Reading in a Foreign Language October 2014, Volume 26, No. 2
ISSN 1539-0578 pp. 1–16

How much input do you need to learn the most frequent 9,000 words?

Paul Nation
Victoria University of Wellington
New Zealand

Abstract

This study looks at how much input is needed to gain enough repetition of the 1st 9,000 words of English for learning to occur. It uses corpora of various sizes and composition to see how many tokens of input would be needed to gain at least twelve repetitions and to meet most of the words at eight of the nine 1000 word family levels. Corpus sizes of just under 200,000 tokens and 3 million tokens provide an average of at least 12 repetitions at the 2nd 1,000 word level and the 9th 1,000 word level respectively. In terms of novels, this equates to two to twenty-five novels (at 120,000 tokens per novel). Allowing for learning rates of around 1,000 word families a year, these are manageable amounts of input. Freely available *Mid-frequency Readers* have been created to provide the suitable kind of input needed.

Keywords: extensive reading, vocabulary learning, repetition, text coverage, input

Although it was long assumed that native speakers increase their vocabulary size largely through the strategy of guessing from context rather than from directly being taught vocabulary, it was only relatively recently (Nagy, Herman, & Anderson, 1985) that there was strong experimental evidence that guessing from context was effective and resulted in vocabulary learning. Surprisingly, the Book Flood studies (Elley & Mangubhai, 1981a, 1981b) with foreign language learners did not use measures of vocabulary size or growth to show the effect of reading on vocabulary knowledge. However, since the early work of West (1955) and later as a result of Krashen's (1985) influential input hypothesis, there has been a strong and growing movement to encourage the use of extensive reading programs for foreign language development (Day & Bamford, 1998; Day & Bamford, 2004; Waring, 2001). However, with one notable exception (Cobb, 2007) there has been no corpus-based study of the feasibility of learning large amounts of foreign language vocabulary through reading. Although reference is made to first language (L1) learning as evidence for the role of reading, there has also been no L1 study which has taken a corpus-based approach to looking at opportunities for vocabulary learning through reading or listening.

Is it possible to learn enough vocabulary just through reading?

There has been a debate, with Cobb (2007, 2008) on one side and McQuillan and Krashen (2008)

on the other, over whether it is possible to learn enough vocabulary solely through reading input. Cobb argued that given the difficulty of the material and the time available, learners could not get through enough reading to meet the words at each level, let alone have enough repetitions to learn them. McQuillan and Krashen argued that it is possible, but the major point of disagreement for them involved the difficulty of the reading material.

McQuillan and Krashen assumed that learners would be able to read a wide range of texts with relative ease and speed. Cobb argued that the difficulty of the texts with their heavy load of unfamiliar vocabulary would make reading very slow and laboured. There were thus two aspects to their disagreement: (a) the heavy vocabulary load of unsimplified text, and (b) the quantity of input needed to repeatedly meet target words.

The first part of the present study temporarily puts aside the vocabulary load issue, and looks solely at the quantity of input needed. So, at first, this article largely ignores the enormous vocabulary load placed on learners when they read and listen to unsimplified texts. It should be noted however that the vocabulary load issue is a very important one that needs to be properly addressed (Hu & Nation, 2000; Nation, 2009; Nation & Deweerdt, 2001; Schmitt, Jiang, & Grabe, 2011) and will be taken up again later in this article.

The focus of the present study is on the 1st 9,000 words, and because research has shown that the 1st 9000 word families plus proper nouns provide coverage of over 98% of the running words in a wide range of texts (Nation, 2006), a vocabulary size of 9,000 words or more is a sensible long-term goal for unassisted reading of unsimplified texts. Schmitt and Schmitt (2012) also suggested applying the term *mid-frequency vocabulary* to the 6,000 word families making up the 4th 1000 to 9th 1000 words, because these along with the 3,000 high frequency words of English and proper nouns provide 98% coverage of most texts.

An essential condition for learning is repetition, and so learners not only need to gradually meet the most frequent 9,000 word families, but they have to meet them often enough to have a chance of learning them.

Repetition and vocabulary learning

There is clearly a relationship between repetition and vocabulary learning (Elley, 1989; Laufer & Rozovski-Roitblat, 2011; Pellicer-Sanchez & Schmitt, 2010; Stahl & Fairbanks, 1986). The amount of repetition of words typically correlates with the chance of them being learned at around .45 (Saragi, Nation, & Meister, 1978; Vidal, 2011) and is the major factor affecting vocabulary learning from reading (Vidal, 2011). Even though repetition is a very important factor, it is still only one of many factors, and as a result there is no particular minimum number of repetitions that ensures learning. For reading, Vidal (2011) found the greatest increase in learning between two and three repetitions. Webb (2007a, 2007b) found at least 10 repetitions were needed to develop something approaching rich knowledge, but Webb used 10 different tests for each word measuring orthography, association, grammatical functions, syntax, and meaning and form, both receptively and productively, thus requiring a fairly high standard of knowledge.

