need [uh, I don't need take the box[mispronounced false start], boxing lessons. I think it is wonderful because I'm really interested in the ballet's[mispronounced] lessons.
J	[˜boxing˜, ˜boxing˜

Table 64: Transcript for LRE 67 (spoken group)

sults, S's knowledge of *boxing* did not change, however her word association responses suggest otherwise. S provided three associations on the post-test for *boxing*: *sport*, *race*, and *competition*. This set of associations is more representative of the contexts that *boxing* occurs in *Billy Elliot*. The LRE in Table 64 took place within the context of *Billy Elliot*, where the sport of *boxing* was central to the plot of the story. The results suggest that the opportunity S had to ask a group member for the meaning of *boxing* in this context assisted her in consolidating her meaning knowledge of *boxing*.

The LRE in Table 64 revealed that J provided S knowledge of the word *boxing*, which means that J had knowledge of the word at this time. Where was this knowledge acquired? On the pretest, J reported knowing the form of *boxing* and provided one word association response: *pack*. This suggests that J had some idea of the meaning of *boxing* as it relates to a storage container. During the Say-it activity, when S initiated the *boxing* LRE in Table 64, J was able to provide an appropriate form for the meaning of *boxing* as a sport. This means that for J, the reading alone was enough to learn the meaning of *boxing*. In addition, J reported knowing the meaning of *boxing* on the post-test, and provided five word association responses: *fighting*, *race*, *members*, *strong*, and *muscle*. These responses have a stronger relationship to the meaning of *boxing* as a sport than they do to the meaning of *boxing* as a container for goods.

This example LRE has shown that extensive reading can facilitate lexical development, reinforcing the discussion in section 5.3. The LRE in Table 64 has also revealed the facilitating nature that language-related episodes can have on lexical development. With the inclusion of the Say-it activity both S and J were able to develop their depth of knowledge for the word *boxing*. This increase in depth is a result of adding an additional meaning - the sport of *boxing* - to a word which was already known to some degree (*boxing* as a storage unit).

The example LRE detailed above was not the only meaning-focused LRE which facilitated development. Table 66 below illustrates another LRE which focused on the mid-frequency target word *lorry*. This word family occurred nine times in the first graded reader and did not occur in the subsequent four graded readers. Specifically, the word type *lorry* occurred eight times, and *lorries* occurred once. Table 65 presents the initial context in which *lorry* occurred in the first graded reader *Jojo's Story*, a book about a young boy's experience during wartime. In the part of the story where *lorry* first occurs, Jojo is hiding in his village which has been ransacked as a result of the war. One day, a group of soldiers travel to his village on a *lorry* when Jojo hears the sound of the *lorry* approaching.

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.

Table 65: Initial context of *lorry* in *Jojo's story* (p. 9)

During the Say-it activity which took place after the students finished reading Jojo's story, one of the spoken groups was discussing this event in the story when an LRE centered on the word *lorry* was triggered. This LRE is shown in Table 66. In the first turn in the table, T is reading a discussion prompt for S to address which includes the word *lorry*. When T finishes reading the prompt, S, unsure of the meaning of *lorry*, asks for clarification of the word. In the next turn, student J provides an explanation

of *lorry* for S which she acknowledges and then addresses the discussion prompt.

Student	Dialogue
T	Yes, uh. You are Jojo. You hear a lorry. You see soldiers in your village. What do you thinking?
S	Hmm, 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.

Table 66: Transcript for LRE 1 (spoken group)

According to her self-report pretest results, S began reading *Jojo's Story* never having seen the word *lorry*. When she was confronted with *lorry* during the Say-it activity she did not know the meaning of *lorry* and asked for help from her group. This means that for S, reading by itself was not enough for to learn that *lorry* referred to a vehicle which is used to transport people and goods. Instead, S needed additional assistance to learn this meaning, and her group member J provided her with this assistance. On the post-test S reported that she knew the meaning of the word *lorry*. In addition, S provided two associations on the post-test: *car* and *truck*. This change in knowledge from *none* on the pretest to *meaning* on the post-test represents a change in breadth of knowledge because S has added a new word to her lexicon. The LRE provided S with the opportunity to address the gap in her knowledge by clarifying the meaning of *lorry*.

S's data reveals that another aspect of knowledge may have been acquired during the language-related episode. As seen in Table 66, the episode was triggered when T used the spoken form of *lorry*. Since S had no knowledge of *lorry* before reading, she was unaware of the spoken form of the word. Reading *Jojo's Story* provided S with the written form of the word, but not the spoken form. The episode may have been S's initial exposure to the spoken form of the word which would explain her clarification. After J provides assistance, S is able to connect the spoken form to the meaning and has

increased her depth of knowledge by adding an additional dimension of knowledge to a word which she already knew to some degree. This means that the language-related episode provided the opportunity for J to be exposed to the spoken form of the word. Referring back to Table 1 which depicts 18 aspects of word knowledge, it can be seen that extensive reading provided receptive knowledge of written form, while the Say-it activity allowed for development in receptive knowledge of spoken form. In other words, the Say-it activity allowed for an additional aspect of word knowledge to develop which would not have been possible with reading only; graded readers cannot verbally provide the spoken form of a word.

S was provided a meaning of *lorry* from J her group member, and upon further inspection of J's data, an interesting picture emerges. On the pretest, J reported no knowledge of the word *lorry*. At the time of the LRE in Table 66, J was able to provide a meaning for *lorry* to S when S was in need of assistance. For J, it seems that reading by itself was sufficient for acquiring meaning knowledge of *lorry*, however her interview data suggests otherwise. Table 67 depicts an excerpt from J's interview when she is discussing *Jojo's Story*. J mentions that *Jojo's Story* was easy to read and that she learned the word *lorry* from the book. When asked what it means, she replies that it is a "big truck". This interview excerpt seems to reinforce the idea that through reading J was able to acquire meaning knowledge of *lorry*. However, at the end of the interview J returns to the word *lorry* and says that she remembered while reading that she thought that a *lorry* resembled a big car, however was unsure and so after finishing *Jojo's Story* looked up *lorry* in a dictionary. In short, a combination of J's pretest data, the LRE shown in the Table 66, and her interview data reveal that extensive reading did contribute to learning, but only partially since she additionally required the use of a dictionary to confirm her understanding of *lorry*.

The focus in this thesis is on the 60 target words which were carefully selected for the study, however it is worth remembering that these target words were just a sample of the potential learning available. The learners could focus on any word which they encountered, and often did as seen in the following LRE depicted in Table 68. This

Person	Dialogue
J	...and this book [Jojo's story] is easy, but I learn a word from this. Lorry
T	Lorry.
J	Lorry. I don't know that before.
T	What does it mean?
J	Mmm, big truck?
T	big...? [asks her to repeat]
J	truck. Lorry. l-o-r-r-y?
T	yeah. a big what?
J	big truck

Table 67: Interview excerpt for student J discussing the word *lorry*.

LRE occurred after the students read the fourth graded reader *Land of my Childhood: Stories from South Asia*. During the Say-it activity in which this LRE occurred J asks R a question involving the word *poverty*. R does not know the meaning of *poverty* and asks for assistance. J provides a Chinese translation of *poverty*, and additionally spells the word in English for R. While J is spelling the word *poverty*, R acknowledges the assistance from J. After J finishes spelling *poverty*, which is spelled without the 't', he informs R that *poverty* is more academic than the word *poor*, and then asks if R agrees. R replies with "maybe" and in J's final turn of the LRE suggests that the word *poverty* is "absolutely" more academic than the word *poor*. *Poverty* was not one of the 60 target words in the study meaning no pretest information was available. However as mentioned in section 3.7.3, the fact that a question was raised about the word indicates that R had difficulty with the word and further consolidation was necessary (Ellis et al., 2001; Swain, 2001). R answered the LRE-based question for *poverty* correctly, suggesting that the LRE in Table 68 led to the development of the meaning of the word *poverty*, assuming R did not come into contact with *poverty* in the approximately two weeks between the time of the LRE and the post-test.

So far the discussion has been centered on meaning-focused LREs, however an LRE

Student	Dialogue
J	do you think all of these story is due to the poverty in the country?
R	poverty? whats mean poverty?
J	[L1 chinese translation] poverty, p o v-
R	oh
J	e r y, yeah it's more academic than poor. do you think? do you agree?
R	maybe
J	maybe! absolutely, you must agree with me. that's all

Table 68: Language-related episode focusing on a non-target word (i.e., poverty)

did not need to have a meaning focus for semantic knowledge development to occur. One example of semantic knowledge developing can be seen in the target word *shaft*. This word occurred 17 times in the fifth graded reader *A Kiss Before Dying*, and is in the fourth 1,000-word frequency band of the British National Corpus. Table 69 shows the initial context that *shaft* occurred in.

