



Project  
**MUSE**<sup>®</sup>

*Today's Research. Tomorrow's Inspiration.*

---

## **Learning L2 Vocabulary through Extensive Reading: A Measurement Study**

Horst, Marlise.

The Canadian Modern Language Review / La revue canadienne des langues vivantes, Volume 61, Number 3, March / mars 2005, pp. 355-382 (Article)

Published by University of Toronto Press  
DOI: 10.1353/cml.2005.0018



➔ For additional information about this article  
<http://muse.jhu.edu/journals/cml/summary/v061/61.3horst.html>

# Learning L2 Vocabulary through Extensive Reading: A Measurement Study

---

Marlise Horst

**Abstract:** Many language courses now offer access to simplified materials graded at various levels of proficiency so that learners can read at length in their new language. An assumed benefit is the development of large and rapidly accessed second language (L2) lexicons. Studies of such extensive reading (ER) programs indicate general language gains, but few examine vocabulary growth; none identify the words available for learning in an entire ER program or measure the extent to which participants learn them. This article describes a way of tackling this measurement challenge using electronic scanning, lexical frequency profiling, and individualized checklist testing. The method was pilot tested in an ER program where 21 ESL learners freely chose books that interested them. The innovative methodology proved to be feasible to implement and effective in assessing word knowledge gains. Growth rates were higher than those found in earlier studies. Research applications of the flexible corpus-based approach are discussed.

**Résumé :** Dans de nombreux cours de langue, on met dorénavant à la disposition des apprenants des textes simplifiés, adaptés à leur niveau de compétence linguistique, pour leur permettre ainsi de lire de plus longs textes dans la nouvelle langue. Les études qui ont évalué ce type de programme de lecture extensive (LE) ont révélé des gains en compétence générale, mais très peu de ces études ont examiné les gains au niveau du vocabulaire, et aucune n'a identifié les mots disponibles pour l'apprentissage dans un programme complet de LE ou évalué dans quelle mesure ces mots ont été appris par les participants. Le présent article décrit une méthode pour relever le défi que présente ce type d'étude : à l'aide d'un numériseur, de profils de fréquence lexicale, et de tests individualisés constitués de réponses à cocher. Un essai pilote a été effectué dans un programme de LE où 21 apprenants d'anglais langue seconde avaient libre choix de livres qui les intéressaient. La méthodologie s'est révélée applicable et efficace pour évaluer les gains en connaissances lexicales. Selon cette méthode, des gains plus importants ont été relevés dans la présente étude que dans les études antérieures. Les réalisations de recherche de l'approche-recueil flexible sont abordées.

## Introduction

Teachers and researchers are increasingly aware of the importance of reading in the development of L2 proficiency and the role reading plays in expanding vocabulary knowledge. Reading is important because comparisons of large corpora consistently show that written texts are richer in lexis than spoken ones. According to analyses reported by Nation (2001), a typical English conversation proves to contain very few words that are likely to be unfamiliar to language learners beyond the most basic stages. There is also evidence that the spoken language learners are exposed to in communicative ESL classrooms contains very few words not found on lists of the most frequent English word families (Meara, Lightbown, & Halter, 1997). This suggests that learners hoping to move beyond basic oral communication skills (e.g., to read academic texts or to function professionally in English) must read written text in order to expand their lexicons.

Of course, there are other, less time-consuming ways to acquire new L2 lexis other than picking up word knowledge incidentally through exposure to print. Classroom tasks and instructional materials that actively involve learners in elaborating on new word knowledge have been shown to be particularly effective. Indeed, such activities consistently result in greater numbers of words learned and higher retention rates than simply encountering new words while reading for general comprehension (for an overview see Hulstijn & Laufer, 2001). But for learners who aspire to achieve high levels of L2 proficiency, there are more words to know than can be readily treated in instructional activities during the time frame of a typical language course. This is an argument for promoting L2 reading fluency so that learners can acquire new word knowledge on their own through independent reading. Fluent readers who can easily process large amounts of written text will encounter new items that they are unlikely to meet through exposure to spoken language (e.g., by watching television). They can be expected to infer the meanings of some of these unfamiliar words, retain the new form-meaning associations, and build an ever larger mental lexicon – much as native speakers do over a lifetime of reading in their L1 (Nagy, Herman, & Anderson, 1985). But the short passages available in typical integrated-skills ESL textbooks hardly amount to a program of reading on the scale required to develop L2 reading fluency. Grabe and Stoller (2002) make this point in a recent book: ‘Most L2 readers are simply not exposed to enough L2 print (through reading) to develop fluent processing. ... Nor do they have enough exposure to build a large recognition vocabulary’ (p. 47).

These concerns have prompted many ESL programs to include an extensive reading component as a supplement to regular language classes. Extensive reading (ER) is defined as reading that exposes learners to 'large quantities of material within their linguistic competence' (Grabe & Stoller, 2002, p. 259). Such reading material is available from publishers like Cambridge, Penguin, Oxford, and others that have produced hundreds of simplified book-length fiction and non-fiction texts graded at varying levels of English vocabulary and structure; these texts are referred to as simplified or graded readers.<sup>1</sup> According to Day and Bamford (1998), the main goal of ER is developing reading fluency; that is, rapid access to known L2 words through encountering them repeatedly. The acquisition of new L2 vocabulary along the way is seen as an important additional benefit. Krashen (1989, 1993) takes a stronger position, advocating 'free voluntary reading' as the main route for acquiring new vocabulary.

To what extent do participants in ER programs experience the vocabulary growth that is expected to go hand in hand with increased amounts of reading? The answer is of interest to teachers and course designers contemplating the addition of an ER component to their programs. From a research perspective, the question addresses an under-explored learning context, one where participants choose freely from reading materials suited to their levels of proficiency. The study reported in this article addresses the problem of how to assess such growth in an accurate and ecologically valid manner. I begin by turning to relevant research precedents and consider two main strands of research. First, I examine what previous studies of ER reveal about L2 vocabulary growth and how experiments assess gains. I then consider studies of incidental vocabulary acquisition through reading with a view to evolving a suitable experimental method for assessing gains achieved through ER. The review is limited to investigations of adult learners.

### *Studies of extensive reading*

The benefits of ER have been widely documented in studies that range in scope from large-scale implementations across whole school districts (e.g., Elley, 1991; Lightbown, 1992) to case studies of single readers (e.g., Cho & Krashen, 1994; Parry, 1991). However, ER research tells us relatively little about the vocabulary-expanding effects of reading extensively in a second language, simply because it tends to focus on other, more general aspects of language development. This can be seen in Table 1, which summarizes published studies of ER programs for adult ESL learners. As the 'Results' column indicates, studies show that

participating in an ER program tends to be associated with improved performance on tests of reading comprehension, writing, and other integrative measures. Only two of the studies listed in Table 1 (Cho & Krashen, 1994; Robb & Susser, 1989) included vocabulary measures, and only Cho and Krashen's study of four learners tested participants on words that had actually occurred in their reading. (The Robb & Susser study appears to have used a standardized vocabulary test.)