Waring and Takaki (2003) found that at least eight repetitions of a word in a graded reader were needed to have a 50% chance of remembering the word three months later. Recognition after three months is a tough measure, and the scores on the immediate posttest were higher. In this study, the moderately safe goal of 12 repetitions is taken as the minimum. This fits with Vidal (2011) and Webb (2007a, 2007b), but according to Waring and Takaki (2003) and Brown, Waring, and Donkaewbua (2008) may be a bit too few. Twelve repetitions, however, are enough to allow the opportunity for several dictionary look-ups, several unassisted retrievals, and an opportunity to meet each word in a wide variety of contexts. It is also hoped that learning through written input will be supported to some degree by learning through spoken input, learning through output, deliberate learning, and fluency development, and so a high standard of learning from input alone is not necessary. Setting too low a number of repetitions would bias the study towards favouring the input position.

Because one aim of this study is to resolve the McQuillan and Krashen versus Cobb debate, the main focus is on reading. However, because input for incidental learning can be of many kinds, the study also looks at what kind of input provides the best opportunities for meeting the most frequent 9,000 word families. What kind of reading material provides the best opportunities? Is reading material better than spoken input? Is a mixture of input preferable?

This study attempts to answer the following research questions:

1. How much input do learners need in order to meet the most frequent 9,000 word families of English enough times to have a chance of learning them?
2. Can learners cope with the amount of input?
3. What kinds of input provide the greatest chance of meeting most of the most frequent 9,000 word families?

Method

The present study uses word family lists created from the British National Corpus and the Corpus of Contemporary American English (COCA) to represent learners' vocabulary sizes. These lists each consist of 1,000 word families and the various lists are ordered according to the frequency and range of the words. So, the 1st 1000 word family list contains the 1,000 most frequent and widely used words. The words in each list are in word families. Here are two example families from the 2nd 1000 list.

(a) ACCESS:
- ACCESSED
- ACCESSES
- ACCESSIBILITY
- ACCESSIBLE
- ACCESSING
- INACCESSIBILITY
- INACCESSIBLE

(b) ACCIDENT:
- ACCIDENTAL
- ACCIDENTALLY
- ACCIDENTLY
- ACCIDENTS

Note that all the words in a family share the same free-standing stem, but can be different parts of speech. Because the focus is on receptive knowledge, the word family (Bauer & Nation, 1993) is the most appropriate unit of analysis. Using word families assumes that when the learner knows at least one member of the family, the other members are accessible through the application of word building rules, or what Anglin (1993) calls morphological problem-solving. The arguments in favour of the use of word families are as follows.

(a) Word families are psychologically real (Bertram, Laine, & Virkkala, 2000), meaning that users of English treat members of a family as belonging to that family.
(b) It is much more sensible than assuming that different parts of speech such as *walk* as a noun and *walk* as a verb, inflected forms like *family* and *families*, or derived forms like *separate*, *separately*, and *separateness* are different words, each requiring separate unrelated learning for the purposes of reading or listening.
(c) Seeing words as members of word families increases the repetitions of words, as the occurrence of a family member is likely to strengthen knowledge of other members of the family (Nagy, Anderson, Schommer, Scott, & Stallman, 1989).

There are arguments against word families.

(a) Learners' word building knowledge and skills change as their knowledge of the language develops, so a flexible description is needed of what is included in a word family. In this study, a conservative description of the word family is used (Bauer & Nation, 1993) where the headword of a family must be a free form (a word in its own right) and only transparent, frequent, regular, and largely productive affixes are allowed.
(b) Some learners have a poor knowledge of English morphology (Schmitt & Zimmerman, 2002) and so some family members may not be obvious members of the word family for them.
(c) Computer-based text analysis programs like Range cannot deal with polysemy, homography, and homonymy, meaning that some families like *bank* as in "the bank of a river" and *bank* as in "the bank that takes your money" are not distinguished. This problem also exists for lemmas and word types (see Schmitt [2010, pp. 189–193] for further discussion of word families).

Nonetheless, the word family is the most suitable unit of analysis for receptive purposes and particularly so where the words are met in context, and so it is used in this study. At present there are twenty-five 1,000 word family lists and they are freely available with the Range program from Paul Nation's web site http://www.victoria.ac.nz/lals/staff/paul-nation.aspx.

The data in the study was processed using the Range program (Heatley, Nation, & Coxhead, 2004). This program compares texts to word lists and provides frequency and range information about how often and in how many texts each word family occurs.