The handsome man looked around him. Each side of the roof was about 150 feet wide. All around the edge was a brick wall, about three and a half feet high and a foot think. But the building wasn't solid. In the middle, it had a big square air shaft. Each side of the air shaft was about 30 feet wide. There was a brick wall around the air shaft too. It was the same height and thickness as the outer wall.

Table 69: Initial context of the target word *shaft* in the graded reader *A Kiss Before Dying* (p. 30)

The word *shaft* arose in an LRE during the last Say-it activity. The relevent excerpt of this LRE is shown in Table 70. During the Say-it activity, F was recalling an event in which the two main characters go to their local municipal building to submit marriage papers. Upon arrival, they find that the office is closed for lunch, and decide to go to the roof of the building and wait for the office to reopen. F begins explaining what

happened when the two characters arrived at the roof, when he mispronounces *shaft*. W interrupts F after this mistake and continues with the narrative explaining what happened in the story. In F's next turn, he repeats the last part of his first utterance, this time pronouncing *shaft* correctly. W responds by saying she does not know the meaning of shaft, to which F repeats *air* and *shaft*, mispronouncing them both. F shifts the focus of the discussion to another Say-it activity prompt which ends the LRE. This example reveals that F is unsure of the pronunciation of *shaft* since he struggles pronouncing it throughout the LRE. On the pretest, F reported form knowledge of *shaft* and did not provide any word association responses. On the post-test, F reported that he knew the meaning of *shaft* yet did not provide any word association responses for *shaft*. These results suggest that F learned additional knowledge for a word which was previously known to some degree, i.e., an increase in depth of knowledge, but as this was not a target word additional data, for example in the form of word association responses, was not available to investigate this development further.

Even though the LRE in Table 70 was triggered as a result of F struggling with the pronunciation of *shaft*, W voiced her lack of knowledge of the meaning of *shaft*. On the pretest, W reported no knowledge of *shaft*. At the time of the episode, W did not know the meaning of *shaft*, which suggest that reading by itself was not sufficient for acquiring the meaning of *shaft*. On the post-test however, W reported *form* knowledge of *shaft* and provided two word association responses: *verb* and *push*. It is unclear why W produced *verb*, however the second association *push* makes more sense; in the story, just as the characters finish their cigarettes on the roof of the municipal building one character gets *pushed* by the other into an air shaft and dies. This example suggests that W's knowledge of *shaft* developed in such a way that she was able to produce two word association responses, one of which (*push*) was directly related to the LRE in Table 70. While W's semantic knowledge of *shaft* showed signs of semantic development, she was unable to provide an accurate meaning for *shaft* on the post-test, instead providing one word, "airplane". In her interview, W mentioned that she was glad she was in the spoken group and not the written group because "if you are in the typing group, you

can see others' answer. But for the speaking group we can't know others' answer so um when I finished the first time uh task, uh I try my best to understanding because I know I need to talk something say something yes according to my real feeling". This quote from W may explain her explicitly voicing her lack of meaning knowledge for *shaft*, as a way to try her best to understand what F was saying.

Student	Dialogue
F	Highest story building. And uh and uh sit sit down to the air[mispronounce] shaft[mispronounce]
W	Uh _take a took a cigarette_ for a minute and then uh
F	sit down the uh shaft.
W	Oh, I don't know the word
F	Air[mispronounce] shaft[mispronounce] and uh I ask the last question, did you take the pills[mispronounce]?
W	Uh huh

Table 70: Transcript for LRE 408 (spoken group)

The second LRE which illustrates semantic knowledge development occurring in a non-meaning-focused LRE occurred with the target word *lorry*. As mentioned at the beginning of this section, *lorry* was a mid-frequency word which occurred nine times in the first graded reader *Jojo's Story*. The LRE centering on *lorry* which occurred during the first Say-it activity is depicted in Table 71. In the episode, R mispronounces the word *lorry* while recalling an event involving Jojo's uncle who may have owned a *lorry*. Realizing a problem in her pronunciation, she stops and corrects her pronunciation of *lorry*, and then finishes her utterance. In the next turn in the episode, C asks R what she meant, causing R to reformulate her utterance. In R's next turn she pronounces *lorry* correctly. In his next turn C agrees that someone in Jojo's family has a lorry. In C's final turn, he repeats the fact that Jojo's uncle had a *lorry*, which then concludes the episode.

This episode reveals that R struggled with the pronunciation of *lorry*. On the

Student	Dialogue
R	I think, _I think because my uncle have lorry[mispronounce], lorry[correct]._ Do you remember?
C	What?
R	her uncle or her family have one? Lorry.
C	Uh, yeah.
R	Do you know the war?
C	Yeah yeah it's just, he uncle has a lorry, and he die on the floor. haha. You see soldiers.

Table 71: Transcript for LRE 477 (spoken group)

pretest, R reported *form* knowledge for *lorry* but did not provide any associations. In addition R was not able to answer the C-test question for *lorry* correctly on the pretest. On the post-test, R reported meaning knowledge of *lorry* and provided three word association responses for *lorry*: *car*, *big*, and *heavy*. In addition, R correctly answered the C-test question for *lorry* on the post-test. These post-test results for R suggest that the extensive reading may have been sufficient for semantic knowledge development of *lorry*, but insufficient for pronunciation knowledge development. During the Say-it activity R was able to realize a gap in her knowledge of *lorry* through mispronouncing it. The language-related episode provided an opportunity to consolidate pronunciation knowledge of the word *lorry*, by allowing R to reflect on the language she produced and correct it. In addition, R was able to correctly pronunce *lorry* on the post-test.

Returning to the episode in Table 71, C questioned R's first utterance, possibly signifying that C was unfamiliar with the spoken form of *lorry*. On the pretest, C reported *meaning* knowledge of *lorry* but did not provide any word association responses. In the language-related episode, C pronounced *lorry* correctly after hearing the spoken form from R. Interestingly, on the post-test, C reported knowing the *form* of *lorry*, and did not provide any associations. This reported knowledge represents a backward learning trajectory in depth of knowledge since C has gone from some knowledge to less

knowledge. On the LRE-based question for *lorry*, C received a score of *partially correct* because he mispronounced *lorry* the first time he said it, but pronounced it correctly the second time he said it (the pronunciation-based LRE questions provided two opportunities for pronouncing a word). C's pretest data, the language-related episode, and his post-test data suggest that he acquired productive knowledge of the spoken form of *lorry* during the episode since he was able to incorporate the correct spoken form into his speech. However, since it took him two times to pronounce the word correctly on the post-test, it appears that this productive knowledge was at an early stage of learning, and highlights the nonlinear nature of lexical development. The episode also reveals that R struggled with the pronunciation of the word *lorry* since she mispronounced it initially. The episode provided an opportunity for her to notice the mispronunciation and to correct it. This is another instance where the Say-it activity provided an opportunity to develop an aspect of word knowledge which was not possible to develop when reading, i.e., the spoken form. However, in contrast to the episode previously discussed in Table 66 which showed development of *receptive* knowledge of spoken form, the episode in table 71 reveals that the Say-it activity provided both R and C with the chance to develop their *productive* knowledge of the spoken form of *lorry*. In other words, the Say-it activity enabled development of both receptive and productive knowledge of spoken form, two aspects of knowledge which are not addressed when reading since reading provides for exposure to the written form, not the spoken form.

The previous LRE revealed the complexities involved in determining the extent that lexical development occurred as a result of reading, the Say-it activity (in the form of an LRE), or a combination of both. The last pronunciation-focused LRE example will make it easier to see semantic knowledge development occurring during a pronunciation-focused LRE. The LRE below centered on the mid-frequency word *kite*. This word occurred occurred 20 times in the fourth graded reader: 16 times as *kite* and four times as *kites*. In the British National Corpus, *kite* occurs in the sixth 1,000-word frequency band. Table 72 shows the initial context of *kite*.

During the Say-it activity in which the episode occurred, W was discussing how the

If war divides your country, village from village, family from family, it is safer to stay at home and not travel anywhere. Only a bird can fly freely over war-torn countries. Or a kite - a toy made of brightly painted paper and wood. It climbs high into the sky, turning and dancing on the wind, as light as the air, as free as a bird...

