Mid-frequency readers

Paul Nation
LALS, Victoria University of Wellington, Wellington, New Zealand
Laurence Anthony
Waseda University, Tokyo, Japan

This article describes a new free extensive reading resource for learning the mid-frequency words of English and for reading well known texts with minor vocabulary adaptation. A gap exists between the end of graded readers at around 3,000 word families and the vocabulary size needed to read unsimplified texts at around 8,000 word families. *Mid-frequency readers* are designed to fill this gap. They consist of texts from *Project Gutenberg* adapted for learners with a vocabulary size of 4,000 word families, 6,000 word families and 8,000 word families. Each text is available at these three different levels. The goal is to have at least fifty such texts at each of the three different levels freely available. The adaptation is done using the BNC/COCA word family lists and the *AntWordProfiler* program. The article also discusses research that needs to be done on learning mid-frequency vocabulary and on creating and using *mid-frequency readers*.

The vocabulary demands of reading

Research on vocabulary comprehension has shown that a learner of English needs to understand around 98% of the running words (tokens) in a text for unassisted comprehension (Hu & Nation, 2000; Schmitt, Jiang, & Grabe, 2011). Using corpora from various genres, Nation (2006) showed that this value equates to around 8,000 word families (see Table 1), which is an ambitious goal for most learners and would require a lot of deliberate and incidental learning of vocabulary.

Table 1: Vocabulary sizes needed to get 98% coverage (including proper nouns) of various kinds of texts (Nation, 2006)

Texts	98% coverage
Novels	9,000 word families
Newspapers	8,000 word families
Spoken English	7,000 word families
Children's movies	6,000 word families

To reach these high vocabulary sizes, extensive reading should play a large role in any vocabulary learning program, both in helping the learning of vocabulary and in improving its use

(see Pigada & Schmitt, 2006; Waring & Takaki, 2003, for reviews). Unfortunately, most graded reading schemes end at around the 3000 word-family level. This means that if learners with a vocabulary size of 3,000 word families or more want to continue doing extensive reading which is at the right level for them, there is no suitable material. The 5,000-6,000 word-family gap between the end of graded readers and the requirements for unassisted comprehension is simply too large. Also, it is possible that even if a learner has the vocabulary knowledge required for unassisted reading, some of the vocabulary will not be accessed quickly enough for fluent extensive reading. Thus, the need to bridge the gap between graded readers and authentic texts is even more important.

In the past, a series to bridge this gap, appropriately called the Bridge series, was published by Longman, Green, and Co. The Bridge series contained 32 titles including fiction works, such as *Animal Farm*, *Lucky Jim*, *Persuasion*, *The Red Badge of Courage*, and *Great Expectations*, and non-fiction works, including *The Mysterious Universe*, *Changing Horizons*, and *Mankind against the Killers*. Although the series is now out of print, the number of printings for some of the books shows that they, at least, sold well. The following is a note describing the series that appeared in the introduction to *Animal Farm*.

> The *Bridge Series* is intended for students of English as a second or foreign language who have progressed beyond the elementary graded readers and the *Longman Simplified English Series* but are not yet sufficiently advanced to read works of literature in their original form.
>
> The books in the *Bridge Series* are moderately simplified in vocabulary and often slightly reduced in length, but with little change in syntax. The purpose of the texts is to give practice in understanding fairly advanced sentence patterns and to help in the appreciation of English style. We hope that they will prove enjoyable to read for their own sake and that they will at the same time help students to reach the final objective of reading original works of literature in English with full understanding and appreciation.
>
> *Technical Note:*
>
> In the *Bridge Series* words outside the commonest 7000 (in Thorndike and Lorge: *A Teacher's Handbook of 30,000 Words*. Columbia University, 1944) have usually been replaced by commoner and more generally useful words. Words used which are outside the first 3,000 of the list are explained in a glossary and are so distributed throughout the book that they do not occur at a greater density than 25 per running 1000 words.
>
> (from the introduction to the Bridge Series edition of *Animal Farm,* 1945)

The Bridge series involves a reasonable amount of glossing (the glossary is in the form of a list with definitions at the back of the book) and a small amount of adaptation. For example, *Animal Farm* contains a glossary of around 880 words which cover approximately 3.3% of the running words in the text. The number of glossed words for *Animal Farm* is high because no words were replaced in the text. Other glossaries range from 120 to 600 words. Although having an extensive glossary at the back of the book could interrupt the flow of reading, glossed words in the Bridge Series are not bolded or marked in any way in the text. Learners are supposed to look up words only when they need to.