The texts used in the study included transcriptions of informal spoken language, scripts of movies and TV shows, novels, academic texts, and popular journal type articles. These specially constructed corpora each one million tokens long were used to cover British and American

English, written and spoken English, and formal and informal language. The corpora came from the following sources – the 25 novels came from Project Gutenberg (http://promo.net/pg/), a one million word novels section was taken from the British National Corpus (BNC World Edition, 2000, www.hcu.ox.ac.uk/BNC), the movies and TV corpus came from movie scripts gathered from the internet, the spoken corpora were taken from the BNC demographic section (unscripted speech) and from the American National Corpus, the journal section came from the American National Corpus, and the academic corpus came from Coxhead's (2000) academic corpus. The sections of the American National Corpus were kindly provided by the editor of the corpus.

Results

Research question 1: How much input do learners need in order to meet the most frequent 9,000 word families of English enough times to have a chance of learning them?

Table 1 uses a corpus of novels to see how many running words (tokens) would have to be read to meet most of the words at a particular 1,000 word family level on average twelve times. To get the different sized corpora shown in column 2 of Table 1, novels were gradually added to the corpus as shown in column 5.

Table 1. *Corpus sizes needed to gain an average of at least twelve repetitions at each of nine 1,000 word levels using a corpus of novels*

1,000 word list level	Corpus size to get an average of at least 12 repetitions at this 1,000 word level (repetitions)	Number of 1 timers / 2 timers out of 1,000)	Number of families met	Number of novels
2nd 1,000 families	171,411 (13.4)	84/99	805 of 2nd 1,000	2
3rd 1,000 families	300,219 (12.6)	83/73	830 of 3rd 1,000	3
4th 1,000 families	534,697 (12.6)	93/73	812 of 4th, 1,000	6
5th 1,000 families	1,061,382 (13.7)	101/79	807 of 5th, 1,000	9
6th 1,000 families	1,450,068 (13.1)	89/82	795 of 6th, 1,000	13
7th 1,000 families	2,035,809 (13.7)	92/63	766 of 7th, 1,000	16
8th 1,000 families	2,427,807 (14.1)	96/70	755 of 8th, 1,000	20
9th 1,000 families	2,956,908 (12.0)	88/78	805 of 9th, 1,000	25

The 25 novels used were *Adam Bede, Alice in Wonderland, Animal Farm, Babbit, Born in Exile, Captain Blood, Castle Rackrent, Cranford, Emma, Far from the Madding Crowd, Glimpses of the Moon, Great Gatsby, Lady Chatterley's Lover, Lord Jim, Main Street, Master of Ballantrae, Middlemarch, More William, Right Ho Jeeves, Scaramouche, Tono Bungay, Turn of the Screw, Ulysses, Walden, Water Babies*. They were all from Project Gutenberg.

In Table 1 we can see that for the 3rd 1,000 word families, learners would need to read just over 300,000 running words in order to meet most of the 3rd 1,000 word families an average of 12.6 times. If they read 300,219 running words they would not meet all of the 3rd 1,000 word families but would meet 830 of them. Note that averages are used, and as columns 3 and 4 show, this

certainly does not mean that all words at this level are repeated at least twelve times.

In Table 1, figures are not given for the 1st 1000 because the average frequency figures would be very misleading given the very high frequencies of the high frequency function words. So, to meet most of the 2nd 1000 families, a corpus size of 171,411 tokens would be needed. This would provide an average of 13.4 repetitions for words at the 2nd 1000 level. To meet most of the 5th 1000 learners would need to read a million tokens.

The number of novels in column 5 of Table 1 is a very rough estimate to project from as the novels used in the study varied in length from 9,000 tokens (*Alice in Wonderland*) to 323,599 tokens (*Middlemarch*), the average being 118,276 tokens.

When averages are used, they should be accompanied by some measure like standard deviations, but standard deviations assume a normal distribution. The standard deviations for the averages in Table 1 are roughly the same as the averages and are thus not very useful. I have used the number of one timers and two timers (words occurring only once or twice) in column 3 because these better reflect the weakness of averages. Just under 10% of the word families occurred only once or twice. When this is combined with the families that did not occur (around 200 at each level, see column 4), it underlines how rough these estimates are. However, the nature of word frequency distributions in natural language, as shown by Zipf's law (Sorrell, 2012), makes such a result unavoidable. These unmet words and one-timers or two-timers will occur and be repeated in other texts. For example in the 13 texts used to measure the 6th 1000 level, there are only 30 one-timers of the 3rd 1000 word families, and 28 two-timers, compared with 83 and 73 in Table 1. Only 63 of the 3rd 1000 families did not occur in the 13 texts compared to 170 in the three texts in Table 1.