Table 72: Initial context of *kite* in Land of my Childhood: Stories from South Asia (p. 30)

main character of the story made a kite for a village festival. During her explanation she mispronounces the word *kite* triggering the LRE. Her group member A assists her by providing the correct pronunciation, however W repeats the mispronounced version of *kite*. In A's next turn, she provides the correct pronunciation of *kite* for the second time. In W's third turn she explicitly acknowledges this assistance, however still does not pronounce *kite* correctly. In W's last turn, she says kite twice, the first time mispronouncing *kite* and the second time correctly pronouncing it. It is clear from this example that W is struggling with the correct pronunciation of *kite*.

On the pretest W reported form knowledge of *kite* and provided two word association responses: "food" and "name". On the post-test W reported knowing the meaning of *kite* and provided five word association responses: "sky", "fly", "story", "children", and "competition". The results suggest that for W the graded reading was sufficient for semantic knowledge development, yet insufficient for accurate pronunciation development since, as mentioned earlier, reading provides the written forms of words, not the spoken. The Say-it activity provided an environment for W to receive feedback on her pronunciation and attempt to correct it. After being provided with the correct pronunciation from A and attempting the correct form five times, W's efforts paid off and she was able to produce the correct spoken form of *kite*. However, on the post-test W incorrectly pronounced the word *kite*. This reveals that even explicit feedback and practice do not necessarily result in retention, which highlights the nonlinear nature of lexical development.

As seen in Table 73, A provided W with the correct pronunciation of *kite*. Interestingly, A reported no knowledge of *kite* on both the pretest and the post-test. Nonetheless, A provides the correct pronunciation which assists W. This may be explained by the fact that the discussion prompt which A's group was discussing included the word *kite*. This means that all A had to do was read the word, not necessarily understand the meaning of it.

Student	Dialogue
W	yes and I made the kite[mispronounce]?
A	Kite.
W	Kite[mispronounce?]
A	Kite
W	Yes I made the kite[mispronounce], and uh yes playing the kite[mispronounce].
F	Mm.
W	And one day I lost my kite[mispronounce] and uh uh but _I remember, I I remembered_ in the end of the story _I, __the_ person the person a person__ gave back to my kite[correctly pronounced] and I find my son finally.

Table 73: Transcript for LRE 369 (spoken group)

To summarize this section, the majority of the LREs lead to semantic knowledge development, emphasizing the facilitating effects that LREs can have on lexical development. The episodes detailed above provided opportunities for students who displayed lack of lexical knowledge to

- realize a gap in knowledge,
- ask for assistance, and
- receive assistance.

In addition, each student who triggered one of the episodes discussed above increased

their semantic knowledge of the word they struggled with in both meaning-focused and pronunciation-focused LREs. Table 74 shows these students' pretest and post-test latent semantic analysis scores for the target words discussed in the aforementioned episodes. The students are listed in the order which they were introduced. The "n.a." fields represent instances where a learner provided no associations meaning no score could be calculated.

Student	Target Word	Pretest	Post-test
J	boxing	-0.02	0.16
S	boxing	-0.02	0.10
T	lorry	n.a.	0.04
S	lorry	n.a.	0.03
J	lorry	n.a.	0.11
F	shaft	n.a.	n.a.
W	shaft	n.a.	0.30
R	lorry	n.a.	0.08
C	lorry	n.a.	n.a.
W	kite	0.01	0.26
A	kite	n.a.	n.a.

Table 74: Summary of semantic knowledge development scores for the students who struggled with an aspect of target word knowledge in an LRE.

5.5 Chapter summary

This chapter discussed how the research design controlled for confounding variables which were critically reviewed in chapter 2. The discussion then focused on the facilitating effects that reading alone had on lexical development. It then shifted to discussion of the supplementary activities and their facilitating nature. Through these supplementary activities, learners were provided opportunities to produce language,

and as a results were able to notice gaps in their knowledge, clarify these gaps in their knowledge, and pool their group's resources in order to further develop their knowledge. That the mid-frequency words showed the greatest development was the result of a combination of frequency of occurrence and a strong relationship to central ideas in the graded readers. This chapter also discussed the nature of the associational knowledge development which took place with these mid-frequency words. It was found that associational knowledge developed at various stages of word learning, and did so irrespective of the aspect of word knowledge focused on in a language-related episode. This can have important implications for language teaching and pedagogy, which will be addressed in the next chapter.

6 Conclusion

6.1 Introduction

This chapter concludes the research study presented in this thesis. Section 6.2 summarizes the main findings of the current research in light of the research questions posed in section 2.7. After summarizing these findings, limitations of the research design are discussed in section 6.3. Subsequently, section 6.4 discusses the extent that this thesis has contributed to theory and pedagogy for language learning and development. Finally, section 6.5 discusses some future research directions that this thesis has found to be promising.

6.2 Summary of the main findings

This section summarizes the main findings in this thesis by returning to the research questions presented in section 2.7. Each question will be addressed separately in light of the results revealed in chapter 4.

6.2.1 How do different approaches to ER affect L2 vocabulary development, specifically an ER-only approach and an ER-plus approach?

The research in this thesis found that both the ER-only approach and the ER-plus approach lead to lexical development. The students in the ER-only group developed their semantic knowledge of the 60 target words after reading the five graded readers. The students in the spoken and written groups also saw development in their semantic knowledge of the 60 target words, however their development was found to be greater than the ER-only group. This greater increase in semantic knowledge for the ER-plus groups was most easily seen in the target words which students explicitly addressed through language-related episodes during the Say-it activities. This incidental focus on form, i.e., explicit attention to language during a meaning-focused activity, created an environment which facilitated lexical development. This environment also allowed

for aspects of knowledge to be acquired which would not have been possible through reading. As discussed in section 5.4, the episodes allowed for development of the spoken form of a word to be learned. The Say-it activity provided an environment for meaning-focused output, and importantly it also allowed for a new dimension of vocabulary knowledge to be learned because meaning-focused output allows for a person to be exposed to the spoken form of a word and this is something not possible to be exposed to when reading. In short, the features of the Say-it activity (see section 3.6) allowed for an environment in which incidental focus on form could occur, and this type of activity adds to lexical knowledge in a way that reading does not.

The greater development seen in the spoken group compared to the ER-only group is measured in terms of the contexts in which the target words were found in the five graded readers used in the study. Up until the point when a language-related episode was triggered during the Say-it activity, learners' attention was focused on the graded reader which they recently completed. The context of this graded reader provided a reference for the target word; when a learner who struggled with one of the target words was provided a meaning of the word, that word was then reanalyzed as it related to the graded reader so that the learner could address the discussion prompt. That this contextual learning occurred was manifested as higher semantic knowledge values for the ER-plus groups compared to the ER-only group. This increase is evidence suggesting that incidental focus on form allows for language development to occur. Additionally, the focus on form occurred during a discussion of the graded readers, and these graded readers provided a context for the newly consolidated knowledge to be applied.

It should be remembered that the 60 target words explicitly tracked in this study were but a subset of the words that the participants were exposed to over the course of their respective treatments. The majority of the language-related episodes located in this thesis involved words which were not explicitly tracked. This highlights the flexibility offered by supplementary discussion activities; participants are provided occasion to address gaps in their knowledge, gaps which may not exist in other learners' lexicons. The large amount of words which were addressed in the LREs highlights the fact that

word knowledge is acquired incrementally and not everyone has the same degree of L2 vocabulary knowledge.

6.2.2 How do different types of interaction following ER affect L2 vocabulary development, specifically spoken interaction and written interaction?

This research has addressed the gap discussed in section 3.6 of this thesis, that of a paucity of research comparing different modes of communication and their effects on language development (Joe, 2006). The research presented in this thesis found that both speaking and writing versions of the Say-it activity provided opportunities for language development, yet to differing degrees. The current study found that the students in the written group had fewer opportunities for development compared to the spoken group. That they had fewer opportunities was in part due to the longer time needed to type a sentence compared to the amount of time it takes to say the same sentence orally. As a result, the written group produced almost 75% less language than the spoken group did in the time available. This complements previous research discussed in section 3.6 which found more opportunities for language learning in the spoken mode compared to the written mode (e.g., Brown, Sagers, & Laporte, 1999). The spoken group encountered numerous opportunities for explicitly focusing on aspects of their L2 which required further consolidation. During the language-related episodes, the students in the spoken group were able to pool their group's knowledge to address these areas of their L2 needing attention.

