Reading in a Foreign Language October 2022, Volume 34, No. 2
ISSN 1539-0578 pp. 232-248

Extensive Reading for a 9,000-Word Vocabulary: Evidence from Corpus Modeling

Clarence Green
Federation University Australia
Australia

Abstract

This paper contributes to a research program within extensive reading (ER) and *Reading in a Foreign Language* using corpora to simulate ER input to develop vocabulary through incidental learning to 9,000 words. This helps researchers/teachers evaluate ER. If corpora indicate no 'pathway' from smaller to larger vocabulary sizes through authentic ER input, with vocabulary recurrence rates sufficient for incidental learning, then graded readers or other pedagogy appear essential. Studies offer different conclusions due to modeling issues. This study replicates previous research on a larger corpus of general fiction, with improved modeling. For every vocabulary size, a substantial amount of comprehensible fiction is found, with enough repetition of vocabulary from subsequent levels that pathways from smaller to larger vocabulary sizes are possible without graded readers. Prior estimates of approximately 3 years to acquire 9,000 words at 1 hour a day are underestimates, with modeling indicating 2 hours a day would be required.

Keywords: extensive reading, graded readers, vocabulary acquisition, coverage comprehension model, comprehensible input, incidental learning

This paper contributes to a research program within extensive reading (ER) that has used corpora to simulate the possibility of ER providing enough input to develop a student's vocabulary via incidental learning up to the first 9,000 words of English. This vocabulary size has often been cited as a goal for general reading comprehension (Nation, 2014). Previous studies have modeled how many of the most frequent 9,000 words of English can be encountered through ER, how many repetitions there are with these targets, and if repetitions occur at rates associated with incidental learning. Additionally, time-frame simulations have evaluated whether incidental learning is likely or efficient through ER. Other important questions addressed include how much comprehensible input for different vocabulary sizes exists in unsimplified general fiction. Some simulations have concluded that ER would necessarily require graded readers for learners with mid-level vocabulary sizes, while others have concluded that combinations of graded readers and authentic fiction are possible given comprehensible input appears available in unsimplified general fiction (McQuillan, 2016a, 2016b).

This research program can be seen as an ongoing conversation amongst some leading vocabulary researchers over the past 15 years, mostly within the pages of *Reading in a Foreign Language* (*RFL*). Key publications include Cobb (2007, 2016), Nation (2014), McQuillan (2016a, 2016b), McQuillan and Krashen (2008), and Schmitt et al. (2017). The findings are highly cited, but with regard to practice, the publications make conflicting recommendations. With regard to theory, these studies test the Comprehension Coverage Model (CCM), which McLean (2021) argues is "the cornerstone of research and pedagogy concerning the importance of vocabulary to reading" (p. 127). The CCM proposes that text becomes comprehensible when a reader's vocabulary size covers most of the words in a text, and comprehensible input is optimal for incidental learning. McLean (2021) focuses on 98% as a coverage proxy for comprehension in his recent *RFL* review, with 95% being the other most cited figure (Nation, 2006; Laufer, 2020). It is argued that for ER to be effective, it should be guided by the CCM.

Researchers have been interested in the potential for ER, based on general fiction self-selected by students, to "do the entire job" (Krashen, 1989, p. 448) of building a vocabulary of 9,000 words incidentally. To offer a quick orientation, Cobb (2007, 2016) and Schmitt et al. (2017) concluded that corpus simulations indicated it is impossible because there is not enough repetition of vocabulary in authentic text in reasonable time frames to incidentally acquire a large vocabulary, nor is there enough reading material that would be consistent with the CCM for different vocabulary sizes. McQuillan (2016a, 2016b) and McQuillan and Krashen (2008) concluded it is possible, based on their corpus simulations. Nation (2014) is somewhere in the middle of this debate. He argued that his corpus simulations, which have been interpreted differently by both sides of the debate, indicate that it may be possible to achieve this but not without mid-frequency graded readers.

To model if an incidental learning pathway to 9,000 words is possible, books that can be read with more than 95%–98% vocabulary coverage need to exist for vocabulary sizes from beginner to proficient (Schmitt et al., 2017). There also needs to be enough books to enable self-selection by interest "at the right level" (Nation, 2014, p. 7). The repetition of the target words needs to be at frequencies associated with incidental learning (Uchihara et al., 2019). Finally, input amounts and timeframe estimates are required to evaluate whether ER is likely an efficient pedagogy within a reasonable amount of time, even if previous conditions are met (McQuillan, 2016b). If simulations cannot show the above, the hypothesis that wide reading alone might do the job through incidental learning and the CCM would appear unlikely (Cobb, 2016).

There is a lack of consensus based on corpus models published thus far. This may be due to some methodological issues. One issue is that some conclusions have been drawn from recomputing numbers in earlier simulations published in *RFL* rather than replicating them to ensure they are accurate. The time estimates given in McQuillan (2016a, 2016b) are that 3years of reading at 1 hour a day may be enough to reach 12 encounters with 9,000-word families, possibly sufficient repetition for incidental learning. However, the estimate is based on the numbers in Nation's (2014) modeling, and that study has several limitations. Another issue is that other than McQuillan (2016b), no attempt has been made at estimating how difficult it is to find comprehensible input in general fiction for readers with different vocabulary sizes. This is an essential question when evaluating if learners could potentially self-select pathways from smaller to larger vocabularies without graded readers. While McQuillan (2016b) found in an ad-hoc sample of juvenile fiction that there are books for each vocabulary size progressively up to 9,000 words, indicative of a CCM pathway, his modeling did not indicate if such books were widespread in a representative sample (i.e., easy to find), nor if it would be possible to navigate a pathway through general fiction. Nation (2014) and Cobb (2007) used small corpora, which do not represent ER well. No study yet has modeled how much input for the next vocabulary-size levels are provided by reading comprehensible input at a specific vocabulary size. For example, a student may have a vocabulary size of 3,000 words and reads comprehensively any material where this vocabulary size covers 95% or more of the words in a book. However, modeling has not checked if reading such comprehensible input will provide input from the next vocabulary size target (e.g., 4,000-word band) at rates associated with incidental learning. This is required to better evaluate whether a pathway is possible through reading 'at the right level.' The current study therefore undertakes a conceptual replication with methodological improvements to previous studies.