An experimental design problem is evident in the 'Other English input' column in Table 1. In almost every study, participants were either taking other English classes, living in an English-speaking milieu, or both, so the extent to which reported language gains can be attributed to ER is unclear. ER research so far appears to document proficiency gains that may be underpinned by vocabulary growth rather than vocabulary growth per se. That is, we can assume that ER (along with other class work and/or other sources of exposure to L2 input) has familiarized the participants in these experiments with L2 words and structures in ways that are reflected in improved performance on integrative measures, but the direct effects of participation in an ER program on vocabulary size are largely unexplored.

#### *Incidental vocabulary acquisition research*

In the second strand of research – studies of incidental acquisition of new vocabulary through reading – participants are typically asked to read a text, usually with the expectation that some sort of evaluation task will follow, but they do not realize that this will be a test of words that occurred in the passage. Test formats include identifying synonyms of target words in a multiple-choice format or providing translation equivalents. These read-and-test studies clearly document gains in learners' knowledge of specific words learners have encountered in reading a text (rather than the more general language gains documented in most ER studies). In addition, steps are usually taken to ensure that the tested words are not previously known to the participants; the reading material can then be assumed to be the sole source of information about the meanings of the test items and the main explanation for any growth that is documented. This experimental control stands in marked contrast to ER studies that assess gains in contexts that offer other opportunities to learn new vocabulary. Details of a number of studies of incidental vocabulary acquisition are summarized in Table 2.

Studies such as those in Table 2 offer a useful methodological model for investigating new word learning through ER in that they focus directly on words met in reading and control for other sources of

TABLE 1

Selected studies of extensive L2 reading (adult learners of English)

Study	Context & population <sup>a</sup>	Reading materials	Mean amount read per week <sup>b</sup>	Other English input	Results: change in ...
Robb & Susser (1989)	125 EFL Japanese university	simplified for NSs	18 pages	EFL classes	• reading comp. • speed • vocabulary
Hafiz & Tudor (1989)	16 ESL Pakistani secondary	graded ESL readers	4.5 hours	UK milieu	• reading comp. • writing
Hafiz & Tudor (1990)	25 EFL Pakistani secondary	graded ESL readers	4 hours	EFL classes	• essay writing
Lai (1993)	250+ EFL Hong Kong secondary	graded ESL readers	4–5 books	English in milieu?	• reading comp. • speed • essay writing • vocabulary
Cho & Krashen (1994)	4 ESL adults	Simplified for NSs	2–6 books	US milieu	• essay writing
Tsang (1996)	48 EFL Hong Kong secondary	graded ESL readers	3 books	EFL classes	• essay writing
Mason & Krashen (1997)	20, 128, 76 EFL Japanese university (3 experiments)	graded ESL readers (+ unsimplified in exp. 2)	1–2 books in exp. 1	EFL classes in exp. 2 & 3	• cloze • reading comp. • summary writing • speed* • cloze**
Walker, C. (1997)	43 ESL adults	graded ESL readers	1 book	UK milieu	• reading comp. • oral reading
Lituañas, Jacobs, & Renandya (1999)	30 ESL Filipino secondary	various unsimplified? materials	1.5 hours	ESL classes & English in milieu	• integrated skills**
Renandya, Rajan, & Jacobs	49 EFL Vietnamese government employees	graded ESL readers	91 pages	EFL classes	• speed • cloze • reading comp.
Bell, T. (2001)	14 EFL Yemeni government employees	graded ESL readers	1.5 + (?) hours	EFL classes	

<sup>a</sup> The absence of Canadian ER studies in this table is striking. As the information in this column shows, most of the studies were conducted in overseas EFL settings.

<sup>b</sup> This column allows for rough comparisons to the recommended one graded reader per week (Day & Bamford, 1998, 2002; Nation & Wang, 1999). In cases where researchers did not report amounts by week, I have used figures reported in the articles to calculate a weekly rate as accurately as possible. For instance, in the case of Robb and Susser (1989), which reports 641 pages read per school year,

TABLE 1 (continued)

---

I have assumed a 35-week school year and calculated a weekly rate of 18 pages per week (641 pages ÷ 35 weeks = 18.31 pages). Rates reported variously in books, pages, or hours in this column reflect the absence of a uniform measure across the studies.

\* assessed using course materials

\*\* no test for significance reported

NS = native speaker of English

---

TABLE 2

Selected studies of incidental L2 vocabulary acquisition through reading

Study	No. of participants	Text type	Text length	No. of words tested	Mean no. of words learned
Ferris (1988)	51	<i>Animal Farm</i> , unsimplified novel	1 book	50	7
Pitts, White & Krashen (1989)	35	<i>Clockwork Orange</i> , unsimplified novel	1 excerpt	28	2 <sup>a</sup>
Day, Omura & Hiramatsu (1991)	200	<i>Mystery of the African Mask</i> , simplified story	1 story	17	3 <sup>a</sup>
Hulstijn (1992) exp.1	65	simplified passage	1 passage	12	1
Dupuy & Krashen (1993)	42	<i>Trois hommes et un couffin</i> , unsimplified drama	Video excerpt + 15 pages	30	7 <sup>a</sup>
Horst, Cobb & Meara (1998)	34	<i>Mayor of Casterbridge</i> , graded reader	1 book	45	5
Rott (1999)	67	simplified passage	6 paragraphs	12	6
Horst (1999)	1	<i>Lucky Luke</i> , unsimplified comic	1 book	300	85
Horst (2000)	1	<i>Der Besenbinder</i> , unsimplified novella	1 novella	300	66

---

<sup>a</sup> gain established by comparison to a control group only (no pre-test)

learning gains. However, other aspects of this methodology are problematic. First, many of the studies offered participants limited opportunities to demonstrate new word learning. Figures in the 'Words tested' and 'Words learned' columns in Table 2 indicate that testing learners on their knowledge of just two or three dozen low-frequency items that occur in a short passage tends to deliver very small, possibly unreliable mean gains, sometimes on the order of just two or three new words (see Horst, Cobb, & Meara, 1998, for a detailed discussion). Of course, there

is reason to expect incidental word learning rates to be low. Nagy, Anderson, and Herman's 1987 study of child L1 readers put the chances of retaining the meaning of a new word met in reading at one in 20. But tiny increments of growth are hardly inspiration for curriculum designers and language instructors contemplating the usefulness of implementing ER programs in their schools. Thus there is a face validity argument for testing participants on larger numbers of words (preferably hundreds rather than dozens) after they have read larger amounts of text (preferably several graded readers rather than a single passage).

Second, figures in the 'Words tested' column of Table 2 attest to the difficulty of administering lengthy word tests to large groups of participants. It is interesting to note the (roughly) inverse relationship between number of participants and number of words tested in the incidental acquisition experiments, with the largest tests (300 items) administered in studies of a single learner. The one ER experiment that assessed participants' knowledge of hundreds of words (Cho & Krashen, 1994, in Table 1) was also a case study of just four learners. Thus the research precedents point to a methodological constraint: So far, close examinations of word learning in ER settings appear to be limited to investigations of a few individuals who are willing to undergo extensive testing.