The results in this thesis also revealed that the mid-frequency target words which were focused on in a language-related episode led to the greatest development in semantic knowledge. It was suggested that one reason this increased development was seen in the mid-frequency words was the result of frequency of occurrence combined with a relevence or importance to understanding the main plot of the graded reader in which the word occurred (see section 5.2.4). In short, this thesis has shown that the spoken and written groups both had opportunity for language development during

the supplementary activities, but to differing degrees, and this development was most clearly seen in the semantic knowledge of the mid-frequency target words which were focused on in a language-related episode.

6.2.3 What additional aspects of language do learners focus on during the interactions?

The results in chapter 4 revealed that the learners in the spoken and written groups focused on a variety of grammatical phenomena during the interactions. There were 314 grammatical language-related episodes which were located during the Say-it activities. These episodes centered on numerous aspects of language, some of which included gender (e.g., *he-she*, in Table 36) and determiners (e.g., *a-the*, in Table 37). Looking at the wide variety of aspects of grammatical knowledge which the students addressed during their interactions, it becomes clear that these episodes provided an environment which facilitated language learning of multiple areas of grammatical and lexical knowledge. That these episodes were conducive to learning across multiple linguistic phenomena is reinforced by the fact that the spoken and written groups were able to correctly demonstrate their understanding of these aspects during the post-test (see Table 60). Finally, it was mentioned that untestable LREs also occurred wherein it was not possible to determine the actual item a learner was struggling with. However, this does not necessarily take away from the facilitating effects which the language-related episodes provide; as discussed in chapter 5, these language-related episodes were still composed of processes which have been found to facilitate language development.

6.3 Limitations

One limitation to be noted in the current study relates to the number of participants from which data was gathered. The research design in this thesis was quasi-experimental involving intact classes in an ESL environment. These classes consisted of at most 16 students, meaning that classes which made up the spoken and written groups were

divided to accommodate both modes. This leaves approximately eight students per class for the spoken group and approximately eight students for the written group. These are low numbers to use when making conclusions about the effectiveness of a certain intervention, especially when conducting statistical analyses with the intent to generalize the results to a larger population. The results presented in this thesis may have been more robust if data was collected from a larger amount of students. By using a larger sample size, it may be clearer to see the benefits that supplementary activities have for developing vocabulary knowledge.

Another limitation in this thesis is the lack of a direct relationship between the researcher (myself) and the students. During the course of the study, my contact with the students was limited to the pretest and post-test, when I delivered the graded readers to each of the classes, during the Say-it activities, and also during the post-test interviews. Having a more direct relationship with the students, for example by being their main class teacher, would most likely have increased the amount of rapport built with the students. This rapport would allow for a better understanding of the students' level of commitment regarding their respective treatments (Macalister, 2008). A greater degree of rapport with the students may have also allowed for further confirmation that they were engaging in the reading to acceptable levels. As the main teacher for example, it is possible to closely monitor the students' level of engagement and if it was determined that motivation levels had dropped, appropriate measures could be taken by, for example, reiterating the benefits of extensive reading to the students.

A third limitation concerns the lack of student choice in determining the graded readers. Controlling for which graded readers the students would read, without receiving their input as to what they think is interesting, goes against one of Day and Bamford's (2002) ten principles for successful extensive reading programs. It is difficult to know the extent that this limitation affected the results of the current study, however it is possible that the students may have enjoyed their time with the graded readers even more if they had the freedom to choose the books they wanted to read (Tabata-Sandom & Macalister, 2009). At the same time, Day and Bamford's (2002) principles

can be thought of as guidelines, not commandments, and when implementing an extensive reading program, caution should be taken to avoid limiting language-learning opportunities (Macalister, 2014, 2015).

A fourth limitation in this thesis concerns the validity of the self-report data collected. As discussed in Section 3.7.1, a test format which allowed for easy collection of data from a large amount of target words was desirable. The Vocabulary Knowledge Scale was approached cautiously as a possible candidate to use in the current study, but in the end was abandoned. Instead, a self-report format was designed specifically for the current study. Even though care was taken when deciding the format of the test, a self-report format by definition, has no demonstrated knowledge requirement, meaning that it was not possible to confirm the degree that the learners were reporting accurately.

The remaining two limitations both relate to Latent Semantic Analysis, the statistical method applied to the word association data for determining the strength of the relationship between the target words and the set of responses a participant provided for a target word. First, as discussed in Section 3.8.2, Latent Semantic Analysis derives its ability to determine semantic relationships between words through a process of dimension reduction. In the current study, the number of dimensions was set at 300 based on previous research (e.g., Landauer & Dumais, 1997). However, certain factors can influence the amount of dimensions which provide optimum evaluation, for example the size of the corpus used to create the semantic space. In other words, the choice to use 300 dimensions, while adopted from previous research, is still arbitrary and may have affected the results. The second limitation regarding Latent Semantic knowledge is the length of one document in the semantic space. In Section 3.8.2 it was mentioned that each document in the semantic space was one paragraph. While some research suggests that this is an appropriate length of one document, it is still an arbitrary length. For example, in some of the graded readers used in the study, long bouts of dialogue took place between the characters in the story and each utterance from a character would often have its own line on the page. While care was taken to ensure that the paragraphs

were sliced in a systematic fashion, it was still an arbitrary partition.

6.4 Contributions

6.4.1 Theoretical contributions

This thesis adds support for the practice of extensive reading as a means of developing vocabulary knowledge. The findings of the study revealed that the graded readers provided a contextualized, authentic environment that the learners were able to use to develop their semantic knowledge of target vocabulary. This finding is in line with previous research which has revealed the language learning benefits of extensive reading (e.g., Elley & Manghubai, 1981; Waring & Takaki, 2003).

This thesis also adds support for a task-based approach to extensive reading as a way of optimizing vocabulary learning. Chapter 5 discussed instances in which extensive reading by itself was insufficient for vocabulary acquisition to occur. For these words, it was the addition of the Say-it activities which led to further consolidation of knowledge. The Say-it activity provided occasion for explicit focus on troublesome words. This focus on form allowed for clarification and scaffolding by which the group pooled their resources to provide the appropriate knowledge aspect that a student was struggling with. This highlights previous research which has shown output to be facilitating for language development (e.g., Watanabe & Swain, 2007). This thesis also highlights the benefits of a task-based approach to extensive reading by revealing that the interactions allowed for form-focused attention to develop additional aspects of vocabulary knowledge that were not present in reading (e.g., pronunciation). In short, the current study has confirmed the language-learning benefits of extensive reading, and it has also revealed the optimizing nature that supplementary meaning-focused activities which promote interaction can have on word learning.

6.4.2 Methodological contributions

The main methodological contribution this thesis makes to the field of second language acquisition is through its approach to assigning word association strength to prompt-response sets. While some research has shown predicable patterns of development (e.g., Fitzpatrick, 2012), other research has found that certain features of the associations are difficult to determine, for example association strength. Meara and Wolter (2004) attempted to address the issue of association strength by incorporating a self-report option into their word association test. In the test, the participants were to first select two words which were associated in some way, and then assign a strength to the association. There was a scale for them to choose a value from one to four, with higher scores indicating greater strength and lower values indicating a weaker strength. The authors noted that this system is limited in its inherent subjectivity; Some test-takers tended to respond to most associations they selected as strong, while others did the opposite. This made the strength of the associations unpredictable, and allowed for the same association (e.g., *animal-badger*) to have different strengths depending on the test-taker. The research in the current study overcomes this issue by using a statistical technique, Latent Semantic Analysis, which was applied to the word association data. In this way, all instances of a given prompt-response set, e.g., *mirage-illusion*, get assigned the same strength. This approach circumvents the tendencies Meara and Wolter (2004) reported and provides a different approach to examining knowledge of word associations. The research design used in the current study allowed for knowledge of the 60 target words to be examined at multiple points int time, and as a result it was discovered that the semantic knowledge developed at different stages of word learning. It was also found that the semantic knowledge developed irrespective of the aspect of word knowledge focused on in a language-related episode. Taken together, this shows that the semantic knowledge as determined via word associations is updating continually as learners are exposed to language, reinforcing previous research suggesting this dynamic nature of learners' semantic networks (e.g., Henriksen, 1999).