Literature Review

A Review of the Core Issues

The core issues in this research program are as follows. Firstly, there has been an interest in understanding how many words need to be known for comprehension, which is often operationalized as what vocabulary coverage of text correlates with comprehension (Schmitt et al., 2017). Since Hu and Nation (2000), the figures 95%–98% vocabulary coverage have been the most cited. Laufer (2020) concludes 95% constitutes the minimal lexical coverage needed, and 98% the optimal. This link between vocabulary and comprehension has come to be known as the Comprehension Coverage Model (McLean, 2021). Understanding what is read is important because comprehensible input (Krashen, 1989) provides optimal conditions for the development of new vocabulary through incidental learning. This research linking vocabulary coverage to comprehension led to corpus research exploring the vocabulary sizes needed by second language learners for reading comprehension.

The most commonly cited vocabulary size is Nation's (2014), which proposes "a vocabulary size of 9,000 words or more is a sensible long-term goal for unassisted reading of unsimplified texts" (p. 5). This estimate has been recommended as one answer to the question of how much vocabulary is needed to use English (Schmitt et al., 2017). The figure was estimated by corpus coverage studies based on the BNC-COCA word level lists. Nation (2014) simulated how much vocabulary coverage was provided by how many 1,000-word levels for a range of different corpora. Such simulations found that to reach 95%-98% coverage of most reading material, the first 9 BNC-COCA lists (i.e., the most frequent 9,000-word families) sufficed.

While 9,000 words may be a good target for general reading comprehension, any text is 'at the right level' below 9,000 words if the vocabulary size of the reader covers 95% or more of the text's vocabulary. Therefore, corpus research has tried to model whether there is 'pathway' of reading material constituting comprehensible input at each stage of vocabulary size development (Cobb, 2016) (i.e., material for smaller vocabulary sizes up to the target 9,000 words) aligned with the CCM at each step. In the absence of finding a pathway, one solution has been graded readers, matched for a student's current vocabulary size (McLean, 2021). But graded readers have been criticised as not being authentic language input and offer limited opportunities for students to align their interests with a book compared to self-selection from a large library (McQuillan, 2016b). Corpus research has also modeled the number of words that would need to be read to build a vocabulary incidentally, and estimated reading time. It is a problem if corpus modeling shows that ER would not provide enough encounters with target vocabulary to support incidental learning in any reasonable amount of reading time. It is also a problem if authentic reading material that is comprehensible for different vocabulary sizes is hard to find, or even non-existent for some vocabulary sizes (Cobb, 2007).

The Cobb, Nation, McQuillan and Krashen Studies

The key publications that have built on each other's methods and findings are Cobb (2007, 2016), McQuillan and Krashen (2008), Nation (2014), McQuillan (2016a, 2016b) and Schmitt et al. (2017). Cobb (2007) provided the initial modeling of whether reading could provide enough repetition of target words for incidentally learning a larger vocabulary in a pedagogically reasonable amount of time. Cobb (2007) took a random sample of words from a wordlist generator in Lextutor (http://www.lextutor.ca) and produced 20 sets of 10-word samples from three frequency bands in the BNC lists: 1,000, 2,000, and 3,000. One of these sets was then selected randomly as possible learning targets representing the frequency band. He computed in different subsections of the Brown corpus (fiction, press, academic), how many times target words occurred (REGEX stemming to search for word families). Despite variation in the number of occurrences in fiction, press, or academic texts, at the 1,000 level, all but 1-2 words of the sample would be met at rates that might support incidental learning, which he set at the threshold of 6 encounters. Beyond the first 1,000 words, Cobb (2007) concluded the simulations paint a bleaker picture. Five of the 2,000-level words were below 6 repetitions; and for the 3,000-level words, none reached six repetitions.

Cobb (2007) then undertook a more specific analysis of fiction, since self-selected pleasure reading has been a core recommendation of Krashen (2004). Cobb (2007) used 300,000 words from seven *Jack London* novels. The results indicated they contained 817-word families from the 3,000 BNC level, but only 469 were repeated 6 or more times, 348 were encountered 5 times or less, 181 were met 2 times or less. Less than half were at the incidental learning threshold set by Cobb (2007), so he concluded that even for a modest vocabulary size of 3,000, corpus modeling showed "the extreme unlikelihood of developing an adequate L2 reading lexicon through reading alone, even in highly favourable circumstances" (Cobb, 2007, p. 28).

There are limitations with Cobb (2007). One issue is the sample size. In the first study, a corpora of only 175,000 words represented fiction. A small sample such as this would restrict the number of occurrences of any target set of words and the possibility of incidental learning. Cobb (2007) argued this sample size was an optimistic representation of how much free reading students can undertake over 1–2 years. In response, McQuillan and Krashen (2008) argued this is a pessimistic projection and unrealistically low for a simulation of ER. They computed that if reading at even a slow rate of 100 words per minute (wpm) this would equate to less than 5 minutes of reading a day per year. Nevertheless, the paper and its conclusions appear to be largely stood by in Cobb (2016) and Schmitt et al. (2017).