The incidental acquisition research detailed in Table 2 also points to the need for an investigation of word learning through reading in true ER contexts. Conditions for reading in these studies differed substantially from ER settings in a number of ways. For instance, as the column labelled 'Text type' shows, many of the experiments presented L2 readers with unsimplified reading material. These texts may not have been understood well enough for the participants to be able to infer the meanings of unfamiliar words they encountered.<sup>2</sup> Nation's (2001) review of studies of the relationship between known-word densities in texts and reading comprehension settles on a known-word minimum of 95%. That is, in order for L2 readers to comprehend a text adequately and infer the meanings of unfamiliar words they encounter, no more than 5% of the running words (or one word in 20) should be unfamiliar. Analyses of graded readers by Nation and Wang (1999) indicate that this 95%-known-word minimum can be met if learners in an ER program select simplified materials appropriately.

It is also clear that the learning results in the studies listed in Table 2 are based on limited amounts of reading – in some cases no more than a single passage or short story (see column labelled 'Text length'). In ER programs, learners usually read a great deal more, though much depends on learners' circumstances and abilities (see 'Mean amount



read' column in Table 1). Day and Bamford's (2002) rule of thumb for designing ER programs is to set a goal of one simplified reader per week. Research by Nation and Wang (1999) confirms the one-per-week figure to be the minimum needed to ensure that learners get the repeated encounters with new words that make them memorable. Finally, levels of motivation can be expected to be different in ER contexts as well: In contrast to the studies in Table 2 where all participants were required to read a prescribed book or passage, students in authentic ER settings choose freely from a collection of materials and read as much (or as little) as they please.

### **Implications for research design**

To summarize, the review of previous research points to a number of design criteria for valid experimental assessment of vocabulary gains achieved in a program of ER. First, word knowledge measures administered before and after an ER treatment should feature words that participants have actually met in their reading. Second, as much as possible, the experimental design should exclude a role for other sources of word learning (e.g., language classes) so that any documented gains can safely be ascribed to the ER treatment alone. Ideally, measures would also be large, sensitive, and easy to administer to a sizable group of participants; in addition, they should reflect the word learning opportunities available in large amounts of text. Finally, to ensure ecological validity, the materials used in the reading treatment should be selected by the participants themselves from a collection of materials suited to their interests and levels of proficiency.

The next section describes an experimental methodology that attempts to meet these criteria through the use of electronic scanning, lexical frequency profiling, and checklist testing techniques. The feasibility of the experimental method is then explored in a pilot study.

### **Design**

#### *Challenge 1: Identifying vocabulary that occurred in participants' reading*

Assessing the extent to which participants (a group of adult immigrant learners of English in an ER program at a community centre in Montreal) acquired new knowledge of words that appeared in the materials they read required selecting words from these materials to serve as test items on measures of word knowledge. However since no two partici-

pants read the same graded readers, this was problematic. Ordinarily in read-and-test studies, researchers pre-test participants on their knowledge of selected words that occur in a given text, ask all participants to read this text, and later administer the same word test again to look for pre-post-test differences. But in this setting there was no single text that all would read and, therefore, no one set of words that all students could be tested on before and after participating in the ER program.

The challenge of determining which words to target on pre- and post-tests in a context where each participant read different text was resolved in the following way. First, a dozen graded readers representing the full range of simplification levels in the collection were scanned electronically. The scanned files were used to create lists of word types that appeared in the texts. Items from these 12 lists were randomly selected to build a test that served as a pre-reading baseline measure of knowledge of words that typically occur in graded readers. The next step in the plan was to scan many more books – eventually the entire collection, if possible – and to create lists of the words that appeared in them. The idea was to use these lists to create a post-reading test for each participant made up of items from the texts he or she had opted to read. Results on the individualized post-tests could then be compared to participants' pre-reading baseline performance. For example, if checkout records showed that a participant had chosen to read simplified versions of *Vanity Fair*, *Treasure Island*, and *Pride and Prejudice*, it would be possible to use the lists of words from these books to create a post-test and then compare performance on this test to performance on the baseline measure. If a bank of test-target lists for *all* the books in the collection could be created, then each participant in the ER program could be tested using an individualized instrument that reflected his or her choice of reading, regardless of which titles had been chosen. This would require a great deal of scanning, but the opportunity it afforded to accommodate learner choice was crucial to the ecological validity of the experiment. Allowing readers to follow their interests and choose freely what they will read is a defining aspect of ER programs (Day & Bamford, 1998; Nation, 1997); it adds the motivational component that is likely to be lacking in reading studies such as those presented in Table 2.

### *Challenge 2: Distinguishing between ER and other sources of L2 input*

To address the problem of distinguishing vocabulary growth achieved through ER from vocabulary growth achieved in some other way (e.g.,

in ESL classes or through media exposure), my research assistants and I turned to lexical frequency profiling (LFP) software. This computer tool devised by Nation and Heatley (1996) uses word lists by West (1953) and Coxhead (2000) to classify the words of a text into four frequency categories: (a) the 1,000 most frequent word families of English; (b) the 1,001–2,000 most frequent zone; (c) the Academic Word List (AWL), a set of 570 word families that occur frequently in university texts across academic disciplines; and (d) off-list words; that is, less frequent words and proper nouns that do not occur on any of the earlier lists. The frequency profiler works with words arranged into families made up of closely related inflected and derived forms (e.g., *sharp*, *sharply*, *sharpening*, and *sharpness*); a word is classified according to the frequency of the family it belongs to. The scanned files of graded readers were analyzed in this way using VocabProfile, the on-line version of the LFP program (see Cobb n.d.).

Although participants were expected to learn words in all four frequency categories as they read, lexical items falling into the off-list category were of particular interest for this project. Though one could not claim with certainty that a learner's knowledge of off-list words like *squire* and *rum* was acquired by reading a simplified version of *Treasure Island*, it was clear that frequency profiling could be used to systematically identify low-frequency test targets that were unlikely to have been met outside of the ER context. Frequency profiling was also key to the pre-post design of the planned experimentation. The idea was to compare participants' knowledge of words in particular frequency ranges before and after participating in the ER program. For instance, if learners indicated knowledge of more words on a post-test of off-list words taken from books they had read during the six-week experimental period than on the pre-reading baseline measure of other off-list items, they would be considered to have demonstrated growth in knowledge of words in this frequency range.

### *Challenge 3: Providing ample opportunity to demonstrate growth*

As is apparent in Tables 1 and 2, reading studies that test learners on their knowledge of large numbers of words are the exception rather than the rule. But a large test is clearly desirable. If, like their L1 counterparts in the study by Nagy et al. (1987), L2 readers retain the meaning of only one new word in 20, then a test of hundreds of words is needed to give ample opportunities to demonstrate learning. Ideally, the test would also be sensitive enough to detect small gains in word knowledge, since research with L1 readers by Nagy et al. (1985) has shown that word

knowledge gained incidentally through reading is acquired in increments.

To address these test design challenges (compounded considerably by the plan to create an individualized post-test for each participant), it was decided to use a modified version of the self-report checklist technique that Horst and Meara (1999) used in a reading study of a single learner. The test simply requires participants to register levels of confidence in their ability to recognize the meanings of listed items by circling one of three rating options: YES (I know what this word means); NS (I have an idea about the meaning of this word, but I am not sure); or NO (I do not know what this word means). The NS option would allow learners to register partial knowledge of words. It was also intended to encourage honesty, since it was assumed that students would over- or under-estimate their word knowledge less if they were not forced to simply choose between either YES or NO. The test is easy to construct, administer, and score, and it allows for quick assessment of a large number of items. Horst's (2000) investigations of adult learners have shown self-ratings to be a reasonably reliable indicator of word knowledge; participants in these studies were able to provide accurate translation equivalents of about 80% of the words they rated 'known.' Table 3 shows a portion of the instrument that was eventually developed by randomly sampling off-list words from 12 readers, including some from *Treasure Island* (items 64–66).