6.4.3 Pedagogical contributions

The findings presented in chapter 4 of this thesis lend support to practitioners who believe that an integrated approach to extensive reading is more beneficial than a traditional reading-only approach. The research design employed in this thesis incorporated 15 minutes of in-class reading per day, and coupled with the addition of the Say-it activity, it is possible to incorporate an integrated approach to extensive reading in a short amount of class time. The greater degree of development found in the spoken group compared to the ER-only group reinforces the benefits that an integrated approach can have on language learning. Similarly, as discussed in Chapter 4, a large amount of words were the focus of language-related episodes in addition to the 60 target words tracked over the course of the study. There was no explicit design implementation which created this environment; it is a feature of meaning-focused activities which gave the learners freedom to decide what aspects of language they needed to focus on. To that extent, the findings in this thesis suggest that teachers can be confident their students will be given numerous opportunities to address the language problems which they themselves face.

Practitioners should be aware that different modes of communication may require different task design features. In the current study, 15 minutes was found to be inadequate for the written group to produce the degree of language output that the spoken group did in the same amount of time. For those advocates of writing, or those practitioners who teach writing courses, a longer amount of should be allocated to a written task in order for students to produce a larger amount of language, more comparable to that found in a spoken group. Similarly, as mentioned in section 6.4.1, the dimensions of knowledge which developed during the Say-it activities included dimensions which were not present in the reading-only condition (e.g., knowledge of pronunciation). This means that practitioners should take into account the aspects of knowledge their students are struggling with when determining appropriate tasks that will allow the possibility for these aspects to be addressed.

In short, the current study highlights the importance of considering task requirements, as well as restrictions on the task (e.g., limited class time), emphasizing that practitioners make decisions which will maximize the opportunities learners have to interact with each other and collectively develop their lexis.

6.5 Future Directions

A number of new research directions would be interesting to pursue with the intent of deepening our understanding of lexical development through integrated approaches to extensive reading. The current study used Macalister's (2014) Say-it activity as a supplement to extensive reading. It was mentioned in section 3.6 that the Say-it activity prompts can be designed for a variety of functions including

- recalling information from a story,
- making inferences based off of a story, and
- drawing on personal experience.

The implementation of the Say-it activity in this thesis was such that only the first function was utilized. Future research may benefit from investigating prompts designed to accommodate the second and third functions above. The addition of these types of prompts may necessitate a wider range of vocabulary knowledge than designing the task according to one function only since the learners would be discussing experiences that did not necessarily occur in the graded readers.

The Say-it activity was the sole activity investigated in this thesis. It would be useful for future research to examine other forms of supplementary activities and their contributions to lexical development. For example, the Say-it activity provided occasion for *incidental* focus on form, occurring during communicative activities designed for general samples of language. Further research could instead implement activities which are conducive to *planned* focus on form, for example a dictogloss or information-gap activity. As mentioned in Section 2.4.1, *planned* focus on form activities are focused on eliciting the use of specific features or words. In this way, it may be possible to hone in

on specific words or aspects of word knowledge which are of interest to the researcher or practitioner. As it relates to this thesis, this focus could be on mid-frequency words because they showed the greatest development.

Regarding the usage of Latent Semantic Analysis, future research could assign strength to word association data using different corpora to create semantic spaces, corpora which are larger than the five graded readers could provide. Using a corpus of general knowledge for example, it may be possible to draw conclusions regarding the effects that extensive reading has on this general learner knowledge. One promising semantic space created from such a corpus is readily available to calculate cosine similarity scores can be found at the University of Colorado Boulder website http://lsa.colorado.edu. A related avenue for future research stems from the fact that Latent Semantic Analysis does not take into account word order. In other words, Latent Semantic Analysis carries out computations without considering grammar. Recently, scholars have begun calculating statistical methods that complement Latent Semantic Analysis which do take account of word order (e.g., Naptali, Tsuchiya, & Nakagawa, 2010). To that extent, future research could utilize these state-of-the-art methods for determining semantic similarity between sets of text.

Finally, the research in this thesis was conducted in an ESL environment, whereby the participants had opportunity for a large amount of exposure to English in their daily lives in addition to the exposure in their classrooms. Future studies could replicate the current research design in different contexts with different populations. One environment which could provide additional insight into semantic knowledge development is the EFL environment, characterized by a limited amount of exposure to the target language compared to an ESL environment. Such a replication may provide a clearer understanding of the language-learning benefits found in approaches to extensive reading which include supplementary activities since there will be less chance that extracurricular exposure affected any language development.

7 Bibliography

Albert, R. & Barabasi, A. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97.

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*(4), 383–401.

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

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*(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

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

Beglar, D., Hunt, A., & Kite, Y. (2012). The effect of pleasure reading on Japanese university EFL learners' reading rates. *Language Learning*, *62*(3), 665–703.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *Society for Industrial and Applied Mathematics Review*, *4*(0), 573–595.

Blanche, P. (1988). Self-assessment of foreign language skills: Implications for teachers and researchers. *RELC Journal*, *19*(1), 75–93.

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variations and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, *45*, 122–159.

Brantmeier, C. (2004). Statistical procedures for research on L2 reading comprehension: An examination of ANOVA and regression models. *Reading in a Foreign Language*, *16*(2), 51–69.

Brown, C., Sagers, S. L., & LaPorte, C. (1999). Incidental vocabulary acquisition from oral and written dialogue journals. *Studies in Second Language Acquisition*, *21*, 259–283.

Bruton, A. (2002). Extensive reading is reading extensively, surely? *The Language Teacher*, *26*(11), 23–25.

Bruton, A. (2009). The Vocabulary Knowledge Scale: A critical analysis. *Language Assessment Quarterly*, *6*(4), 288–297.

Bygate, M. (1996). Effects of task repetition: Appraising the developing language of learners. In J. Willis & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 136–146). London, UK: Heinemann.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23–48). New York, NY: Routledge.

Bygate, M., Skehan, P., & Swain, M. (2001). *Researching pedagogic tasks: Second language learning, teaching and testing*. Applied Linguistics and Language Study. Longman.

Clark, H. H. (1970). Word associations and linguistic theory. *New Horizons in Linguistics*, *1*, 271–286.

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

Craik, F. I. & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology*, *104*(3), 268–294.

Creswell, J. (2012). *Educational research: Planning, conducting, evaluating research*. New York, NY: Pearson.

Cronbach, L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *The Journal of Educational Research*, *36*(3), 206–217.

Crossley, S. A., Salsbury, T., McCarthy, P., & McNamara, D. S. (2008). Using latent semantic analysis to explore second language lexical development. *American Association of Artificial Intelligence*, *21*, 136–141.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English*, *42*, 895–901.

Day, R. R. & Bamford, J. (1998). *Extensive reading in the second language classroom* (1st edition). Cambridge University Press.

Day, R. & Bamford, J. (2002). Top ten principles for teaching extensive reading. *Reading in a Foreign Language*, *14*(2), 136–141.

de la Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary. *Studies in Second Language Acquisition*, *24*, 81–112.

de Bot, K., Paribakht, T. S., & Wesche, M. B. (1997). Toward a lexical processing model for the study of second language vocabulary acquisition: Evidence from ESL reading. *Studies in Second Language Acquisition*, *19*, 309–329.

Deerwester, S., Dumais, S. T., Fumas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Dobao, A. F. (2014). Attention to form in collaborative writing tasks: Comparing pair and small group interaction. *The Canadian Modern Language Review*, *70*(2), 158–187.

Dóczi, B. & Kormos, J. (2015). *Longitudinal developments in vocabulary knowledge and lexical organization*. New York, NY: Oxford University Press.

Donato, R. (1994). Collective scaffolding in second language learning. In J. P. Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (pp. 33–56). Norwood, NJ: Albex.

Elgort, I., Candry, S., Eyckmans, J., Boutorwick, T. J., & Brysbaert, M. (2016). Contextual word learning with form-focused and meaning-focused elaboration. *Applied Linguistics*. doi:10.1093/applin/amw029

Elley, W. B. & Mangubhai, F. (1981). *The impact of a book flood in Fiji primary schools*. Wellington, NZ: New Zealand Council for Educational Research.

Ellis, N. & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning*, *43*(4), 559–617.

Ellis, R. (1994). *The study of second language acquisition*. Oxford Applied Linguistics. Oxford University Press.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford Applied Linguistics. Oxford University Press.

Ervin, S. M. (1961). Changes with age on the verbal determinants of word association. *The American Journal of Psychology*, *74*, 361–372.

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

Fitzpatrick, T. (2007). Word association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics*, *17*(3), 319–331.

Fitzpatrick, T. (2012). Tracking the changes: Vocabulary acquisition in the study abroad context. *The language learning journal*, *40*(1), 81–98.