Nation (2014) built on Cobb (2007) with a more extensive study. He framed his paper as investigating whether it is possible to learn the first 9,000 words solely by reading, taking into account calculations of the amount of input and time needed to meet targets at rates associated with incidental learning. He selected 12 repetitions based on a review of incidental learning in experimental literature. Nation (2014) examined: a) the number of words that needs to be read to encounter the 9,000 words 12 times; b) if that is a reasonable/unreasonable amount of input given the time it would take based on reading rates of 150-200 wpm; and c) whether the answers to (a) and (b) are better or worse if different combinations of input are considered (e.g., fiction and non-fiction, fiction and movies, etc.) Nation (2014) constructed several 1-million-word corpora, which included novels, spoken language transcripts, movie and television scripts. He then computed occurrences of the BNC-COCA word family lists, up to the first 25 thousand most used word families of English.

Of particular importance is Nation's (2014) fiction corpus modeling. As a proxy, 25 novels from Project Gutenberg represented ER (e.g., *Alice in Wonderland, Animal Farm, The Great Gatsby, Ulysses*). Nation's (2014) simulations showed for the 2,000 level, 805 target words would be encountered on average 12 times with 171,411 words of input. To make further calculations, he computed an average novel length in his corpus, and suggested that this input equates about 2 novels. The equivalent of 3 additional novels containing 300,219 words of input would provide 830 of the 3,000 level words with an average of 12 times. Nation (2014) continued for each level up to 9,000 and estimated that about 2,956,908 words (25 novels) would provide even 805 of the 9,000 level word family targets at an average of 12 encounters. Nation's (2014) simulations are encouraging for ER and he concludes that "if learners read a total of 3 million tokens, then they would meet the first 9,000 words often enough to have a chance of learning them" (p. 7). He then asked if this amount of input is possible. Modeling input at reading rates of 150 and 200 wpm, he concluded that about 8 hours 20 minutes a week for 40 weeks a year for the slower reader, and 6 hours 10 minutes per week for 40 weeks a year for the faster reader, could be sufficient. Nation (2014) concluded this was a "largely positive study of the opportunities for learning through input" (p. 13). Nation (2016) clarified his time frame estimate in a later discussion article, however, noting, "My wishful thinking based on the data in the 2014 article is that with appropriate reading material and encouragement, it is feasible to learn close to 1,000 words a year through reading" (p. 305) (i.e., approximately 9 years for 9,000 words).

By examining a few novels in the corpus, Nation (2014) showed that they cannot be read with a mid-range vocabulary size of 5,000-8,000 words and only become comprehensible when a vocabulary size of 9,000 words was reached. Nation (2014) therefore concluded that finding comprehensible input would be challenging because "unsimplified text clearly provides poor conditions for reading and incidental vocabulary learning for learners whose vocabulary sizes are less than 9,000-word families" (p. 13). He recommended that, for ER, teachers should use graded readers for the mid-frequency vocabulary ranges.

Nation (2014) acknowledged "serious problems with the crude calculations" (p. 12). This study has a problematic corpus. While larger than Cobb (2007), it is still only 25 dated novels from the public domain. Another problem is the use of an average frequency of 12 for the target words in a frequency band, which undermines the stipulation of a word needing a minimum of 12 encounters. Another issue is that Nation (2014) assumed the reading material is "at the right level" (p. 7), that is comprehensible input that students can read with 95% or above vocabulary coverage. The corpus indicates this is unlikely. No language learner is reading *Ulysses* as comprehensible input while progressing to a vocabulary size of 9,000 words. Modeling from a corpus not representing reading material at the right level for different vocabulary sizes cannot offer evidence whether it is possible to read 3 million words and progress along a pathway of comprehensible input; nor does it suggest that the words needed to move up a level in reading comprehension are available in reading material at the right level for smaller vocabulary sizes. Nation (2014), to be fair, clearly acknowledges several of these issues.

McQuillan (2016a) interpreted Nation (2014) as showing that with 11 million words of input, at 150 wpm, a student could reach 9,000-word families in about 1,200 hours, equivalent to about 1 hour a day over 3 years: “very doable for a motivated adult or adolescent acquirer” (p. 65). Using Nation’s (2014) figures, McQuillan (2016a) modeled the input as sequential steps, in that a reader needs to read not 3 million only in total for sufficient exposure to the 9,000 words, but 3 million if they have acquired the first 8,000 words, 2.4 million if they have a vocabulary of 7,000, 2 million if they have a vocabulary size of 6,000 and so forth to a total of approximately 11 million (see table 3 of Nation, 2014). McQuillan (2016a) tested whether there was a pathway of comprehensible input from smaller to larger vocabulary sizes given Nation’s (2014) conclusion of a limited availability of authentic texts. He modeled ER on juvenile fiction, consisting of some full-length novels and some chapter selections. He found that learners with 4,000- to 5,000-word vocabulary sizes would be able to read with 98% coverage books such as *Goosebumps* and *Harry Potter*. Books written for older teens such as *Twilight* reached 98% vocabulary coverage at a 6,000-word vocabulary size, and books such as *Hunger Games* at an 8,000-word vocabulary size. At 95% coverage, more books were comprehensible with smaller vocabulary sizes (e.g., *Hunger Games* was readable with 4,000 words). McQuillan (2016a) isolated series-books such as *Sweet Valley High*, finding books in the series that can be read with 4,000-, 5,000- and 6,000-word vocabularies. McQuillan (2016a) concluded there was material comprehensible in the mid-frequency range and a pathway for students from smaller vocabularies to approximately 9,000 words. This data counters Nation’s (2014) conclusion that mid-frequency graded readers are essential, indicating comprehensible authentic fiction is available between graded readers for learners with smaller vocabularies (less than 3,000-4,000 words) and more challenging texts.