To determine whether this ambitious plan for measuring vocabulary growth through ER was feasible, we explored three methodological aspects in the first experiment: scanning on a large scale, finding unusual words in the simplified materials, and implementing the innovative testing. The research questions were as follows:

- 1 How feasible is scanning whole texts? Is it possible to scan enough to create post-tests that reflect the individual reading experiences of a group of participants each choosing freely from a large collection of titles?
- 2 To what extent do off-list items (and words of other frequencies) occur in graded readers? Do simplified materials, which by definition contain large proportions of frequent words, contain enough infrequent items to be useful in making an experimental case for word learning through ER?
- 3 Is making individualized self-report checklist tests that systematically sample participants' reading choices realistic? How many words can reasonably be tested, and what does the pilot testing reveal about new word learning in an ER context?

TABLE 3  
Self-report checklist test, sample items

51. amnesia	YES	NS	NO
52. receptionist	YES	NS	NO
53. pub	YES	NS	NO
54. fridge	YES	NS	NO
55. helmet	YES	NS	NO
56. monument	YES	NS	NO
57. poppies	YES	NS	NO
58. untidy	YES	NS	NO
59. beach	YES	NS	NO
60. nasty	YES	NS	NO
61. magnifying	YES	NS	NO
62. studio	YES	NS	NO
63. drawing	YES	NS	NO
64. rum	YES	NS	NO
65. pirate	YES	NS	NO
66. squire	YES	NS	NO
67. candle	YES	NS	NO

Before I address each of these questions in turn, the participants and the reading program are described briefly.

## Method

### *Participants*

The 21 adult immigrant learners were ESL students at a community centre in Montreal.<sup>3</sup> First language backgrounds included Arabic, Chinese, Farsi, Korean, Polish, Spanish, and Russian. Some learners were recent immigrants, whereas others had been in Canada for as long as five years but had prioritized learning French, the language of everyday life in Quebec. Students were divided into two groups on the basis of oral and written answers to a short in-house placement questionnaire; proficiency levels ranged from elementary to high intermediate. The reading activities were offered as a supplement to the regular three-hour classes participants attended twice each week.

## Materials

The heart of the mini-library created for the reading project was a donated set of 35 graded readers at various levels of simplification.<sup>4</sup> Additional books were ordered to allow students of varying proficiency levels increased opportunities to choose books that interested them. The

collection eventually included about 70 different titles with at least two exemplars of each title, totalling around 150 books. Most of the readers were either new or classic fiction in the romance, suspense, and mystery genres.<sup>5</sup> The books varied in level of simplification from 400 to 3,800 headwords. The term *headwords* refers to lists of basic vocabulary used to guide the writing of simplified materials and to grade them. A writer working at, say, the 1,000-headword level may introduce other less frequent items that are key to a narrative, but these are normally explained in the text, glossed, or clarified in illustrations (Oxford University Press, 2003). Headword lists appear to vary from publisher to publisher and are not readily available. They should not be confused with research-based frequency lists used by Nation and Heatley (1996) to create LFP software. In the collection used in the experiment, there were five or more titles to choose from at each of six rough levels of simplification: 400–600, 800–1,000, 1,300–1,400, 1,800–2,100, 2,300–5,000, and 3,000–3,800 headwords.

During each week of the experimental period, participants had the opportunity to check out books during a break in their regular ESL classes. About an hour of class time each week was devoted to activities that supported ER. These included discussing books in pairs, completing worksheets, adding entries to vocabulary notebooks, or simply reading silently (although most of the reading was done independently at home). The number of books checked out by the 21 students who participated in the full six-week program from beginning to end varied widely. One exceptional student took home 33 books; another selected only three. The mean figure in the group was 10.52 books ( $SD = 6.71$ ). If we assume that students read the books they checked out, the average amount read per student was just under two titles per week ( $10.52 \text{ books} \div 6 \text{ weeks} = 1.75$ ). Most participants appeared to be enthusiastic about the program; some parents chose books for themselves and simpler ones to read aloud to their children. One eager participant involved her husband and children in reading *The Thirty-Nine Steps* and watching the video version *en famille*.

## Procedures

### *Scanning*

Several weeks before the reading project started, research assistants began scanning graded readers so that a word knowledge pre-test containing words from a sample of books in the collection could be created. Initially progress was slow, and it took more than three hours

to scan longer books, some of which were more than 100 pages long. To meet the goal of sampling words from at least a dozen books on the pre-test – two from each of six levels of simplification – it was decided to limit scans to the first 20 pages. In cases where books were short, this meant that books were scanned in their entirety, but usually some lesser proportion of a book was scanned. In 30 graded readers, with numbers of pages ranging from 20 to 108, the proportion of scanned pages ranged from 100% down to 18%, with a mean of 41.43% ( $SD = 22.75$ ). Although scanning entire books would have been preferable, the 20-page compromise was deemed reasonable; it was assumed that a large proportion of the lexical items occurring in a particular text would have already been introduced in the first 20 pages. The extent to which this is the case was explored in four books of varying lengths that were scanned in their entirety. The results are shown in Table 4, where the number of off-list word types appearing in 20 pages can be compared to the number in the entire book. Scanning 20 pages appears to capture a sizable proportion of the off-list content of the books, but it is also clear that the available word learning opportunities are under-represented, especially in longer books – a point that will be taken into account later, in the discussion of the learning results.

The decision to scan 20 pages per book meant that it was possible to scan 12 graded readers in the three weeks available before the ER program began and to prepare the pre-reading word knowledge test as planned. Once the 21 participants in the ER program had been pre-tested, they started checking out books from the collection. Selected titles were closely tracked. Early on, it became apparent that certain titles were favourites; the most-checked out book proved to be a romance simplified at the 1,300-headword level entitled *Just Good Friends*, chosen by 11 students. Popular books were prioritized for scanning; since there were at least two copies of each title, a book could be taken away for scanning for a day or two while its duplicate(s) remained in circulation. By monitoring checkout records to identify frequently chosen titles and prioritizing these for scanning, it proved possible to scan the first 20

TABLE 4  
Numbers of off-list types in entire graded readers and in 20 pages

Title	Off-list types in first 20 pages	Off-list types in whole book	Total pages in book
<i>John Doe</i>	32	35	42
<i>A Picture to Remember</i>	67	113	48
<i>Lady in White</i>	37	108	79
<i>A Love for Life</i>	79	278	112

pages of a large proportion of the different titles that the 21 participants chose to read over the six-week period. The records show that a total of 222 books were checked out during the experiment, representing 62 different titles. Almost all of the books circulated around the group, passing from student to student; only nine titles were chosen just once, and a smaller handful was not chosen at all. Of the 62 titles read by two or more participants, 37 were scanned, such that these 37 titles accounted for 149 of the 222 books that were checked out in the entire experiment. In other words, through judicious selection of widely read books for scanning, two thirds (149 of 222 = 67.12%) of the material participants selected to read during the six-week ER program could be evaluated for its vocabulary content and used to prepare the planned individualized post-tests.