Fitzpatrick, T. & Izura, C. (2011). Word association in L1 and L2. *Studies in Second Language Acquisition*, *33*, 373–398.

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.

Foss, P. (2010). *Vocabulary use and development in a corpus of Japanese learner blogs* (Doctoral dissertation, Victoria University of Wellington, Wellington, NZ).

Foster, P. & Ohta, A. S. (2005). Negotiation for meaning and peer assistance in second language classrooms. *Applied Linguistics*, *26*(3), 402–430.

Gass, S. (1988). Integrating research areas: A framework for second language studies. *Applied Linguistics*, *9*, 198–217.

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*(4), 395–427.

Graesser, A., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). Autotutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*, *36*, 180–193.

Green, C. (2005). Integrating extensive reading in the task-based curriculum. *ELT Journal*, *59*(4), 306–311.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*, *69*(4), 626–653. doi:10.1080/17470218.2015.1038280

Haastrup, K. & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics*, *10*(2), 221–240.

Hafiz, F. M. & Tudor, I. (1989). Extensive reading and the development of language skills. *ELT Journal*, *43*(1), 4–13.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 117–127). Cambridge: Cambridge University Press.

Hatch, E. (1978). *Second language acquisition*. Rowley, MA: Newbury House Press.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*, 303–317.

Henriksen, B. (2008). Declarative lexical knowledge. In *Vocabulary and writing in a first and second language: Processes and development* (pp. 22–66). London, UK: Palgrave Macmillan. doi:10.1057/9780230593404_2

Higginbotham, G. M. (2010). Individual learner profiles from word association tests: The effect of word frequency. *System*, *38*(3), 379–390.

Horst, M. (2000). *Text encounters of the frequent kind: Learning L2 vocabulary through reading* (Doctoral dissertation, University of Wales, Swansea).

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *The Canadian Modern Language Review*, *61*(3), 355–382.

Horst, M. & Cobb, P., T. Meara. (1998). Beyond a Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*(2), 207–223.

Horst, M., Cobb, T., & Nicolae, I. (2005). Expanding academic vocabulary with an interactive on-line database. *Language Learning and Technology*, *9*, 90–110.

Horst, M. & Collins, L. (2006). From faible to strong: How does their vocabulary grow? *The Canadian Modern Language Review*, *63*(1), 83–106.

Hosmer, D. W. & Lemeshow, S. (2000). Special topics. In *Applied logistic regression* (pp. 260–351). John Wiley Sons, Inc. doi:10.1002/0471722146.ch8

Hu & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430.

Huckin, T. & Coady, J. (1999). Incidental vocabulary acquisition in a second language: A review. *Studies in Second Language Acquisition*, *21*, 181–193.

Hulstijn, J. H. (2013). Incidental learning in second language acquisition. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 2632–2640). Chichester: Wiley-Blackwell. doi:10.1002/9781405198431.wbeal0530

Hunt, A. & Beglar, D. (2005). A framework for developing EFL reading vocabulary. *Reading in a Foreign Language*, *17*(1), 23–59.

Ishii, T. & Schmitt, N. (2009). Developing an integrated diagnostic test of vocabulary size and depth. *RELC Journal*, *40*(1), 5–22.

Janopoulos, M. (1986). The relationship of pleasure reading and second language writing proficiency. *TESOL Quarterly*, *20*(4), 763–768.

Jessup, E. & Martin, J. (2001). Taking a new look at the latent semantic analysis approach to information retrieval. *Computational information retrieval*, 121–144.

Joe, A. (1995). Text-based tasks and incidental vocabulary learning. *Second Language Research*, *11*(2), 149–158.

Joe, A. (1998). What effect do text-based tasks promoting generation have on incidental vocabulary acquisition. *Applied Linguistics*, *19*(3), 357–377.

Joe, A. G. (2006). *The nature of encounters with vocabulary and long-term vocabulary acquisition* (Doctoral dissertation, Victoria University of Wellington, Wellington, NZ).

Jung, C. G. (1910). The association method. *The American Journal of Psychology*, *31*, 219–269.

Kantrowitz, M., Mohit, B., & Mittal, V. (Eds.). (2000). Stemming and its effects on TFIDF ranking, Athens, Greece, *22*. Conference on research and development in information retrieval.

Kelly, L. G. (1969). *25 centuries of language teaching: An inquiry into the science, art, and development of language teaching methodology, 500 bc - 1969*. Rowly, MA: Newbury House Press.

Kent, G. H. & Rosanoff, A. J. (1910). A study of association in insanity. *The American Journal of Insanity*, *68*(1), 37–390.

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

Kowal, M. & Swain, M. (1994). Using collaborative language production tasks to promote students' language awareness. *Language Awareness*, *3*(2), 73–93.

Krashen, S. (1998). Comprehensible output? *System*, *26*, 175–182.

Krashen, S. D. & Terrell, T. D. (1983). *The natural approach: Language acquisition in the classroom*. San Francisco, CA: The Alemany Press.

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

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*(2), 211–240.

Landauer, T. K., McNamara, D., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahawah, NJ: Lawrence Erlbaum Associates.

Laufer, B. & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436.

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

Laufer, B. & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51.

Laufer, B. & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30.

Laufer, B. & Yano, Y. (2001). Understanding unfamiliar words in a text: Do L2 learners understand how much they don't understand? *Reading in a Foreign Language*, *13*(2), 549–566.

Lee, J. F. (1995). Using task-based activities to restructure class discussions. *Foreign Language Annals*, *28*(3), 437–446.

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

Leroy, G. & Kauchak, D. (2014). The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association*, *21*(1), 169–172.

Lifchitz, A., Jhean-Larose, S., & Denhiere, G. (2009). Effect of tuned parameters on a LSA multiplce choice questions model. *Behavior Research Methods, Psychonomic Society, Inc. 41*(4), 1201–1209.

Lizza, M. & Sartoretto, F. (2001). A comparative analysis of LSI strategies. *Computational information retrieval*, 171–181.

Llach, G. (2009). Exploring the increase of receptive vocabulary knowledge in the foreign language: A longitudinal study. *International Journal of English Studies*, *9*(1), 113–133.

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

Long, M. (1989). Task, group and task-group interactions. *University of Hawaii Working papers*, *8*(1), 1–26.

Long, M. H. (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics*, *4*(2), 126–141.

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.

Luo, D. & Koolaard, J. (2014). Predictmeans: Calculate predicted means for linear models. *R package version 0.99. http://CRAN. R-project. org/package= predictmeans*.

Lyster, R. & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition*, *19*, 37–66.

Macalister, J. (2008). Integrating extensive reading into an English for academic purposes program. *The Reading Matrix*, *8*(1), 23–34.

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

Mason, B. (2004). The effect of adding supplementary writing to an extensive reading program. *International Journal of Foreign Language Teaching*, *1*(1), 2–16.

McKeown, M. G. & Curtis, M. E. (1987). *The nature of vocabulary acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McNeill, D. (1970). Language before symbols: Very early children's grammar. *Interchange*, *1*(3), 127–133.

Meara, P. (1996). The dimensions of lexical competence. Swansea: Lognostics.

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

Meara, P. & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, *28*, 19–30.

Meara, P. & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society* (pp. 80–87). London, UK: Centre for Information on Language Teaching and Research.

Meara, P. & Wolter, B. (2004). V_links: Beyond vocabulary depth. *Angles on the English Speaking World*, *4*, 85–96.

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

Milton, J. & Fitzpatrick, T. (2014). *Dimensions of vocabulary knowledge*. New York, NY: Palgrave Macmillan.

Murphy, G. L. (2004). *The big book of concepts*. Cambridge: MIT Press.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, *20*(2), 233–253.

Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*, 133–142.

Nakanishi, T. (2015). A meta-analysis of extensive reading research. *TESOL Quarterly*, *49*(1), 6–37.

Naptali, W., Tsuchiya, M., & Nakagawa, S. (2010). Topic dependent class based language model evaluation on automatic speech recognition. *Spoken Language Technology*, *8*(1), 395–400.

Nation, I. S. P. (1980). Activities for the language laboratory. *Guidelines*, *4*, 41–46.

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

Nation, I. S. P. (1984). Understanding paragraphs. *Language learning and communication*, *3*(1), 61–68.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82.

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

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

Nation, P. & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Nation, P. & Crabbe, D. (1991). A survival language learning syllabus for foreign travel. *System*, *19*(3), 191–201.