McQuillan’s (2016a) then reused Nation’s (2014) computations and this is potentially problematic. He used Nation’s (2014) simulations for how many words need to be read to reach a 9,000-word vocabulary size (i.e., 3 million for an average of 12 repetitions of the 9,000-word band, 2.5 million for 8,000, 1 million for the 5,000-word band, etc.). These numbers were used to develop a model of how many books (i.e., how much available input) would be comprehensible at the different vocabulary sizes. He computed, for example, at 5,000 words, there might be approximately 1.9 million words of fiction in his corpus that could be read with that vocabulary size. For a 9,000-word vocabulary size, there might be approximately 5.4 million available words. The crucial point is that he suggested that because the number of words for the level in his modeling was greater than Nation’s (2014) recommended amount of input for meeting the level’s words 12 times, there is enough material at each vocabulary step to acquire the next 1,000 words progressively to a 9,000-word vocabulary. McQuillan (2016a) concluded “there is sufficient input each step of the way” and that with 1 hour a day “a little over three years of reading takes readers all the way to the 9,000-word family level” (p. 72).

These papers are innovative with methods, and explore important questions. However, they need to be built on. Neither Cobb (2007), Nation (2014) nor McQuillan's (2016a) models show how many, for example, 6,000-level words are available in 5,000-word 'right level' texts, or any other band. Further, it remains unclear how easy self-selection at the right level could be, as the corpora used have often not been representative of ER. McQuillan (2016a) illustrated that juvenile fiction contains texts that are comprehensible for mid-frequency vocabulary, so the strong claim that graded readers are essential does not appear correct. But, it has not been shown if this is true of general fiction, and not all L2 readers will find pleasure in fiction written for teens. By modeling on a corpus that represents a virtual library of fiction, as in the following study, we can better estimate how much comprehensible input is out there for different vocabulary sizes, and better address the core issues in this research program.

Assumptions in Corpus Modeling: CCM, Incidental Learning, Counting Units

Because modeling requires assumptions, some might reject them a-priori as simplifying complexity. They certainly simplify complexity, but should not be rejected. Assumptions are required for any computational modeling (Sterratt et al. 2011). It is simply necessary within the rich and diverse field of computer science to simplify complex behaviours to test hypotheses about the real world. There are several assumptions used in previous research and the current paper. All of these assumptions should be acknowledged, but also these assumptions are grounded in research, as in any area of modeling.

Firstly, there is experimental research supporting the CCM and the link between comprehension and vocabulary coverage from 90%–98%. However, no percentage guarantees comprehension (Schmitt et al., 2011). Using specific figures in the models published in *RFL* has been a simplifying assumption. In this study, 95% is used, which is consistent with McQuillan (2016a) and following Laufer's (2020) conclusion that this is the minimal comprehension threshold required to infer the meaning of new words. It represents 1 in 20 words unknown. The number is also debatable depending on how ER pedagogy is implemented. For example, 98% may make more pedagogical sense if dictionary look up is recommended for unknown words (Nation, 2015); but for pleasure and general understanding, Schmitt et al. (2011) show that "comprehension may not be easy when there is more than 1 unknown word in 10, [but] learners can still achieve substantial comprehension" (p. 35).

Secondly, recent debate has been had about the Word Family (WF) (specifically the 'level 6' family constituted by a lemma and its derivations) as a counting unit. McLean (2021) is critical of the WF "despite the published evidence" (p. 136) indicating not all participants understand the meaning of all derivational forms. He suggests the lemma (root wordform and inflections) or flemma (root wordform and inflections, disregarding part of speech) may be more precise for measuring vocabulary size and linking it to what a student could read. Other scholars argue the WF may be imprecise but is still a robust and valid unit for approximating lexical knowledge and coverage (Laufer et al., 2021). Laufer and Cobb (2020) concluded that 95% text coverage "reaching the lexical thresholds for reading does not require the knowledge of most of the derived words in a word family" (p. 971). For this reason and consistency with prior published models allowing for comparisons across the research record, the WF (level 6) is used in the following study.

Thirdly, no precise number of exposures guarantees incidental learning (Uchihara et al., 2019). Webb (2007) reported experiments indicating 10 repetitions supported incidental learning with students potentially having acquired "spelling, meaning, part of speech… grammatical accuracy in a sentence" (p. 62). Vidal (2011) reported gains in a listening condition for as few as 2–3 repetitions. In a meta-analysis, Uchihara et al. (2019) reported that for simple L2 form-meaning mappings, learners may need only 2-4 repetitions, but for knowledge of collocations and associations, between 10–17 repetitions may be required. Nation (2014) argued that based on the experimental research "the moderately safe goal of 12 repetitions" (p. 3) is a reasonable simplifying assumption, a number used in several corpus studies and therefore the current study. Finally, it should be noted that incidental learning is modeled as a function of repetition, but in the real world is affected by variables beyond repetition such as salience and quality of attention (Webb & Nation, 2017).

This Study

This study develops a corpus model of incidental vocabulary learning through ER that responds to the limitations of previous research. It computes how many authentic texts can be read at 95% coverage or above in the Corpus of Contemporary American English (COCA) (Davies, 2010) general fiction corpus for vocabulary sizes from 1–9,000 words. This study also addresses what starting size vocabulary might be needed before authentic material can be used in ER, and if it would be difficult to find input matched to vocabulary size through self-selection. It extracts all novels consistent with comprehensible input for different vocabulary sizes. With these sub-corpora representing reading material 'at the right level,' it computes how many occurrences of the words from the next 1,000-level frequency bands occur, and the input required to meet this vocabulary 12 times, with estimates of timeframes based on reading at 150 wpm. The research questions are:

1) How many unsimplified novels can be read at 95% or above in the COCA general fiction corpus at the different frequency bands from 1,000–9,000?
2) What does the answer to RQ1 suggest about how difficult it might be to self-select pathways through general fiction from smaller to larger vocabulary sizes?
3) Does fiction at the right level for a vocabulary size contain the words in subsequent frequency bands, and how often do they occur?
4) What does a corpus model based on the data from RQ3 suggest about the possibility of incidentally learning words and progressing from smaller to larger vocabulary sizes through general fiction reading?