With the growth of personal computers and the development of word family lists and computer programs that use them, the study of the vocabulary load of text has become increasingly more detailed. For example, Nation (2009) looked in detail at the number of changes that would need to be made to adapt texts for learners at various vocabulary size levels. In Table 2, we can see that to adapt the Project Gutenberg version of the novel *Lord Jim* by Joseph Conrad for a learner who knew 4,000 word families, 5% of the word families would need to be glossed and 0.75% of the word families would need to be replaced. In column 3, the "target word families to gloss" is arbitrarily set at a maximum of 5%. If this percentage is lowered, then the percentage in Column 4 needs to be increased. Several unknown words will be easy to guess from context, and words which are easy to guess should not be chosen for replacement. The lowest frequency level words are replaced unless they are

Table 2: Percentage of target word families to support and word families to replace in *Lord Jim* at various levels of previous knowledge

Assumed known word families	% coverage of known word families	% target word families to gloss	% of word families to replace	Total %
2,000	88.32	5.0	6.68	100
3,000	92.16	5.0	2.84	100
4,000	94.25	5.0	0.75	100
5,000	95.65	4.35	0	100
6,000	96.56	3.44	0	100
7,000	97.20	2.80	0	100
8,000	97.74	2.26	0	100
9,000	98.15	1.85	0	100

repeated within the text or they are easy to guess.

Table 2 shows that as learners' vocabulary size increases, the percentages of changes that need to be made become small. However, the weakness of this method of calculating changes is that a small percentage can still be a large number of word families. *Lord Jim* is 132,413 tokens long, so 5% of the tokens equals 6,621 tokens. This is well over 2,500 word families, which is far too heavy an unknown vocabulary load for a reader. What this means is that the number of words replaced needs to be greater so that only a small percentage of the running words (well under 2%) are unknown words. The critical figure is the raw number of unknown word families that need to be dealt with by the reader, not the percentage coverage of text by unknown words.

High-, mid-, and low-frequency vocabulary

It is useful to distinguish three broad frequency levels of vocabulary: high-frequency vocabulary, mid-frequency vocabulary, and low-frequency vocabulary. The idea of high-frequency words has a long history, and Michael West's (1953) A General Service List of English Words containing around 2,000 word families is the most well-developed and well-known example.

Following the lead of Schmitt and Schmitt (2012), here we consider the high-frequency vocabulary to include the most frequent and wide ranging 3,000 word families of English (see Table 3). The arguments in favour of including the first 3,000 word families in the high-frequency level are that the 3,000 word-family level is needed to gain 95% coverage of the running words in most texts (when the coverage of proper nouns and marginal words is included), and that most graded readers end at around the 3,000 word-family level. Note that this figure differs from Nation (2001) who considered this level to contain only the first 2,000 word families.

In Table 3, the mid-frequency vocabulary consists of around 6,000 word families, which when added to high-frequency vocabulary adds up to 9,000 word families. The reason for making the arbitrary cut-off point between mid-frequency and low-frequency vocabulary after the 9th 1000 word-family level is because 9,000 word families provide 98% coverage of most texts, when the coverage of proper nouns and other marginal words is also included.

Table 3: High-frequency, mid-frequency, and low-frequency vocabulary

Vocabulary level	Word family levels (and total)	Nature of the vocabulary
High-frequency	1st 1000-3rd 1000 (3,000)	Wide range, very high-frequency, essential, general purpose vocabulary
Mid-frequency	4th 1000-9th 1000 (6,000)	Wide range, moderate frequency, general purpose vocabulary
Low-frequency	10th 1000 on	Narrower range, low-frequency, some technical vocabulary unique to a particular discipline

In order to create the word family lists reported in the Nation (2009) study, an untagged version of the British National Corpus (BNC) was used. This was divided along genre divisions into 10 roughly equally sized sections each 10,000,000 word tokens long. At around the 9,000 word-family level, the range figures for the most frequent words changed from a value of 10 to a value of 9. That is, at around the 9,000 word-family level, the word families did not occur in all 10 sections of the BNC, but in only 9 of them. This can be seen as marking a change from generally useful vocabulary to more narrowly focused vocabulary.

Table 4 shows examples of word families from a revised list of mid-frequency word family level lists that were developed for this study on the basis of frequency information from the BNC combined with that from the Corpus of Contemporary American English (COCA) kindly supplied by Mark Davies (Nation, 2012). The word families in Table 4 are taken from the lists beginning at the letter *b* and are shown here so that readers of this article can get a feel for the kinds of words in the mid-frequency vocabulary.

Table 4: Example word families from the six 1000 mid-frequency word-family levels using the BNC/COCA lists

Word family frequency level	Example word families
4th 1000	*ballet, balloon, ballot, bankrupt, barn, barrel, baseball*
5th 1000	*badge, bail, bait, balcony, bald, banner, Baptist*
6th 1000	*babe, bachelor, baffle, bandage, banish, banquet, barb*
7th 1000	*badger, bale, ballad, bamboo, baptism, baptize, barbarian*
8th 1000	*babble, backfire, baggy, ballistic, banal, bandit, barber*
9th 1000	*backlog, bailiff, bandwagon, banister, banter, barbaric, bard*

Mid-frequency words are commonly known by adult native speakers of the language, and we would expect native-speaking children beginning secondary school to know many of these words to some degree. Note that the related words *Baptist, baptize*, and *baptism* in Table 4 are separate word families. This is because the stem form of these words is a bound form, not a free-form. That is, there is no word *Bapt* which stands as a free word. Note also that compound words, such as *backfire* and *bandwagon*, are included. This is because these are not transparent compounds where the meaning of the word can be explained directly from the word parts. The test for transparent compounds is to see if it is possible to state the meaning of the compound using the parts with few if any further content words needed. For example, your *birthday* is the *day* of your *birth*.