Nemati, A. (2010). Proficiency and size of receptive vocabulary: Comparing EFL and ESL environments. *International Journal of Education Research and Technology*, *1*(1), 46–53.

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

Nguyen, B. T. (2013). *Tasks in action in Vietnamese EFL high school classrooms: The role of rehearsal and performance in teaching and learning through oral tasks* (Doctoral dissertation, Victoria University of Wellington, Wellington, NZ).

Nicholas, H., Lightbown, P. M., & Spada, N. (2001). Recasts as feedback to language learners. *Language Learning*, *51*(4), 719–758.

Palermo, D. S. (1971). Characteristics of word association responses obtained from children in grades one through four. *Developmental Psychology*, *5*(1), 118–123.

Paribakht, T. S. & Wesche, M. (1996). Enhancing vocabulary acquisition through reading: A hierarchy of text-related exercise types. *The Canadian Modern Language Review*, *52*(2), 155–178.

Paribakht, T. S. & Wesche, M. (1997). Vocabulary enhancement activities and reading form meaning in second language vocabulary development. In J. Coady & T.

Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 174–200). New York, NY: Cambridge University Press.

Pica, T. (1996). The essential role of negotiation in the communicative classroom. *JALT Journal*, *18*(2), 241–268.

Pica, T., Holliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition*, *11*(1), 63–90.

Pigada, M. & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, *18*(1), 1–28.

Politzer, R. B. (1978). Paradigmatic and syntagmatic associations of first-year French students. In V. Honsa & M. J. Hardman-deBautista (Eds.), *Papers on linguistics and child language: Ruth Hirsch Weir memorial volume* (3rd edition, Chap. 2, Vol. 1, pp. 80–87). Mouton: The Hague.

Randall, M. (1980). Word association behaviour in learners of English as a foreign language. *Polyglot*, *2*(2), 1–26.

Read, J. (1988). Measuring the vocabulary knowledge of second langauge learners. *RELC Journal*, *19*(2), 12–25. doi:10.1177/003368828801900202

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

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

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* (Chap. 11, pp. 209–227). Amsterdam: John Benjamins.

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

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

Richards, J. C. & Rodgers, T. S. (2001). *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.

Riegel, K. F. & Zivian, I. W. M. (1972). A study of inter- and intralingual associations in English and German. *Language Learning*, *22*(1), 51–63. doi:10.1111/j.1467-1770.1972.tb00073.x

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, *21*, 589–619.

Russell, D., Peplau, L. A., & Ferguson, M. L. (1978). Developing a measure of loneliness. *Journal of Personality Assessment*, *42*, 290–294.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*(1), 33–54.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, *48*(2), 281–317.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43.

Schmitt, N. & Meara, P. (1997). Researching vocabulary through a knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *20*, 17–36.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing*, *18*(1), 55–88.

Shen, Z. (2008). The roles of depth and breadth of vocabulary knowledge in EFL reading performance. *English Language Teaching*, *4*(12), 135–137.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*(1), 38–62.

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

Skehan, P. & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*(1), 93–120.

Smith, K. (2006). A comparison of pure extensive reading with intensive reading and extensive reading with supplementary activities. *The International Journal of Foreign Language Teaching*, *2*(2), 12–15.

Sommer, R. (1901). *Diagnostik der Geisteskrankheiten*. Berlin, Germany: Urban und Schwarzenberg.

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.

Swain, M. (1995). Three functions of output in second language learning. Oxford, UK: Oxford University Press.

Swain, M. & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, *16*(3), 371–391.

Swain, M. & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal*, *82*(3), 320–337.

Swenson, E. & West, M. (1934). *On the counting of new words in textbooks for teaching foreign languages*. Toronto, CA: University of Toronto.

Tabata-Sandom, M. & Macalister, J. (2009). That 'Eureka feeling': A case study of extensive reading in Japanese. *New Zealand Studies in Applied Linguistics*, *15*(2), 41–60.

Tanaka, H. & Stapleton, P. (2007). Increasing reading input in Japanese high school EFL classrooms: An empirical study exploring the efficacy of extensive reading. *The Reading Matrix*, *7*(1), 115–131.

Tudor, I. & Hafiz, F. (1989). Extensive reading as a means of input to L2 learning. *Journal of Research in Reading*, *12*(2), 164–178.

Waring, R. (2006). Why extensive reading should be an indispensible part of all language programs. *The Language Teacher*, *30*(7), 44–47.

Waring, R. & Nation, P. (2004). Second language reading and incidental vocabulary learning. *Angles on the English-Speaking World*, *4*, 11–23.

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

Watanabe, Y. & Swain, M. (2007). Effects of proficiency differences and patterns of pair interaction on second language learning: Collaborative dialogue between adult ESL learners. *Language Teaching Research*, *11*(2), 121–142.

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

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*, 79–95.

Webb, S. A. (2002). *Investigating the effects of learning tasks on vocabulary knowledge* (Doctoral dissertation, Victoria University of Wellington, Wellington, NZ).

Widdowson, H. (1979). New starts and different kinds of failure. In H. Widdowson (Ed.), *Explorations in applied linguistics 2* (pp. 54–67). Oxford, UK: Oxford University Press.

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

Wilkins, D. A. (1976). Approaches to language syllabus design. Oxford, UK: Oxford University Press.

Wilks, C. & Meara, P. (2002). Untangling word webs: Graph theory and the notion of density in second language word association networks. *Second Language Research*, *18*(4), 303–324.

Wilks, C. & Meara, P. (2007). Implementing graph theory approaches to the exploration of density and structure in L1 and L2 word association networks. In H. Daller, J.

Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 167–181). New York, NY: Oxford University Press.

Williams, J. (2001). Learner-generated attention to form. *Language Learning*, *51*, 303–346.

Willis, D. & Willis, J. (2007). *Doing task-based teaching*. Oxford, UK: Oxford University Press.

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.

Yamashita, J. (2008). Extensive reading and development of different aspects of L2 proficiency. *System*, *36*, 661–672.

Zareva, A. & Wolter, B. (2012). The 'promise' of three methods of word association analysis to L2 lexical research. *Second Language Research*, *28*(1), 41–67.

Zhao, S. Y. & Bitchener, J. (2007). Incidental focus on form in teacher-learner and learner-learner interactions. *System*, *35*, 431–447.

Zortea, M., Menegola, B., Villavicencio, A., & De Salles, J. F. (2014). Graph analysis of semantic word association among children, adults, and the elderly. *Psicologia Reflexão e Crítica*, *27*(1), 90–99.

8 Appendices

8.1 Appendix A: Say-it activity discussion prompts

The following sections list the discussion prompts used in each of the Say-it activities. The sections occur in the order that the graded readers were read.

8.1.1 Jojo's Story

1. You are Jojo. Describe your thoughts when you are alone and hiding in the village.

2. You are Chris. Describe your first meeting with Jojo.
3. You are Jojo. Talk about your understanding of the war and the fighting.
4. You are Duck. Talk about Jojo. What did you think about him at first? What did you think about him when you found your gun was stolen?
5. You are Jojo. You hear a lorry. You see soldiers in your village. What are You thinking?
6. You are Doctor Nicky. Talk about the Children's House and your work there.
7. You are Chris. Talk about your home. Why can't you take Jojo to your home?
8. You are Jojo. Why do you leave the Children's House?
9. You are Chris. Talk about your job.

8.1.2 Dead Cold

1. You are Flick. Talk about the first time you worked with your ex-partner, Scott.
2. You are Flick. Talk about what happened at Alpine ski resort when you went skiing.
3. You are Clark Johnson, governor of Colorado. Talk about why you wanted Janine dead.
4. You are Teresa, Janine's friend. Talk about the day you found Janine in the pool.
5. You are Susan the movie star. What did you talk about with Flick at your house?
6. You are Flick. You have just arrived at Alpine Resorts. Talk about why you lied and said you were a tourist.
7. You are Eddie Lang. Talk about your job. Why you were at the Alpine ski resort?
8. You are Flick. Talk about how you felt after discovering who Janine's killer is.
9. You are Janine. Talk about what you wrote in your notebook.

8.1.3 Billy Elliot

1. You are Billy's father, Jackie. Why did you sell your wedding ring? Why did you return to work?

2. You are Billy's father. What problems do you have in your life?
3. You are Billy. Talk about what you do after school at the gym. How does your family feel about this?
4. You are Billy. You have just finished your first ballet lesson. Talk about how you feel.
5. You are Billy. Talk about why you are going to London. How does your family feel about this?
6. You are Billy's brother, Tony. Talk about why you went to jail.
7. You are Billy's brother, Tony. Talk about how you feel about the strike.
8. You are Billy. Talk about what happened when Mrs. Wilksinson told your family that you were taking ballet lessons.
9. You are Billy. How was your relationship with your family before you danced? How is it now?