Method

COCA metadata indicated two codes for fiction: 114 (general fiction, including romance, etc.) and 116 (science fiction, fantasy fiction). All samples matching these codes were extracted as the target corpus (offline version, dataset range 1990–2012). The COCA corpus contains first chapter samples of novels, argued to be representative of whole books. This was tested and confirmed in McQuillan's (2016a) study. However, the assumption likely has exceptions, as some introductory chapters can be written in a different style than the others. Using ad-hoc python code, I split the COCA into a flat structure with one novel sample per file. All chapters containing more than 1,000 words were used in the study, a methodological decision to minimize the problem that lexical diversity is negatively impacted by short texts. Chujo and Utiyama (2005) reported that samples of this size and above are most reliable in text coverage studies.

The COCA contains noise. Inspection indicated issues with word segmentation, presumably due to scanning errors, non-standard speech representation and other non-words. The issues were substantial enough to warrant correction because these problems can lead to lower estimates of word coverage. Coverage tools work by matching against fairly clean lists of words. To increase measurement precision, pre-processing involved checking corpus strings were real English words by matching them against the SCOWL (Spell Checker Oriented Word Lists) (http://wordlist.aspell.net/) and Collins's Scrabble List (2019). Any non-matching strings were stripped out (e.g., *aaparallevarleunpedazode*). Following Nation (2014), character names and proper nouns are included in the corpus modeling that follows. They are added to the cumulative total of the BNC-COCA wordlists. For example, a book that was 93% covered by the wordlists with an additional 3% coverage of names/proper nouns was designated as comprehensible input above 95% (Chujo & Utiyama, 2005; Laufer & Cobb, 2020). To create an extensive list of names/proper nouns, I combined the BNC-COCA list 31 with a lexicon containing further 79,703 names (Kantrowitz, 1994). Further pre-processing included: stripping out *'s* as possessive, converting *'ve* to *have*, *'re* to *are*, separating hyphens into two words, coding *n't* to *not*, *won't* to *will not*, *'d been* to *had been*, *'ll* to *will*. Pre-processing steps are often not very detailed in research, but have effects on word coverage accuracy. Pre-processing decisions, imperfect as they may be, are needed and I have put them on record in the interest of evaluation and reproducibility.

The above methods produced a final corpus of 33,078,808 words and 7,393 unique books. AntWord Profiler (Anthony, 2014) was used to estimate vocabulary coverage based on the 1,000–9,000 BNC-COCA word-family lists (Nation, 2014). It was first computed which COCA books were readable with 95% vocabulary for each frequency level of the BNC-COCA lists (Laufer, 2020). These were extracted, creating subcorpora 'at the right' level for different vocabulary sizes. From the subcorpora, wordlists were generated using Wordsmith v.7 (Scott, 2016) and computed for each wordlist were the occurrences of words from the subsequent levels' lists. I transformed the BNC-COCA lists to word family conversion lists usable in Wordsmith, which converted tokens to their word family headword, then deleting any words in the wordlists generated for each vocabulary size's corpora that did not match the headword. This left the words from the next bands up in the target corpus and their frequency. From these numbers the final model was built to simulate how long a student with a particular vocabulary size might need to reach 12 encounters of the words from subsequent vocabulary levels. Modeling is based on parameters used in the previous research, namely reading at 150 wpm at 60 minutes per day with a target of 12 encounters (McQuillan, 2016a). In a review of reading rates in L2 learners across 11 published papers, McQuillan and Krashen (2008) reported 150 wpm is consistent with beginner/intermediate reading rates in a second language.

In the interest of open science, all data generated underlying the corpus modeling has been made open access at the Open Science Framework (https://osf.io/4fc29/?view_only=abfd2ed07a9349ea84ef9678f58156af). The data contains the raw output and all computations for every vocabulary item and every vocabulary size from 2,000 to 9,000 words. Other researchers can check the current modeling, and develop new models based on this data, for example, at different input rates, for different vocabulary targets, or for different coverage thresholds.

Results

Table 1 reports how many books in the corpus were potentially comprehensible with 95% coverage or more by readers with different vocabulary sizes up to the 9,000-word target.

Table 1

Number of Comprehensible Books for Different Vocabulary Sizes

Vocabulary size	# Books comprehensible at 95% coverage	# Words
1,000 level	5	53,617
2,000 level	393	1,899, 629
3,000 level	1990	9,433,322
4,000 level	3470	16,677,233
5,000 level	5600	26,596,608
6,000 level	6713	30,817,348
7,000 level	7111	32,204,716
8,000 level	7270	32,716,224
9,000 level	7349	32,920,690

Table 1 indicates only 5 books could be read with a vocabulary of 1,000 words, suggesting that a starting vocabulary of 2,000 words might be needed for ER of authentic material. An interesting fact about these texts (listed in the end notes) is they are mostly first-person narratives. Possibly the lexical load of books written in this style tends to be lighter, which may be useful information when helping learners choose books. With a 2,000-word vocabulary, a student would be able to read 393 books at 95% coverage. With 7,393 books sampled, this is approximately 5%, which suggests self-selecting authentic material for a 2,000-word vocabulary is challenging but possible (i.e., one of every 20 books that a student might explore in a library would potentially be at the right level). Both McQuillan (2016a) and Krashen (2019) emphasize selection can be made easier by teacher-librarian support, so a student may not need to try 20 books to find suitable material. Notably, Table 1 indicates many books are consistent with comprehensible input in the mid-range frequency bands. Once a student has a vocabulary size of at least 3,000 words, there is plenty of comprehensible fiction available, as long as they have access to those books. At around 5,000-word families, about 75% of fiction can be read. Table 2 reports how many target words from subsequent levels occur within authentic fiction appropriate for the different vocabulary sizes.