### *Lexical profiling*

In order to target learning that may fairly safely be ascribed to ER rather than to other sorts of exposure, the experimentation focused mainly on infrequent words, although it was clear that more frequent words were also likely to be learned. In this study, 'infrequent' was operationalized as off-list, that is, words that do not appear on lists of the 2,000 most common word families of English (West, 1953) or on the AWL (Coxhead, 2000). As might be expected, lexical profiling reveals that simplified texts contain far fewer off-list items than texts written for native speakers. For example, a reader of the original version of *Treasure Island* would meet the following 145 off-list items in the initial 3,000 running words of the novel:

abominable admiral ah alarmed alongside assizes awaited awful bacon  
bade barrow beach berth bleared briskly buccaneer capstan cheek chorus  
clasp cocks coltish connoisseur cove curtained cutlass decks desirous  
diabolical dreadful effectual eternal exhausted fawning fetch filthy  
flapped fog folk fourpenny frost fury gales glared gray grog grumbling  
hamlet handspike handy haunted hawker hedge hilltops hilt hip ho hoar  
horn horrible incivility indignation insist knuckled lapping leap leer  
legged lingering livid loosened magistrate mast mate mighty mingled  
monstrous mought napkin nightmares oath overriding palm parlor  
particulars passion patched patted peering personage pigtail pirate  
plankplodding plucked quit ragged rapped rebuff reeled resumed rheu-  
matics ripple risen ruffian rum sabre sailorly scarecrow scarred scoundrel  
seafaring seaward sharer sheath shivering signboard sittuated skipper  
slammed slap smack sneering snort sonny spy squire stabling stare stormy



strode stroll surf tallowy tarry terrified threshold tilted tottering trundled  
tyrannized villainous wringing yo

By contrast, a reader of the simplified version, an Oxford Progressive Reader simplified to the 1,400-headword level, would meet only 11 off-list items:

admiral ail anchor cabin crew crutch knelt mate pirate rum squire

This pattern recurs, as can be seen in Table 5, which compares the vocabulary content of the first 3,000 running words of four original literary works available in machine-readable format (available at <http://gutenberg.net>) and their simplified counterparts. The graded novels in Table 5 were chosen to reflect off-list content in materials simplified to levels ranging from beginner (600 headwords) to advanced (2,500 headwords). Proper names (e.g., names of characters and places) have been excluded from the counts of off-list items in both original and simplified excerpts.

Although off-list items in the simplified texts are rather small in number, even the most simplified novel considered in this initial analysis appears to contain 11 off-list types among its first 3,000 words. This was reassuring, given the concern that such items might not occur at all in the more basic graded readers or might appear too rarely to be useful as targets in the investigation. Off-list items were found to occur in 20-page segments of all 37 readers that were scanned and profiled in this experiment, although in some instances numbers were small, as we will see in the discussion of test development. Another point that stands out in Table 5 is the low density of off-list words in all four simplified texts. The column showing the extent to which a simplified text consists of off-list words suggests that the conditions for learning these items are good; that is, they are surrounded by large numbers of common words

TABLE 5  
Off-list words in 3,000-word excerpts of original and simplified novels

Title	Original version		Simplified version		Headwords
	No. of types	% of running words	No. of types	% of running words	
<i>Anne of Green Gables</i>	153	6.67	11	1.83	600
<i>Treasure Island</i>	145	6.78	11	2.08	1,400
<i>Wuthering Heights</i>	383	13.45	49	3.31	1,800
<i>Pride and Prejudice</i>	79	3.58	36	1.94	2,500

and proper nouns (which can be supposed to carry a minimal learning burden). However, it should be noted that for many learners – beginners, in particular – many words are likely to be unfamiliar, not just those identified as off-list. Nonetheless, the simplified texts compare favourably to the much more lexically dense literary originals and clearly offer ESL learners improved, if not optimum, conditions for new word learning.

### *Testing*

#### Procedure

For experimental design reasons, we were most interested in the participants' pre- and post-test knowledge of off-list items that occurred in the ER materials. But we were concerned that participants might be overwhelmed by a test made up entirely of off-list words; it was possible that zones of more common words might prove to be an important area for growth for these beginning and intermediate learners. Therefore lists of potential test targets from three frequency ranges (1,001–2,000, AWL and off-list) were created for each of the 37 books. Words in the 0–1,000 frequency zone were not included; these were considered too likely to be already known or learned through other activities.

The development of the lists proceeded as follows. Once 20 pages of a reader had been scanned, the computer file was entered into VocabProfile to identify the items in the four frequency ranges. A portion of the VocabProfile raw output for off-list words in *The Thirty-Nine Steps* (Oxford University Press, 1,400 headwords) is shown in Table 6. First, all proper names were removed from the output. Among the items that remained, different forms of a word (e.g., *moor* and *moors*) were treated as two members of a single word family, following guidelines outlined by Bauer and Nation (1993).<sup>6</sup> Only word families that occurred two or more times qualified for inclusion on the test lists (see items in bold in Table 6). Any items that occurred only once were removed. These were seen as poor test targets, since a single reading encounter was unlikely to result in a learning event (though of course it is possible that such items recurred in the unscanned portions of a book). The results of the selection process for the selected sample appear at the bottom of Table 6.

Selection of test targets in this manner resulted in lists of varying lengths. An examination of 24 lists, representing the full range of levels of simplification in the collection, showed that the mean number of off-list text targets was 16.24 ( $SD = 6.17$ ). Numbers ranged from five to 26; there appeared to be a slight tendency for scans of more advanced

TABLE 6

Sample of VocabProfile output (off-list words) and selected test targets

Raw *VocabProfile* output for off-list words:

africa africa africa africa american american **armchair armchair** avocado bark bathroom  
 believable berlin biscuits boots **bored bored boring brandy brandy** britain britain britain  
 britain britain british british chase constantine crawled czechenyi digby digby **doorman**  
**doorman** dumfries dumfries dumfries england europe europe europe europe europe  
 europe fortnight franklin franklin galloway galloway galloway german german germans  
 germany greece greek greek hamburg hannay hannay hawk headache hofgaard hood  
 insides jar julia karolides karolides karolides karolides karolides karolides langham  
 langham lisped london london london london london london london london london  
 london london lunchtime luneville matabele meters **milkman milkman milkman milkman**  
**milkman** mirror **moor moor moor moor moor moors**

Remaining test targets:

armchair bored brandy doorman milkman moor

readers to produce longer lists of off-list words than readers at more basic levels. Very few AWL words were found in the readers, so these were not included in the testing. Target lists of words in the 1,001–2,000 frequency range were larger than the lists of off-list targets and varied considerably in length. Examination of 24 lists indicated a mean length of 46.42 words ( $SD = 23.75$ ).

The lists produced in the manner illustrated in Table 6 were used in preparing a 100-item pre-test that presented each target word along with three rating options (NO, NS, YES) as shown in Table 3. Fifty of the tested words came from lists of items in the 1,001–2,000 frequency range; four or five words were randomly chosen to sample each of 12 scanned books (two from each of six levels of simplification). The other 50 items were chosen in a similar fashion from lists of off-list words that occurred in the 12 readers.<sup>7</sup> The 100-word test seemed long enough to provide useful information without taking up undue amounts of class time. It was expected to take 20 minutes to complete.