Table 2 provides further data to support the idea that reading at later levels adds to meetings at more frequent levels. Table 2 shows a full set of data from the 3 million word corpus of novels. This corpus was used only to obtain the figures for the 9th 1000 level. Note however that the words at the other frequency levels would also be met a lot when reading the 25 novels (see column 4 in Table 2). Consider also that having a diverse corpus rather than just novels would result in most of the 9,000 word families being met.

Table 2. *Number of tokens and word families occurring at each of nine different 1000 word family level in a 3 million token corpus of novels*

1000 word list level	Number of tokens	Number of word families occurring	Average frequency per word family
2nd 1,000 families	170,279	994	171
3rd 1,000 families	77,148	972	79
4th 1,000 families	54,739	945	58
5th 1,000 families	34,188	929	37
6th 1,000 families	24,038	904	27
7th 1,000 families	16,436	857	19
8th 1,000 families	13,346	817	16
9th 1,000 families	9,669	805	12
Total	2,956,908 whole corpus	8,219 out of 9,000	

Research question 2: Is this amount of input possible?

In order to read without unknown vocabulary becoming too much of a burden, no more than 2% of the running words should be beyond the learners' knowledge (Hu & Nation, 2000; Schmitt, Jiang, & Grabe, 2011). This means that on average there would be just under 50 words of context around each unfamiliar word, which would allow guessing from context.

Native speakers of English appear to increase their vocabulary at the rate of around 1000 word families per year (Biemiller & Boote, 2006; Goulden, Nation, & Read, 1990), with a typical educated native speaker vocabulary size being around 20,000 words. If we expect second language learners to increase their vocabulary at around the same yearly rate, then they will need to increase the amount they read each year, starting for the 2nd 1000 word level at under 200,000 tokens and rising to 3,000,000 tokens a year for the 9th 1000 level. This may be asking too much, as there is no published research to support this figure for learners of English as a foreign language. However, it is an optimistic goal to aim for. With this optimistic aim, let us assume that each row in Table 1 represents a year's reading.

Would learners be able to read the amounts shown in column 2 of Table 1 if the material were at the right level for them so that the target words would make up around 2% or less of the running words in the text and if the words beyond the target level were largely replaced? Table 3 converts the token figures into minutes of reading per school week. The calculation of the time in column 3 of Table 3 assumes that a learner reads at a speed of 200 words per minute for 40 weeks of the year. Chung and Nation (2006) and Tran (2012) show that university level learners of English as a foreign language completing a speed reading course can read easy texts at average rates of 200 words per minute with many easily exceeding this rate.

Table 3. *Amount of reading in tokens and time per week to meet the 1000 word families around 12 times*

1000 word list level	Amount to read (tokens)	Minutes per week @ 200 wpm
2nd 1000	171,411	21 minutes
3rd 1000	300,219	38 minutes
4th 1000	534,697	1 hour 5 minutes
5th 1000	1,061,382	2 hours 12 minutes
6th 1000	1,450,068	3 hours
7th 1000	2,035,809	4 hours 5 minutes
8th 1000	2,427,807	5 hours 3 minutes
9th 1000	2,956,908	6 hours 10 minutes

If learners read a total of 3 million tokens, then they would meet the 1st 9,000 words often enough to have a chance of learning them. However, Table 1 shows that if you see the learning of vocabulary through reading as a set of staged steps, then after learners know the 1st 2000 words, the next step to learn the 3rd 1000 words would involve reading around an additional 300,000 tokens (Table 3, column 2, row 3). The next step, to learn the 4th 1000 words, would require reading another half million tokens, and after that reading an additional 1 million tokens

to learn the 5th 1000 words.

If we assume that each 1000 word family step takes one year, then by the time the learners reached the 6th 1000 they need to read one and a half million tokens in that year, and 2 million tokens in the next.

These are manageable amounts of reading in terms of the time needed. Column 3 in Table 3 assumes that learners read at a rate of 200 words per minute. If learners read for 40 weeks of the year, five times a week, then at the speed of 200 words per minute, learners expanding their vocabulary at the 3rd 1000 level will need to read for thirty-eight minutes per week, or eight minutes a day five days a week. This is easily achievable, even without home work but just using a set silent reading time in class.

Given the approximate nature of the calculations, there is some justification in rounding off the amount-to-read figures to make them more memorable. Table 4 rounds the figures to the nearest 100,000 running words, and provides figures for a slower reading speed of 150 words per minute. Table 4 provides not only weekly time requirements, but also daily (five days a week) time requirements.