Table 2
Reading at the Right Level: Exposure to more Advanced Vocabulary

Reading at 95% coverage	**3,000 level**	**4,000 level**	**5,000 level**	**6,000 level**	**7,000 level**	**8,000 level**	**9,000 level**
2,000-level books	947	894	847	765	692	611	574
3,000-level books	*	997	997	987	971	953	935
4,000-level books	*	*	998	998	989	982	972
5,000-level books	*	*	*	999	996	989	986
6,000-level books	*	*	*	*	998	992	990
7,000-level books	*	*	*	*	*	993	992
8,000-level books	*	*	*	*	*	*	993

* Vocabulary acquired; sufficient size to read at this level.

What Table 2 reveals is that books at the right level for any given vocabulary size contain the words in the next level, typically above 99%. More surprising is that the majority of all vocabulary from every subsequent level up to the 9,000-word goal also occurs in reading material appropriate for every vocabulary size, approximately 95–99%. As long as a student keeps reading, they will meet most of the words from all levels. For example, if a student read the 1.8 million words in the corpus representing 2,000-level books, approximately 5,330 of the target 9,000-word-vocabulary size might have been acquired, partially acquired or met at least once within comprehensible input. While this has not been shown before, it is not an adequate model of the learning potential of ER, as it ignores repetition in the input.

Table 3 reports the final corpus model replicating Nation (2014), which simulates whether comprehensible input would provide enough encounters in a reasonable amount of time to progress from one vocabulary size to the next up to the 9,000-word target. He modeled how long it would take to meet most of the words (approximately 800) from every progressive 1,000-word-families on average 12 times. Table 3 further reports how many words from all other levels might potentially be acquired by having at least 12 encounters when reading at the right level for 60 minutes a day at 150 wpm. The model works as follows. The target vocabulary sizes are normalised by occurrences per million words in the subcorpora, following standard practice with COCA (Davies, 2010). From this figure, the input (in words) needed was extrapolated, then converted to reading hours based on 150 wpm. The model is cumulative, meaning that it assumes prior learning. For example, a learner would read at the 2K level for the estimated amount of time to learn 800 words from the 3K level, and during this time would have also acquired a substantial number of 4K words. Therefore, the model computes how much additional input would be needed to reach 800 words of the 4K level, given that prior learning of 4K words already encountered 12 times. This additional time estimate is also used to compute how many more words are added from higher levels as they also reach 12 encounters. In Table 3, the first column represents books at the current vocabulary size, the next two columns show the number of words and time needed to read to reach 12 or more encounters with 800 words from the next vocabulary size, and the remaining columns report how many other words from subsequent levels also occur more than 12 times during this timeframe. For exposition, reading time estimates are given in both hours, and the equivalent years and months.

Table 3
Time/input for the Majority of Targets in the Next Vocabulary Size

Reading at 95%	**# Words needed**	**Reading @ 150 wpm**	**3K 12+**	**4K 12+**	**5K 12+**	**6K 12+**	**7K 12+**	**8K 12+**	**9K 12+**
2,000	4,559,110	507 hours; 1 year, 4.5 months	**800**	683	573	461	314	226	167
3,000	1,731,721	193 hours; 6 months		**800**	727	602	490	353	295
4,000	1,490,518	165 hours; 5 months			**800**	688	597	470	383
5,000	3,039,612	337 hours; 11 months				**800**	738	638	583
6,000	2,876,285	319 hours; 10 months					**800**	710	659
7,000	4,098,782	455 hours; 1 year, 3 months						**800**	756
8,000	3,039,443	338 hours; 11 months							**800**

The input estimates in Table 3 are larger than Nation (2014). This might be due to the current corpus being levelled for vocabulary size whereas the previous study's corpus was not and the data indicate that as the reading level goes up, there are more repetitions of advanced vocabulary. Another reason for the difference could be that Nation's (2014) modeling used an average of 12 repetitions whereas this uses a minimum of 12. Based on Table 3, in about 1 year, 4.5 months, a reader at the 2,000 level would have potentially acquired the majority of the 3,000 level. Additionally, they would have met at least 12 times 2,424 words from higher bands. Taken together, this is a possible vocabulary size of 5,224 under a year and a half, sufficient for a range of tasks (Schmitt et al., 2017). A student would only need to read at the 3,000 level for an additional 6 months to reach 12 encounters with 800 4K words, and would have added 843 words from higher frequency bands. This is growth from a 2,000-word-vocabulary size to 6,067 words in about 2 years. The modeling suggests moving from a 4,000-word-vocabulary size to 5,000 would take 5 months, 11 months from 5,000 to 6,000 words, 10 months from 6,000 to 7,000 words, 1 year, 3 months from 7,000 to 8,000 words, and 11 months from 8,000 to 9,000 words.