At the end of the six weeks, enough scanning had been accomplished to allow the development of individualized post-tests for 17 of the 21 participants. Each of these 100-item tests was identical to the pre-test in format but was made up entirely of words known to have occurred in four of the books a participant had chosen to read – again there were 50 words from the 1,001–2,001 list and 50 off-list words. The four participants for whom no individualized test could be made took the same default test that sampled words from graded readers they had not read. The discussion of the results below is based on the group of 17 who took the personalized tests.

## Results

Concerns about the test format proved unfounded. The participants were undaunted by the long list of words, took the ratings task seriously, and finished the test in 10 minutes or less. To assess the learning gains, numbers of words rated YES (I know the meaning of this word) on the pre- and post-tests were tallied and compared. Words in the potentially interesting NS (not sure) category were not included in these counts. All participants registered higher numbers of YES ratings on both 50-word sections of the test at the end of the six weeks than they had at the beginning. Analysis with *t*-tests for paired samples indicated a significant mean post-test increase of about seven words rated YES ( $M = 6.59$ ,  $SD = 5.47$ ) on the measure of words from the 1,001–2,000 most frequent list and a significant mean increase of about 10 words on the test of off-list items ( $M = 10.29$ ,  $SD = 7.62$ ). These results are shown in Tables 7 and 8.

If students are reporting their knowledge honestly, the figures in the last columns in Tables 7 and 8 point to an overall mean gain of about 17 words ( $6.59 + 10.29 = 16.88$ ). For the purposes of evaluating vocabulary growth through ER, the 10-word increase in knowledge of off-list words is the more interesting finding, since we can be reasonably confident that these more unusual items were learned through encountering them in the reading materials rather than through other exposure. The 10-word mean gain is more impressive than it may seem: A 17-word-sized gap in the participants' pre-ER baseline knowledge of off-list items (50 tested – 33.30 rated YES = 16.70 unknown) appears to have been reduced

TABLE 7  
1,001–2,000 words rated YES of 50 words total ( $n = 17$ )

	Pre-test	Post-test	Difference
<i>M</i>	41.35	47.94	6.59
<i>SD</i>	5.38	1.89	5.47

$t = 5.47$   $p < .001$

TABLE 8  
Off-list words rated YES of 50 words total ( $n = 17$ )

	Pre-test	Post-test	Difference
<i>M</i>	33.80	43.59	10.29
<i>SD</i>	8.18	4.30	7.62

$t = 2.78$   $p < .001$

to only seven unknown words (50 tested – 43.59 rated YES = 6.41 unknown). In other words, these results suggest that participants learned well over half of the unfamiliar off-list words they encountered in the simplified readers. Participants also appear to have acquired knowledge of well over half of the unknown 1,001–2,000–level words they encountered, although there may be a role for other sources of exposure in learning these more common words. The findings must be treated with caution, since the pre-post differences are based on comparisons of performance on two different tests, the baseline measure with words from 12 readers, and individualized post-tests with words from four readers.

As a check on reported learning gains, participants were asked to complete a second individualized measure. The format of this measure was based on Wesche and Paribakht's Vocabulary Knowledge Scale (1996). The off-list words selected for testing in this manner met two criteria: (a) the participant had rated the word NO (I do not know this word) on the pre-test, and (b) the word occurred in a book that the participant was known to have selected during the six-week experiment. This meant that in addition to rating confidence levels at the end of the experiment, participants had an opportunity to actually demonstrate new word knowledge.<sup>8</sup> A sample item is shown in Table 9.

Not many items met the criteria. As the mean knowledge score in Table 8 indicates, participants rated many words YES on the pre-test, leaving a small remainder of unknown or partially known off-list words to choose from. Also, since students were free to choose their reading, there was no guarantee that they would eventually read books that had been sampled in the pre-testing and thus meet those target words – a cost of situating the study in an authentic ER setting. Fortunately, 16 of 17 participants did select a pre-tested book, and it proved possible to test them on at least one word they had initially rated NO and subsequently encountered. Some students were tested on as many as three words. Of

TABLE 9  
Sample item requiring demonstration of word knowledge

---

**pirate**

- 1 I don't know what this word means.
  - 2 I have seen this word before, and I think it means \_\_\_\_\_
  - 3 I know this word. It means \_\_\_\_\_. (Give the meaning in English, French or your language.)
  - 4 I can use this word in a sentence. (Write a sentence.)
- 

(If you choose 4, please also complete 3.)

---

the 35 words tested, there was an indication of new full or partial knowledge in 18 instances.<sup>9</sup> Responses such as 'one kind of drink' for *rum* or 'way' for *lane* were considered evidence of partial knowledge. More complete responses, such as the definition 'the open place where green grass grows' for *moor*, accompanied by a meaningful and grammatically acceptable sentence, 'There is a *moor* nearby that mountain,' were considered to be evidence of full knowledge. This result is based on a rather small body of data, but it lends credibility to the gains participants reported on the self-assessment measure. Again, we see growth in more than half of the cases of previously unknown off-list words (18 of 35 = 51.43%). This sizable result stands in marked contrast to gains documented in other studies of word learning through reading, where gain proportions tend to be much lower (e.g., the one new word in 20 documented by Nagy et al., 1987).

### Discussion

On the basis of this pilot study, it is safe to conclude that the proposed methodology for investigating ER is feasible. In a relatively short period (about two months), it was possible to build a large computer-readable corpus of ER materials and identify the word-learning opportunities that occurred in them in a systematic manner. The flexible individualized testing technique that can be applied to any text a learner chooses to read offers a way of overcoming a major difficulty in assessing vocabulary gains achieved through ER – the problem of the source of learning. Perhaps the most promising aspect is the possibility of identifying a large number of words that an ER participant does *not* know at the outset using the checklist method and evaluating what changes occur in knowledge of these items after the participant encounters them in the ER material he or she has chosen. In the pilot study, participants' post-treatment ability to provide synonyms and/or use the target items in sentences was explored using a modified version of the Vocabulary Knowledge Scale (Wesche & Paribakht, 1996). But other kinds of measures could be used to explore other aspects of word knowledge; participants' ability to recognize collocations, recall additional meanings, produce various morpho-syntactic forms, or simply spell the target words are among the many possibilities. In addition, the corpus of simplified texts offers the possibility of exploring text characteristics that may help explain learning outcomes: Interesting variables include the number of times target words occur in the materials, their importance to the events of the narrative, and the extent to which their meanings are supported by the language of the contexts in which they occur.

One of the main benefits claimed by proponents of ER is increased lexical access speed (Grabe & Stoller, 2002); indeed, the main value of ER may be the opportunities it offers to develop more rapid recognition of frequent words rather than the opportunities it affords for learning the meanings of infrequent words. The methods piloted in this study can be easily applied to investigating gains in automaticity achieved in a program of ER.<sup>10</sup> With complete scans of a bank of graded readers in place, it is a fairly simple matter to identify a set of words from a given frequency range (e.g., words on the 1,001–2,000 most frequent list) that occur in materials a learner has read. The learner's reaction times for these words can be compared to those for words from the same frequency range that do not occur in the book to determine the effects of exposure on recognition speed. The expanded study currently underway explores this intriguing avenue.