Table 4. *Amount of reading input and time needed to learn each of the most frequent nine 1000 word families*

1000 word list level	Amount to read	Time needed for reading per week (per day) at a reading speed of 150 words per minute
2nd 1000	200, 000	33 minutes (7 minutes per day)
3rd 1000	300,000	50 minutes (10 minutes per day)
4th 1000	500,000	1 hour 23 minutes (17 minutes per day)
5th 1000	1,000,000	2 hours 47 minutes (33 minutes per day)
6th 1000	1,500,000	4 hours 10 minutes (50 minutes per day)
7th 1000	2,000,000	5 hours 33 minutes (1 hour 7 minutes per day)
8th 1000	2,500,000	6 hours 57 minutes (1 hour 23 minutes per day)
9th 1000	3,000,000	8 hours 20 minutes (1 hour 40 minutes per day)

Note. The per week figure is based on forty weeks, and the daily rate is based on 5 days.

Table 4 shows that from the 4th 1000 level on, the increase required in the amount of reading is 500,000 words per year. From the 7th 1000 level on, over an hour a day five days a week, forty weeks of the year would need to be devoted to reading. This is a lot, but it assumes that this quantity of input is coming only through reading. Spoken sources are of course possible but these provide less intensive input. It takes around two hours to watch a typical 10,000 token movie (a rate of around 83 words per minute, or just over half of a reading rate of 150 words per minute). Nonetheless, an hour to an hour and forty minutes five times a week is possible.

Reading texts at the right level. These figures all assume that learners are reading at the right level where no more than 2% of the tokens are unfamiliar. Even with excellent free computer-based support programs like Read with resources (www.lextutor.ca), the look-up and synonyms functions in the right-click menu of Microsoft Word, and the touch to look up function in Kindle, most text beyond the 3000 word level of graded readers series is very difficult for foreign

language learners. This is because in most novels a very large number of different words occur beyond the learners' current vocabulary knowledge. Added to this is the very low number of repetitions of most of these words in any particular novel. About half of them would occur only once.

Table 5 provides detailed data from one novel, *Captain Blood*, in order to see how many unknown words there are and how often these words are repeated. *Captain Blood* was chosen because it is close to the average length of the 25 novels used in this study at 115,879 words long. If a reader knew 9,000 word families already, then 2.06% of the words (tokens) would be unknown. However, if the learner had a vocabulary of 5000 words, then 4.5% of the tokens would be unknown, many occurring only once. Allowing around 300 words per page this would mean that there would be over 13 unknown words on every page. This is manageable in terms of unknown vocabulary load if the learner could consult a dictionary, but there would be a total of 2,019 unknown words in the whole book with half of them beyond the 9th 1000 word level. 663 of the 1,047 words beyond the 9th 1000 level occur only once.

Table 5. *Number of different word families at various frequency levels in Captain Blood*

Word level	Cumulative coverage	Number of families
1st 1000	81.54%	913
2nd 1000	88.17%	752
3rd 1000	91.63%	591
4th 1000	94.17%	467
5th 1000	95.50%	384
6th 1000	96.43%	310
7th 1000	97.06%	247
8th 1000	97.52%	206
9th 1000	97.94%	189
		4059

Note. The % coverage figures (but not the number of families figures) include proper nouns, transparent compounds (*seaman, mainmast, bloodthirsty*) and marginal words like *oh*, *ah*, *ha!* (3.29%).

Here are some of the frequent words from the 10th 1000 level: *broadside, dyke, deliverance, frigate, kinsman, mirth, tawny, haughty, vindictive, yeoman, archipelago, chagrin, headland*. Of the 159 word families at this level, 71 occurred only once. This low occurrence is typical of words at the lower frequency levels.

Unsimplified text clearly provides poor conditions for reading and incidental vocabulary learning for learners whose vocabulary sizes are less than 9,000 word families.

Supporting reading. Graded readers provide suitable reading material up to vocabulary sizes of 3,000 word families. There is now free adapted reading material available, *Mid-frequency Readers* (see Paul Nation's web site http://www.victoria.ac.nz/lals/staff/paul-nation.aspx), that can provide at least 98% coverage of the tokens for learners with vocabulary sizes larger than 3000 word families. Each reader is available in three versions, at the 4000, 6000, and 8000 levels.

However, although 98% is a useful minimum coverage figure, when looking at the simplification of text with one of the goals being vocabulary learning, we need to give priority to the actual number of unknown words because we do not want the guessing and look-up load to be too heavy. For example, in *Lord Jim* (see Table 6), 1.52% of the running words are low frequency words (10th 1000 on). This is less than 2% and would seem to be a manageable coverage. However, this works out to be 1,089 words, which are far too many to expect learners to deal with in one novel. A more sensible goal would be around 300-400 words. This is why the coverage percentages in the far right column (look-up) of Table 5 are well under 2% with most under 1%. Table 6 shows that not a lot of adaptation is needed to provide such material, but it is essential if learners are to read under the most favourable conditions for vocabulary learning through input.