The model suggests 6.3 years to move from reading at the 2,000 level to having acquired enough vocabulary to read at the 9,000 level if the student reads 1 hour a day. If the student reads for 2 hours a day, it would be approximately 3 years, as estimated by McQuillan (2016a). Note the model assumes that students entering at the 2000 level the ER program have not acquired a proportion of the vocabulary from other levels, which the data indicates is unlikely given the results from higher levels. In reality, a student with a 2,000-word-vocabulary size will not be starting from zero knowledge of the 3,000 level, so the time estimate at the first level is an overestimate. The model could also be looked at as follows. Let us say a student enters at the 3,000 level, and we assume prior learning similar to the data in Table 3 (i.e., he not only has 3,000 words but also quite a few words from subsequent levels as he starts ER). In this case, to move from the 3,000 level to the 9,000 level, the model in Table 3 suggests 4.9 years at 1 hour a day, or 2.45 years at 2 hours a day. In fact, the model indicates that by the time they are reading at the 8,000 level, they could have acquired a vocabulary of 8,556 words, arguably close enough to 9,000-word target, and this progression would be achieved in 4 years at 1 hour a day (2 years at 2 hours a day).

Discussion

The findings uphold and strengthen aspects of the previous research yet do not agree entirely with them either. Schmitt et al. (2017) suggested that for ER targeting 10 encounters to support incidental learning, "not much progress with the third 1,000 words could be expected from …substantial exposure to natural (ungraded) text, at least not in the year or sometimes two that are normally available" (pp. 220-221). Like Nation (2014), they recommended graded readers because authentic self-selected ER "in most cases will NOT provide unknown words in such a friendly ratio" (Schmitt et al. 2017, p. 221). The current modeling, however, suggests that conditions for incidentally learning the words from the subsequent vocabulary sizes are quite good when reading unsimplified fiction at the right level for a 3,000-word-vocabulary size and beyond. Schmitt et al. (2017) concluded that recommending graded readers in ER is reasonable based on the research they reviewed, particularly Cobb (2007). However, they expressed concern that with simplified material, there is "no guarantee that the particular target lexis (the third 1,000-word families, or beyond) is actually present in such materials" (p. 222). The current study therefore speaks to this and shows that if learners can access a wide range of books, the target lexis are definitely present in authentic fiction.

With regard to RQ1 and RQ2, these results support McQuillan (2016b) and McQuillan and Krashen (2008) in showing many texts are available for different vocabulary sizes, which suggests it would not be difficult to self-select pathways through general fiction from smaller to larger vocabulary sizes. The evidence appears contrary to Nation's (2014) position that mid-frequency graded readers are necessary for an ER program—a conclusion drawn from his own simulations. Other than with very small starting-vocabulary sizes, authentic material exists from the 2,000- up to the 9,000-word-vocabulary sizes with 95% vocabulary coverage at each step.

The study shows that a starting vocabulary for an authentic ER program might only require 2,000 words. This is smaller than suggested by McQuillan (2016a, 2016b). Admittedly, there are fewer books that can be read with a vocabulary of only 2,000-word families, so a recommendation of a 3,000-word-starting size is reasonable, but it is still possible since according to the current data about 5% of general fiction library may be readable with about 2,000 words. With librarian and teacher support, and access to books, it is possible that 1 of every 20 books in a library could provide comprehensible input. The data indicate that once a 4,000-word-vocabulary size is reached, the challenges of finding comprehensible input are greatly reduced, with about 1 in every 2 fiction books a student finds at a library potentially being comprehensible. While the availability of comprehensible general fiction suggests mid-frequency graded readers may not be essential, Nation (2016) argues graded readers also have value if they control what the unknown words are. For example, they can be designed to increase the repetitions of vocabulary at the subsequent frequency band, which might reduce ER time estimates for acquiring a 9,000-word vocabulary.

While this study supports McQuillan (2016a, 2016b) and McQuillan and Krashen (2008) rather than Nation (2014), Cobb (2007, 2016) and Schmitt et al. (2017) regarding the availability of authentic material, the ease of self-selection, and the lack of a need for graded readers for mid-frequency vocabulary, it does not support McQuillan's (2016a) estimate that reading approximately 3 years at 1 hour a day is sufficient for a learner to reach a vocabulary size of 9,000 words. Rather, the modeling in this study suggests 6.3 years would be more accurate, and that McQuillan's (2016a) estimate might be revised to approximately 3 years if a student reads 2 hours a day rather than 1. The estimate of this paper is also not 9 years, as suggested by Nation (2016). With regard to larger debates in applied linguistics concerning explicit and incidental vocabulary pedagogy, thus far, no intervention study nor computational modeling built on published findings has indicated explicit vocabulary instruction might develop a 9,000-word vocabulary with 2 hours of instruction a day over 3 years. Such studies may be done in the future, but from this perspective, the current corpus model supports pedagogy that promotes wide-reading to learn vocabulary.

Limitations and Future Research

While the computational modeling and methodological sides of this paper arguably improve on prior research, it has still posited assumptions in the modeling. These assumptions are reasonable based on the research record, but also imperfect. Another limitation is the spacing issue, which remains to be established. It is likely not going to be 12 occurrences within a week, or 12 in sum regardless of how many years pass. The use of 12 repetitions is simply one measure based on previous studies, but of course no number is a threshold for incidental learning (Uchihara et al., 2019). If words in the final few thousand bands need fewer repetitions, which is suggested by the Uchihara et al. (2019) meta-analysis, then incidental learning pathways to the first 9,000 words become easier and timeframes reduced. Finally, students do not jump vocabulary sizes 1,000 words at a time. It is incremental and this is not captured well in the current simulations or previous studies. Readers can add a few hundred words and are then self-selecting material not at the thousand-word level but a little above. The study has used 95% vocabulary coverage, while some studies used 98%. While a researched-based decision, there is no real threshold for comprehensible input and the relationship is correlational. Another issue is the COCA corpus, which consists of first chapters. While the corpus is considered representative of the fiction construct and is the most widely used corpus in the world, in reality learners are reading whole books and there is more repetition internal to a book than across samples. A subsequent study should try to build a model based on a large corpus of full-length books suitable for each vocabulary size and compare to the current modeling.