8.1.4 Land of my Childhood: Stories from South Asia

1. You are the girl in the first story. Talk about your life and your relationship with Vijay.
2. You are the inspector in the second story. Talk about your job and the school you visited.
3. You are the daughter in the third story *The Intelligence of Wild Things*. Talk about your relationship with your brother, Tarun, and how it has changed over the years.
4. You are Ahmed Rasool, the kite maker. Talk about the kite festival and what happened with your son.
5. You are Lakshman, the boy in *The Stepmother*. Talk about how you feel about your stepmother and why you broke the doors.
6. You are the main character in *The Night Train at Deoli*. Talk about your experience at the train station with the girl selling baskets.

7. You are Rakesh's father in *A Devoted Son*. Talk about your son's life.
8. You are Missiya, the wild one. Talk about your life in the village.
9. You are Raju in the final story. Talk about your experience in the hospital.

8.1.5 A Kiss Before Dying

1. You are Dorothy's boyfriend. Talk about your relationship with Dorothy.
2. You are Bud Corliss. Talk about what happened at Leo Kingship's factory.
3. You are Ellen. Talk about why you think Dorothy was killed.
4. You are Bud Corliss. Talk about your plans to kill Dorothy.
5. You are Ellen. Talk about your relationhip with Dwight Powell.
6. You are Leo Kingship. Talk about your relationship with your daughters.
7. You are Marion. Talk about how you feel about your father.
8. You are Dorothy. Talk about your future plans with the young man.
9. You are Bud Corliss. Talk about the research you did so Marion would like you.

8.2 Appendix B: Target word information

Table 75 shows information about the target words used in this thesis. Asterisks next to a word refer to those words also in the pilot study.

Word	Band	Frequency	Range	BNC	Familiarity	Concreteness	Imageability	Meaningfulness
acrobat	off	0	0	9	431	566	583	420
alsatian	low	1	1	13	0	0	0	0
arsenic*	low	6	1	9	0	0	0	0
asphalt	off	0	0	9	488	583	569	385
audition*	mid	28	1	6	479	370	395	397
babu	low	2	1	22	0	0	0	0
back	high	127	5	1	587	540	483	418
bawl	off	0	0	10	429	384	417	370
blouse*	mid	8	2	5	562	640	595	530
boxing*	mid	27	1	4	0	0	0	0
campus*	mid	13	1	4	0	0	0	0
carapace	low	5	1	12	0	0	0	0
carat	off	0	0	11	383	488	345	313
cinder	off	0	0	10	382	579	519	379
closet*	mid	7	1	6	540	599	525	415
clucking	low	1	1	11	0	0	0	0
commode	off	0	0	12	354	482	447	376
copper*	mid	15	1	4	491	547	548	350
cupboard*	mid	7	3	5	0	0	0	0
dandruff	off	0	0	13	495	546	554	385
dead	high	57	5	1	581	429	520	497
down	high	137	5	1	546	339	459	444
egad	off	0	0	22	0	0	0	0
envious	low	1	1	9	470	0	361	0
felicity	low	2	1	12	0	0	0	0
fell	high	31	5	1	546	407	431	403
flick*	mid	19	1	4	0	0	0	0
forceps	off	0	0	12	402	585	553	366
frigid	off	0	0	9	466	411	463	445
gelatin	low	4	1	12	0	0	0	0
give	high	50	5	1	595	326	383	465
gloves*	mid	18	2	4	0	0	0	0
goat*	mid	7	2	4	469	636	585	402
handsome*	mid	28	1	4	0	0	0	0
hard	high	40	5	1	595	425	460	497

hibiscus	low	1	1	13	0	0	0	0
hopper	low	1	1	9	0	0	0	0
jasmine	low	3	1	9	0	0	0	0
jeep*	mid	7	1	6	564	622	659	477
kite*	mid	16	1	6	481	592	624	408
long	high	59	5	1	579	381	471	492
lorry*	mid	8	1	8	198	420	383	236
minstrel	off	0	0	11	390	530	522	367
morgue	off	0	0	10	434	572	589	452
name	high	50	5	1	573	405	475	474
napkin	off	0	0	9	495	585	582	357
near	high	41	5	1	582	337	408	465
need	high	38	5	1	589	314	327	473
papa	mid	13	1	5	0	0	0	0
perjury	off	0	0	9	405	323	353	367
pharmacy*	mid	15	1	6	0	0	0	0
picket*	mid	8	1	6	0	0	0	0
pliable	off	0	0	12	411	377	364	335
prawns	low	4	1	9	0	0	0	0
quiet	high	36	5	1	577	389	426	451
raffle	low	4	1	9	0	0	0	0
read	high	45	5	1	568	420	499	467
rhapsody	off	0	0	13	321	414	409	329
right	high	53	5	1	599	361	372	413
round	high	35	5	1	563	438	559	489
rupees	low	5	1	10	0	0	0	0
sahib	low	5	1	15	0	0	0	0
sans	low	1	1	10	0	0	0	0
saris	low	2	1	11	0	0	0	0
shaft*	mid	17	1	4	0	0	0	0
stop	high	53	5	1	563	308	452	485
stupid	high	34	5	1	550	351	381	487
sward	off	0	0	15	155	364	207	0
tell	high	147	5	1	596	306	350	465
thing	high	37	5	1	587	350	358	479
trailer*	mid	9	1	5	528	597	587	363
tripod	off	0	0	9	363	577	574	338
turban	low	2	1	10	0	0	0	0
vacation	mid	11	1	5	495	414	559	0
vats	low	1	1	12	0	0	0	0
vestment	off	0	0	12	318	404	365	318

well	high	68	5	1	550	467	522	418
white	high	43	5	1	590	472	566	464
whoosh	low	1	1	10	0	0	0	0
zenith	off	0	0	10	416	430	449	317

Table 75: Target words used in the main study (* words also in pilot study)

8.3 Appendix C: Reading log example

Name_____ **Date_____**

Reading Log

Write down the reading you do outside of class

Date	Title	Pages
July 18th	Alice in Wonderland	10

8.4 Appendix D: Participant consent form

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

Reading and second language learning

Consent form

Please sign and date this form to indicate you are willing to participate in this project.

I have read the information sheet for this research project and the details of the research project have been explained to me. I have also had the opportunity to ask questions about the research project.

I understand the way in which results from this project may be used.

I understand that participation in this study is voluntary.

I give my consent to be audio-recorded during this project.

I understand that I can withdraw from this project for any reason up to two weeks after the date that I sign this consent form.

I understand that any data I provide will be confidential, and that no names will be used in any report of the project.

Name: .....

Signature: .....

Date: .....

8.5 Appendix E: Participant information sheet

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

Reading and second language learning

Information sheet

I am a Ph.D. student in the School of Linguistics and Applied Language Studies, and I will be conducting research on reading and second language learning. The Victoria University of Wellington Human Ethics Committee has granted ethical approval for this research.

This research project consists of the following. First, at the beginning of the project I will give you a short test to determine your current knowledge. Then, during the project, you will read about 5 graded readers in class. After finishing each graded reader, some of you will be put into small groups and will have a discussion based on the graded reader either through speaking, through typing on the computer, or both. I will audio-record the oral discussions and collect the written ones electronically. At the end of the project, I will give you another test to see if your knowledge has changed. In addition, you will be asked to keep a reading log of all the reading you do during the research project.

Your participation in this project is voluntary. If you decide you want to withdraw from this research project after you have given consent, that is fine. You may withdraw up to **two weeks** after you have signed the consent form. If you choose to withdraw, your records will be removed from the data.

No names will be used in this study, and all data will be presented in a confidential manner. Findings may be presented at seminars, conferences or in publications. Also, all data will be destroyed after three years. During the research project all data will be stored securely and only myself and my supervisor will have access to it. In addition, choosing to participate (or choosing not to participate) will in no way affect your assessment in class. If you have any questions about this research project, please feel free to contact me or my supervisor, Dr. John Macalister. Our contact details are below.

Sincerely,

TJ Boutorwick
tj.boutorwick@vuw.ac.nz
Tel: 463-5233 ext.8029
office: VZ410

Dr. John Macalister
John.Macalister@vuw.ac.nz
Tel: 563-5609
office: VZ211