In Table 6, the three percentage figures in each row in the three left-hand columns add up to 100%. The first column gives what we assume learners know, so a learner who already knows 5000 words will have 96.22% coverage of *Lord Jim*. The 3.78% unknown words (columns 2 and 3) are made up of 1.52% low frequency words (1,133 tokens) and 2.26% mid-frequency words (6000-9000 words inclusive). For such a learner who knows 5000 words, the target words for learning would be the 6th 1000 which make up 0.83% of the tokens of the novel (see column 5) and these would be the words to be guessed or looked up (358 different words in *Lord Jim*). That means that the words from the 7th 1000 on would need to be replaced (2.95% of the novel being 1,515 word families and 3,410 tokens). In Table 6, proper nouns and transparent compounds are included in the coverage figures in column 1. In each row, the words in columns 2 and 3 (mid- and low frequency levels) add up to the same number as the words in columns 4 and 5 (ways of dealing with the words) and are of course the same words.

Table 6. *Coverage and treatment of mid- and low-frequency word families in a typical unsimplified text (Lord Jim) given various vocabulary sizes*

Words assumed known (cumulative % coverage)	Remaining mid-frequency (% coverage)	Low frequency (% coverage)	Replacement	Look-up
3000 (92.69%)	5.79%	1.52%	5th 1000 on (5.05%)	2.26% (552)
4000 (94.95%)	3.53%	1.52%	6th 1000 on (3.78%)	1.27% (444)
5000 (96.22%)	2.26%	1.52%	7th 1000 on (2.95%)	0.83% (358)
6000 (97.05%)	1.43%	1.52%	8th 1000 on (2.36%)	0.59% (302)
7000 (97.64%)	0.84%	1.52%	9th 1000 on (1.89%)	0.47% (239)
8000 (98.11%)	0.37%	1.52%	Low frequency (1.52%)	0.37% (230)
9000 (98.48%)	0.00%	1.52%	12th 1000 on (1.08%)	0.44% (295)

Only twelve of the low frequency words (words beyond the 9th 1000) have ten or more repetitions in *Lord Jim* – for example, *rajah* (55), *stockade* (34), *schooner* (28), *infernal* (18), and these would not be replaced in an adaptation of the novel.

The replacements in column 4 of Table 6 are not difficult to do, particularly with the new version of AntWordProfiler (Anthony, 2012) which allows direct editing into a marked up text along with access to a thesaurus. Using this freely available software, a *Mid-frequency Reader* can be produced in a few hours.

Problems with the calculations. There are some serious problems with the crude calculations used in this study. First, they assume that the input is comprehensible so that learners can learn from it. We know however that learners need a vocabulary size of around 7000 or 8000 words before unsimplified written input is likely to be comprehensible without outside support such as a dictionary (Nation, 2006). Similarly, learners need a vocabulary size of around 6000 words before movies become comprehensible input, although Webb and Rodgers (2009) argue that 3000 words may be sufficient. However, the more vocabulary known, the better comprehension is likely to be (Schmitt, Jiang, & Grabe, 2011).

Second, if texts were written at the right level, the repetitions would increase slightly, because the words beyond the level would be replaced by known words or target words. The increase in repetitions however would only be small, because the words being replaced would be low-frequency words of which many would be one-timers.

Third, there is also the problem of actual repetitions (not averages) of the target words. Average repetitions for each 1000 word level have been used, but there is a wide range of repetitions at each level, and these repetitions are not evenly distributed among the word families at each level but to a small degree follow Zipf's law which means that some words may well occur only once or twice. In the total data for the novels used to produce the figures in Table 1, there are 89 words at the 6th 1000 level that occur only once, and 82 that occur only twice. At the 4th 1000 level, there are 93 words that occur only once, and 73 that occur twice. Fortunately these numbers are not large enough to have a major effect on the overall learning goal of learning the vocabulary at a particular 1000 word level. They are partly compensated for in any particular novel by the topic-related words beyond that level that occur ten times or more and are not replaced in the adaptation. For example, in the 6,000 word level *Mid-frequency Reader* adaptation of *Glimpses of Unfamiliar Japan* by Lafcadio Hearn, there are 13 word families occurring 10 times or more beyond the 8th 1000 word level.

Fourth, there is the problem of the spacing of the repetitions. Many words gather enough repetitions by occurring in a range of novels not just in one novel. This means that the spacing between repetitions may be quite large particularly as learners move through the later mid-frequency word levels. If the spacing is too large, memory for the previous meeting may disappear before the word is met again.

A positive view. There is however a positive side to the calculations. First, reading at a later level will provide plenty of repetitions for vocabulary met at the earlier levels, so that the reading at a particular level, say the 5th 1000 word level, will not only have the effect of helping the students learn the target words at that level but will also strengthen knowledge of words at the 4th 1000, 3rd 1000 and other levels. What is missed early on can be picked up later. Second, the amount of input required per year is feasible. Third, there is a very large quantity of excellent material within a controlled vocabulary to help learn the very important high frequency words of the language up to the 3,800 word level, and the *Mid-frequency Readers* (Nation & Anthony, 2013) now extend this to the 8000 word level. Fourth, there are very rapid advances in the kind of support that computer technology can give to learning from reading. A very good example of this is Tom Cobb's *Read with resources* on his Compleat Lexical Tutor web site (www.lextutor.ca).