Conclusion

Previous studies have contributed exciting and important insights by using corpora to simulate ER and the likelihood of incidentally learning vocabulary. Given the importance of this previous research, the current study addressed some limitations in previous simulations. These limitations included the use of the average and the investigation of books not at the right level for different vocabulary sizes in Nation (2014), the reuse of estimates from one study to inform time estimates in McQuillan (2016a) and the use of corpora and simulation parameters that do not represent ER well (Cobb, 2007). What the current study has shown is that practitioners can be assured that if access to books is possible, there is for every vocabulary size authentic material that can be comprehensible for students. There is a navigable pathway from smaller to larger vocabulary sizes within the mid-frequency range. As opposed to previous suggestions, ER seems particularly promising for mid-level vocabulary. On the other hand, it is not likely that 3 years could build a 9,000-word vocabulary size, if only reading for 1 hour a day. A revised estimate of 2 hours a day is suggested, and overall, the corpus modeling in this study leads to the same conclusion as Nation (2014): a largely positive picture of the potential of ER with unsimplified text for incidental vocabulary acquisition.

Note

1. The five books that reach 95% coverage at 1,000-word families are: Stephen Chbosky's *The Perks of Being a Wallflower*; Mildred D. Taylor's *The Land*; Ann Patchett's *The Patron Saint of Liars*; Kim Antieau's *Broken Moon*; and Steven Brust's *Tiassa: A Novel of Vlad Taltos*.

References

Anthony, L. (2014). *AntWordProfiler* (Version 1.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.html

Chujo, K., & Utiyama, M. (2005). Understanding the role of text length, sample size and vocabulary size in determining text coverage. *Reading in a Foreign Language, 17*(1), 1–22. https://doi.org/10125/66598

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, *11*(3), 38–63. https://doi.org/10125/44117 Cobb, T. (2016). Numbers or numerology? A Response to Nation (2014) and McQuillan (2016). *Reading in a Foreign Language*, *28*(2), 299–304. https://doi.org/10125/66904

Collins Dictionaries (2019). *Collins official scrabble words*. Collins.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–464. https://doi.org/10.1093/llc/fqq018

Hue, H. C., & Nation, P. (2000). Unknown word density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430. https://doi.org/10125/66973

Kantrowitz, M. (1994). *Name corpus: List of male, female, and pet names* (Version 1.3). Retrieved from: https://www.cs.cmu.edu/Groups/AI/util/areas/nlp/corpora/names/0.html

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal*, *73*(4), 440–464. https://doi.org/10.1111/j.1540-4781.1989.tb05325.x

Krashen, S. (2019). Do libraries and teacher librarians have the solution to the long-term English language learner problem? *Synergy*, *17*(1), 1–78. https://www.slav.vic.edu.au/index.php/Synergy/article/view/v1711910

Krashen, S., & Mason, B. (2020, May 18). The optimal input hypothesis: Not all comprehensible input is of equal value. *CATESOL Newsletter*, *53*(6). https://www.catesol.org/v_newsletters/article_151329715.htm

Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly*, *54*(4), 1076–1085. https://doi.org/10.1002/tesq.3004

Laufer, B., & Cobb, T. (2020). How much knowledge of derived words is needed for reading? *Applied Linguistics*, *41*(6), 971–998. https://doi.org/10.1093/applin/amz051

Laufer, B., Webb, S., Kim, S. K., & Yohanan, B. (2021). How well do learners know derived words in a second language? The effect of proficiency, word frequency and type of affix. *ITL-International Journal of Applied Linguistics, 172*(2)*,* 229–258. https://doi.org/10.1075/itl.20020.lau

McLean, S. (2021). The coverage comprehension model, its importance to pedagogy and research, and threats to the validity with which it is operationalized. *Reading in a Foreign Language*, *33*(1), 126–140. https://doi.org/10125/67396

McQuillan, J. (2016a). Time, texts, and teaching in vocabulary acquisition: A rebuttal to Cobb (2016). *Reading in a Foreign Language, 28*(2), 307–318. https://doi.org/10125/66905

McQuillan, J. (2016b). What can readers read after graded readers? *Reading in a Foreign Language, 28*(1), 63–78. https://doi.org/10125/66715

McQuillan, J., & Krashen, S. (2008). Commentary: Can free reading take you all the way? A response to Cobb (2007). *Language Learning & Technology*, *12*(1), 104–108. https://doi.org/10125/44133

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language*, *26*(2), 1–16. https://doi.org/10125/66881

Nation, P. (2016). Response to Tom Cobb. *Reading in a Foreign Language*, *28*(2), 305–306. https://doi.org/10125/66906

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, *50*(2), 212–226. https://doi.org/10.1017/S0261444815000075

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. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Scott, M. (2016). *WordSmith tools* (Version 7). Lexical Analysis Software Ltd. https://www.lexically.net/wordsmith/downloads/

Sterratt, D., Graham, B., Gillies, A., & Willshaw, D. (2011). *Principles of computational modelling in neuroscience*. Cambridge University Press.

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*(3), 559–599. https://doi.org/10.1111/lang.12343

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*(1), 219–258. https://doi.org/10.1111/j.1467-9922.2010.00593.x

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/applin/aml048

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

About the Author

Clarence Green is a Senior Lecturer in the School of Education of Federation University Australia. He holds a PhD in Linguistics (University of Melbourne) and a Master of Applied Linguistics. He has taught and published widely in second language acquisition, literacy, corpus linguistics, psycholinguistics, and English grammar.
Email: c.green@federation.edu.au