Performance on both measures used in the pilot study indicated that participants gained new knowledge of more than half of the unfamiliar words that occurred in the ER materials they selected. This is an impressive result, given findings of previous investigations of word learning through reading. As mentioned, a study of school-age L1 readers of English by Nagy et al. (1987) identified a mean growth rate of one unfamiliar word in 20; an overview by Horst et al. (1998) points to a mean rate of about one in 12 across studies of L2 readers.

However, a number of caveats are in order. First, comparison to rates in previous studies of incidentally acquired vocabulary knowledge is unwarranted for reasons outlined earlier in this article: The conditions of ER differ substantially from those of most reading studies. Also, while it is possible that the ER setting in the pilot research was more conducive to new word learning than the settings of earlier reading studies, it is important to point out that acquiring new knowledge at the rate of one unfamiliar word in two represents a small amount of growth if new words are scarce in the input. The large proportions of off-list words rated YES on the pre-reading measure (see Tables 7 and 8) suggests that the participants did not have many opportunities to meet unfamiliar words in the ER materials. This problem can be addressed in a future study by administering a measure of vocabulary size such as the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001); participants can then be guided to choose ER materials at levels of simplification that offer more word learning challenges, following guidelines by Nation and Wang (1999).

Second, the rate itself is questionable. The finding is based on a small number of self-rated words (100 items) and an even smaller test of demonstrated knowledge (one to three items per participant). Testing on

a much larger scale is warranted and can be managed; participants finished the rating task and follow-up test more quickly than expected and could easily have been asked to do more. A larger pre-reading checklist measure would allow more opportunities for NO responses to words and, in turn, more opportunities to investigate growth after these unfamiliar words have been encountered in the graded readers. A second project that evaluates pre-reading word knowledge gains using a larger test of 250 off-list words is currently underway.

It is also unclear whether the gains made through ER were long lasting. The pilot study participants have moved on in their lives as language learners, so it has not been possible to administer a delayed post-test; this is an omission that a future study should attempt to address. A closer examination of changes in knowledge of words rated NS (not sure) can also contribute to a more accurate picture of growth. Matrix modeling of vocabulary learning through reading (Horst & Meara, 1999; Horst, 2000) has delineated the effects of reading encounters on levels of partial word knowledge and traced the evolution toward more complete knowledge, but this, too, requires delayed post-testing. The lack of information about what participants did as they read and how this may have affected learning outcomes is also problematic. Many of the learners kept vocabulary notebooks; some participated in paired reading discussions during the regular ESL classes; and some probably also used dictionaries extensively. The impact of these activities (all integral parts of a well-designed ER program) on the learning results is not clear; a future study should take these factors into account.

Perhaps the most important aspect to address is the incomplete corpus of graded readers. Until it is completed, no definitive conclusions about word learning can be drawn using the proposed methodology. Testing based on 20-page excerpts of scanned material cannot adequately sample the word learning opportunities available to participants who read the entire texts. Ideally, the lists of words used to make the word knowledge tests should be based on whole books and feature all the items that appear in them (including words that occur only once). The analyses shown in Table 4 indicate that the 20-page compromise meant that vocabulary learning opportunities were substantially under-represented in the pilot study. It is hoped that the problem of scanning in volume may be solved by persuading publishers to supply machine-readable versions of entire texts; initial forays in this direction have not (yet) proved fruitful. In the meantime, research assistants are engaged in the task of scanning entire graded readers in an ever-increasing collection.

Finally, the discussion returns to its starting point – reading in large



amounts. It is worth pointing out that none of the participants in the ER program investigated here read only one book. The mean gain figure of about 10 new words reported in Table 8 was based on reading four books. If these modest gains are extrapolated to the average 10.5 titles read in the pilot study and to whole books rather than 20-page excerpts, it is possible to claim that the participants learned dozens of new words during the six-week experiment. But volume is clearly crucial. Requiring ESL learners to read a single simplified reader or two per semester as a way of enriching an existing course is hardly to be discouraged; but such minimal programs of reading cannot be expected to result in great amounts of new vocabulary knowledge. For L1 and L2 learners alike, the project of acquiring a sizable mental lexicon appears to involve exposure to a great deal of written text, and native speakers of a language who have been reading and building vocabulary knowledge since their school days have a huge head start. For adult L2 learners aspiring to levels of proficiency beyond the most basic, making up the lexical distance presents a considerable challenge. ER programs may help to narrow the gap – but only if they can motivate learners to read in large amounts.

**Marlise Horst** is Assistant Professor in the Department of Education at Concordia University, where she teaches courses in L2 vocabulary acquisition, language testing, TESL methodology, and the history of English. Her research interests include L2 vocabulary learning through reading. She has taught English in Egypt, Saudi Arabia, Oman, Hong Kong, and North America. She holds a PhD in Applied Language Studies from the University of Wales in Swansea, UK. Contact information: marlisehorst@yahoo.ca

### Acknowledgements

I would like to recognize the invaluable contributions of two research assistants, Sumanthra Govender and Ioana Nicolae, both graduates of the MA in Applied Linguistics program at Concordia University. I am also grateful to Tom Cobb, Michèle Plomer, and the anonymous *CMLR* reviewers for their constructive suggestions, and to Petronella Beran and ESL students at the Tyndale St.-Georges Community Centre for their enthusiastic cooperation. The research was supported by funding provided by Fonds québécois de la recherche sur la société et la culture (FQRSC) and Concordia University.

### Notes

- 1 As Nation (2002, March) observes, the publication of graded readers has been largely a British undertaking. A six-title collection (*Ladder* series)

published by the United States Information Service appears to be an exception; I am not aware of other graded ESL series of North American provenance.

- 2 High unknown word densities may also have been a problem in the ER studies in Table 1 in cases where learners read materials simplified for native speakers of English. See column labelled 'Reading materials.'
- 3 In fact, 34 students were originally pre-tested, but 13 either left the program or were not present for post-testing. Others joined the program and read enthusiastically but were not included due to lack of pre-test data.
- 4 I am grateful to Dr. Patsy Lightbown and Cambridge University Press for donating these materials.
- 5 Some non-fiction titles were also included following recommendations by Gardner (1999). His analyses indicate that the chances of repeated encounters with infrequent lexical items are higher in this genre than in fiction.
- 6 A word family was defined following guidelines for Level 3 in the scheme devised by Bauer and Nation (1993). In this scheme, a family consists of a baseword along with inflected forms and derived forms using frequent and regular affixes (*-able, -er, -ish, -less, -ly, -ness, -th, -y, non-, un-*).
- 7 The on-line *Vassarstats* randomizer (available at <<http://faculty.vassar.edu/lowry/VassarStats.html>>) was helpful in selecting words at random from lists of varying lengths.
- 8 The format of the VKS was adjusted to suit the context of the study. The original version by Wesche and Paribakht (1996) includes self-report options that read 'I don't remember having seen this word before,' and 'I have seen this word before but I don't know what it means.' We omitted some of the language about having previously seen the words because it did not seem very meaningful in our context where students had indeed seen the targeted items on another measure (the ratings instrument) moments before taking the demonstration test.
- 9 Because of the very small data set, the full and partial knowledge categories were conflated into a single 'known' category, following Paribakht and Wesche's use of a basic known/not known dichotomy in their study of word learning through reading (1997, p. 189). Words that were either defined correctly or defined correctly and used appropriately in a sentence were considered known and were assigned one point. Words that were either indicated as not known or that were defined incorrectly were considered unknown and were assigned a score of zero. Two native speaker raters evaluated the responses; rater agreement was 97%. In a larger future study, I expect to use this type of instrument to measure changes in levels of word knowledge in a more nuanced way.
- 10 There are research precedents for investigating increased automaticity in studies of ESL readers; several of the ER studies summarized in Table 1

included measures of reading speed (Bell, 2001; Lai, 1993; Mason & Krashen, 1997; Robb & Susser, 1989). However, as in the case of assessing new word gains, these studies do not measure access speeds for words readers are known to have encountered in the ER materials.