This program provides easy dictionary look-up, pronunciation, and concordance examples for words met in the text. In Microsoft Word, there is a look-up option for dictionary access in the right-click menu, and tablet computers and electronic books usually allow easy electronic access to a dictionary.

Laufer (2003) questions whether learners of English as a foreign language in fact learn much vocabulary through reading. Certainly, the reluctance of many teachers to incorporate extensive reading programs in their language courses supports Laufer's scepticism, and thus the opportunities to learn vocabulary through extensive reading are in many places very underutilised. This present piece of research however shows that with the right material such learning is feasible. It is also important to realise that learning vocabulary through extensive reading is just one of a range of opportunities for vocabulary learning, although it can be one of the most effective and enjoyable opportunities.

Research question 3: What kinds of input allow the most words to be met?

So far, we have only looked at a corpus of novels, but there are many other kinds of input than novels. Would other kinds of spoken or written input provide better opportunities for meeting the most frequent 9,000 word families? To answer this question, let us look at a range of different corpora each exactly 2,000,000 running words long.

A diverse corpus is one that is made up of texts from different genres and topic areas. A major distinction between corpora is the spoken/written distinction (Biber & Conrad, 2009). A homogeneous corpus consists of texts which are similar because they are all spoken or all written, or they make up a particular kind of writing, such as novels or academic texts.

If learners want to meet as many different words as possible, should they stick to similar texts or should they get input from a wide variety of different kinds of texts? Table 7 compares the number of word families from the 1st 9,000 word families of English occurring in fourteen different corpora, each 2 million tokens long. In order to make a sensible comparison, the same sub-corpora are used where appropriate in the various corpora. So, *Spoken BNC* refers to material taken from the British National Corpus demographic section which consists of unscripted spoken material. The same 1 million token part of the British National Corpus demographic is used wherever the term *Spoken BNC* is used in Table 7 with additional material from the British National Corpus demographic section where the whole 2 million word corpus is spoken material. Similarly, the same 1 million token collection of novels is used, and the same 1 million token collection of journal material from the American National Corpus. The journal section is made up of public letters such as those requesting donations, letters to the editor, and news summaries from Ezines like *Slate*.

Table 7. *Number of words from the 1st 9,000 word families occurring in different corpora each exactly 2 million tokens long*

Corpus	Number of word families from the 1st 9,000 occurring in the corpora
Journals ANC + Novels BNC	8,631
Journals ANC only	8,603
Journals ANC + Movies	8,512
Journals ANC + Spoken BNC	8,410
Novels BNC + Academic	8,328
Movies + Novels BNC Table 1	8,276
Movies + Academic	8,006
Novels only	8,005
Spoken BNC + Novels BNC	7,828
Spoken BNC + Movies	7,631
Spoken BNC + Academic	7,584
Academic only	7,500
Spoken BNC only	6,457

There are problems in such a comparison, because there are several factors that affect the richness of vocabulary in a corpus. The spoken/written distinction is clearly an important one (Biber & Conrad, 2009; Shin, 2007). The diversity of topics covered is another very influential factor, with a diverse corpus having a much richer vocabulary than a more homogeneous corpus (Sutarsyah, Nation, & Kennedy, 1994). The degree of formality of the text is likely to be important too, with less formal text (such as letters or friendly conversations) having a less rich vocabulary. In comparisons of various corpora, it is extremely difficult to control for this diversity of variables.

It is also necessary to bear in mind that most novels contain a mixture of narrative and dialogue and they should be seen as not truly representing written text in the same way that academic text does.

As Table 7 shows, a mixed written and spoken corpus provides better opportunities to meet most of the 1st 9,000 word families of English. Six of the top eight two million word corpora in Table 7 contain a mixture of spoken and written texts, if we consider novels to include some spoken text. The top five all contain a journals sub-corpus. It may be the diversity of topics in the journals corpus that results in the richness of vocabulary. Except for the journals only corpus, all the other homogeneous corpora are low in the table. Academic text does not provide the highest inclusion of the top 9,000 word families probably because many technical words are beyond the 9th 1000 level. Spoken corpora alone provide the lowest inclusion. Using data not shown in Table 7, a BNC spoken only, an ANC spoken only and an ANC plus BNC spoken mix all provided similarly low inclusion. The best advice to learners for vocabulary inclusion might be to read lots of magazines, newspapers and novels, and watch plenty of movies.

It is important that this largely positive study of the opportunities for learning through input is not taken as an argument against deliberate vocabulary learning. A well-balanced vocabulary

learning program contains both message-focused opportunities for learning as well as an appropriate amount of deliberate vocabulary learning (Nation, 2007). The major values of the present study are to show that learning through input is feasible for learners of English as a foreign language if texts at the appropriate level are available, and to show what amounts of reading would need to be done to learn the words at each 1000 word level. The material is now available for such learning, and teachers and learners need to take up the challenge**.**

References

Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development Serial No. 238, 58* (10 Serial No. 238), 1–165.

Anthony, L. (2012). *AntWordProfiler [Computer software] Version1.3.1*. Tokyo: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/.

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

Bertram, R., Laine, M., & Virkkala, M. (2000). The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology*, *41*, 287–296. doi: 10.1111/1467-9450.00201

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in the primary grades. *Journal of Educational Psychology*, *98*, 44–62.

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*, 136–163.

Chung, M., & Nation, I. S. P. (2006). The effect of a speed reading course. *English Teaching*, *61*, 181–204.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, *11*(3), 38–63.

Cobb, T. (2008). Commentary: Response to McQuillan and Krashen. *Language Learning & Technology*, *12*(1), 109–114.

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

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

Day, R. R., & Bamford, J. (2004). *Extensive reading activities for teaching language*. Cambridge: Cambridge University Press.

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly*, *24*, 174–187.

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

Elley, W. B., & Mangubhai, F. (1981b). The long-term effects of a book flood on children's language growth. *Directions*, *7*, 15–24.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, *11*, 341–363.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2004). *Range [Computer software] Version 1.32*. Wellington: Victoria University of Wellington. Available from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403–430.

Krashen, S. (1985). *The Input hypothesis: Issues and implications*. London: Longman.

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review*, *59*, 567–587.

Laufer, B., & Rozovski-Roitblat, B. (2011). Incidental vocabulary acquisition: The effects of task type, word occurrence and their combination. *Language Teaching Research*, *15*, 391–411. doi: 10.1177/1362168811412019

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.

Nagy, W. E., Anderson, R., Schommer, M., Scott, J. A., & Stallman, A. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly*, *24*, 263–282.

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

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

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching*, *1*, 1–12. doi: 10.2167/illt039.0

Nation, I. S. P. (2009). New roles for L2 vocabulary? In W. Li & V. J. Cook (Eds.), *Contemporary Applied Linguistics Volume 1: Language Teaching and Learning* (pp. 99–116). London: Continuum.

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

Nation, I. S. P., & Deweerdt, J. (2001). A defence of simplification. *Prospect*, *16*(3), 55–67.

Pellicer-Sanchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do *Things Fall Apart*? *Reading in a Foreign Language*, *22*, 31–55.

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*, 72–78.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke: Palgrave Macmillan.

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

Schmitt, N., & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 1–20. doi:10.1017/S0261444812000018

Schmitt, N., & Zimmerman, C. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*, 145–171. doi: 10.2307/3588328

Shin, D. (2007). The high frequency collocations of spoken and written English. *English Teaching*, *62*, 199–218.

Sorrell, C. J. (2012). Zipf's law and vocabulary. In C. A. Chapelle (Ed.), *Encyclopaedia of Applied Linguistics*. Oxford: Wiley-Blackwell.

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based

meta-analysis. *Review of Educational Research*, *56*, 72–110.
Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based study. *RELC Journal*, *25*, 34–50.
Tran, Y. T. N. (2012). The effects of a speed reading course and speed transfer to other types of texts. *RELC Journal*, *43*, 23–37. doi: 10.1177/0033688212439996
Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*, 219–258. doi: 10.1111/j.1467-9922.2010.00593.x
Waring, R. (2001). Research in extensive reading. *Studies in Foreign Language and Literature (Notre Dame Seishin University, Kiyo)*, *25*, 1–25.
Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*, 130–163.
Webb, S. (2007a). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research*, *11*, 63–81. doi: 10.1177/1362168806072463
Webb, S. (2007b). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*, 46–65. doi: 10.1093/applin/aml048
Webb, S., & Rodgers, M. P. H. (2009). The lexical coverage of movies. *Applied Linguistics*, *30*, 407–427. doi: 10.1093/applin/amp010
West, M. (1955). *Learning to read a foreign language* (2nd ed.). London: Longman.

About the Author

Paul Nation is emeritus professor in the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand. He writes and researches in the areas of second language vocabulary learning and language teaching methodology. His most recent books on vocabulary are Nation, I.S.P. and Webb, S. (2011) *Researching and Analyzing Vocabulary*. Boston: Heinle Cengage Learning, and Nation, I.S.P. (2013) *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press (second edition). E-mail: Paul.Nation@vuw.ac.nz