## References

- Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, 6, 253–279.
- Bell, T. (2001, April). Extensive reading: Speed and comprehension. *The Reading Matrix*, 1(1). Retrieved May 30, 2003, from <http://www.readingmatrix.com/articles/bell/index.html>
- Cho, K. S., & Krashen, S. D. (1994). Acquisition of vocabulary from the Sweet Valley Kids series: Adult ESL acquisition. *Journal of Reading*, 37, 662–667.
- Cobb, T. (n.d.). *The complete lexical tutor* [Web site]. Available: <http://www.lextutor.ca>
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34, 213–238.
- Day, R. R., & Bamford, J. (1998). *Extensive reading in the second language classroom*. Cambridge: Cambridge University Press.
- Day, R. R., & Bamford, J. (2002). Top ten principles for teaching extensive reading. *Reading in a Foreign Language*, 14, 136–141.
- Elley, W. B. (1991). Acquiring literacy in a second language: The effect of book-based programs. *Language Learning*, 41, 375–411.
- Gardner, D. I. (1999). *Vocabulary acquisition through reading: Assessing the lexical composition of theme-based text collections in upper-elementary education*. Unpublished doctoral thesis. Northern Arizona University, Flagstaff.
- Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. Harlow, UK: Longman.
- Horst, M. (2000). *Text encounters of the frequent kind: Learning L2 vocabulary through reading*. Unpublished doctoral thesis. University of Wales, Swansea, UK.
- Horst, M., & Meara, P. (1999). Test of a model for predicting second language lexical growth through reading. *The Canadian Modern Language Review*, 56, 308–328.
- Horst, M., Cobb, T., & Meara, P. (1998). Beyond a Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, 11, 207–223.
- Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, 51, 539–558.
- Krashen, S. (1993). *The power of reading: Insights from the research*. Englewood, CA: Libraries Unlimited.

- Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *Modern Language Journal*, 73, 440–464.
- Lai, F. K. (1993). The effect of a summer reading course on reading and writing skills. *System*, 21, 87–100.
- Lightbown, P.M. (1992). Can they do it themselves? A comprehension-based ESL course for young children. In Courchène, R., St. John, J., Thérien, C., & Glidden, J. (Eds.), *Comprehension-based second language teaching: Current trends* (pp. 353–370). Ottawa, ON: University of Ottawa Press.
- Mason, B., & Krashen, S. (1997). Extensive reading in English as a foreign language. *System*, 25, 91–102.
- Meara, P., Lightbown, P. M., & Halter, R. (1997). Classrooms as lexical environments. *Language Teaching Research*, 1, 28–47.
- Nagy, W. E., Anderson, R. C., & Herman, P. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, 24, 237–270.
- Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, 20, 233–253.
- Nation, I. S. P. (1997, May 21). The language learning benefits of extensive reading. *The Language Teacher Online*. Retrieved May 30, 2003, from <http://www.jalt-publications.org/tlt/files/97/may/benefits.html>
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Nation, I. S. P. (2002, March). *Replacing the General Service List: Words from the British National Corpus*. Paper presented at the meeting of the Second Language Vocabulary Colloquium, Leiden, Germany.
- Nation, I. S. P., & Heatley, A. (1996). *VocabProfile, Word, and Range: Programs for processing text*. Wellington, New Zealand: LALS, Victoria University of Wellington.
- Nation, I. S. P., & Wang, K. M. (1999). Graded readers and vocabulary. *Reading in a Foreign Language*, 12, 355–380.
- Oxford University Press (2003). *How are Oxford readers graded?* Retrieved May 30, 2003, from [http://www.oup.com/elt/global/catalogue/readers/article\\_readers/](http://www.oup.com/elt/global/catalogue/readers/article_readers/)
- Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady and T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 174–200). Cambridge: Cambridge University Press.
- Parry, K. (1991). Building a vocabulary through academic reading. *TESOL Quarterly*, 25, 629–653.
- Pitts, M., White, H., & Krashen, S. (1989). Acquiring second language vocabulary through reading: A replication of the Clockwork Orange study using second language acquirers. *Reading in a Foreign Language*, 5, 271–275.

- Robb, T.N., & Susser, B. (1989). Extensive reading vs. skills building in an EFL context. *Reading in a Foreign Language*, 5, 239–251.
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing* 18, 55–88.
- Wesche, M., & Paribakht, T.S. (1996). Assessing vocabulary knowledge: Depth versus breadth. *The Canadian Modern Language Review*, 53, 13–39.
- West, M. (1953). *A general service list of English words*. London: Longman, Green & Co.

## Appendix A

### *Graded ESL readers referred to in the study*

- Austen, J. (2000). *Pride and prejudice*. Oxford Bookworms Library, Stage 6. Oxford: Oxford University Press.
- Brenan, F. (2000). *The fruit cake special and other stories*. Cambridge English Readers, Level 5. Cambridge: Cambridge University Press.
- Brontë, E. (2000). *Wuthering heights*. Oxford Bookworms Library, Stage 5. Oxford: Oxford University Press.
- Buchan, J. (2000). *The thirty-nine steps*. Oxford Bookworms Library, Stage 4. Oxford: Oxford University Press.
- Campbell, C. (1999). *Lady in white*. Cambridge English Readers, Level 4. Cambridge: Cambridge University Press.
- Hancock, P. (1999). *A love for life*. Cambridge English Readers, Level 6. Cambridge: Cambridge University Press.
- Hancock, P. (1999). *Just good friends*. Cambridge English Readers, Level 3. Cambridge: Cambridge University Press.
- Hardy, T. (1979). *Mayor of Casterbridge*. Nelson Readers, Level 5. Walton-on Thames, UK: Nelson.
- Montgomery, L.M. (2002). *Anne of Green Gables*. Penguin Readers, Level 2. Essex, UK: Pearson Longman.
- Moss, A. (1999). *John Doe*. Cambridge English Readers, Level 1. Cambridge: Cambridge University Press.
- Scott-Malden, S. (1999). *A picture to remember*. Cambridge English Readers, Level 2. Cambridge: Cambridge University Press.
- Stevenson, R.L. (1992). *Treasure Island*. Oxford Progressive English Readers, Level 1. Hong Kong: Oxford University Press.
- Thackeray, W. (1991). *Vanity fair*. Longman Classics, Simplified Edition, Stage 3. Essex, UK: Longman.