The low-frequency words of the language are a very large group. The BNC/COCA lists go up to the 25th 1000 word-family level, but the low-frequency words stretch far beyond this. It is not easy to say how many low-frequency word families there are in English, but various estimates put the number at somewhere around 100,000 word families (Nation, in press). The current BNC/COCA word family lists going up to and including the 25th 1000 plus the four lists of proper nouns, marginal words, transparent compounds and abbreviations provide over 99% coverage of the tokens in most texts and corpora. At least half of the words outside the lists turn out to be proper nouns, and a large number of the remainder are transparent low-frequency members of word families already in the existing lists but which have not yet been added to the families (Nation, in press).

Table 5 shows the typical coverage of high-frequency, mid-frequency and low-frequency word families. The high-frequency words, mid-frequency words, and proper nouns, exclamations, transparent compounds and abbreviations add up to over 98% of the running words in the text. The high-frequency words, proper nouns, exclamations, transparent compounds and abbreviations add up to around 95% of the running words.

Table 5: Coverage of the British National Corpus (BNC) by high, mid- and low-frequency word families

Type of vocabulary	% coverage
High-frequency (3,000 word families)	90%
Mid-frequency (6,000 word families)	5%
Low-frequency (10th 1000 word-family level on)	1-2%
Other (Proper nouns, exclamations, transparent compounds, abbreviations)	3-4%
Total	100%

Table 6: Distribution of high, mid- and low-frequency word families in a variety of genres

Level	Spoken	TV/Movies	Children's reading	Novels
High-frequency - 3,000	92.73%	92.16%	90.37%	90.21%
Proper nouns etc	5.10%	4.31%	3.35%	2.99%
High-frequency plus proper nouns	97.83%	96.47%	93.72%	93.20%
Mid-frequency - 6,000	1.68%	2.5%	4.52%	4.67%
Low-frequency	0.49%	1.03%	1.76%	2.13%

Table 6 shows the range of coverages in a variety of million-token corpora that were created for this study. The spoken corpus consists of one million tokens from the spoken demographic section of the BNC and represents informal conversation. The children's reading material is from the New Zealand School Journal. The range of coverage by the mid-frequency word families (1.68-4.67%) varies in line with the coverage of the high-frequency word families. Generally, the higher the coverage by high-frequency word families, the lower the coverage by mid-freq-uency word families. Low-frequency word families follow a similar pattern.

Mid-frequency readers

In order to fill the gap left by the Bridge series and to enable learners to more easily master vocabulary up to the 9,000 word-family level, one of the authors (Nation) has begun developing a set of *mid-frequency readers*. *Mid-frequency readers* are books within a controlled vocabulary for advanced learners of English as a foreign or second language. They are adapted from the original texts using the profiling and simplification tools of *AntWordProfiler* (Anthony, 2012) and are designed to provide interesting, comprehensible reading to fill the gap of 6,000 word families between the end of graded readers and the demands of unsimplified text.

Mid-frequency readers can be used in extensive reading programs or for individual study and enjoyment. Also, they can help learners learn mid-frequency vocabulary and read texts that would otherwise be too difficult. Each book is available at three levels. There is one level for learners who know 4,000 word families, another for learners who know 6,000 word families, and another for those who know 8,000 word families. The first ones to be made available on Paul Nation's web site are *The Art of War*, *More William*, *Glimpses of Unfamiliar Japan*, *Alice's Adventures in Wonderland*, and *Metamorphosis*. Note that two of these books are adaptations of translations, so they could be called friendly translations rather than simplifications.

The *mid-frequency readers* are available free, and can be used for any purpose without permission, as long as they are not offered for sale or offered on the web where payment must be made for access to the site. The adaptation is done in the same spirit that was behind the setting up of the tremendous resource, *Project Gutenberg*. It has been done without payment and without the wish for financial profit.

Table 7: Corpus sizes needed to gain an average of at least ten repetitions at each of the six mid-frequency 1000 word-family levels using a corpus of novels

1000 word-family list level	Corpus size to get an average of at least 10 repetitions at this 1000 word-family level (repetitions)	Number of words appearing once/ twice (out of 1000)	Number of families met	Number of novels
4th 1000 families	534,697 (12.6)	93/73	812 of 4th 1000	6
5th 1000 families	1,061,382 (13.7)	101/79	807 of 5th 1000	9
6th 1000 families	1,450,068 (13.1)	89/82	795 of 6th 1000	13
7th 1000 families	2,035,809 (13.7)	92/63	766 of 7th 1000	16
8th 1000 families	2,427,807 (14.1)	96/70	755 of 8th 1000	20
9th 1000 families	2,956,908 (12.0)	88/78	805 of 9th 1000	25

At present only a few *mid-frequency readers* are available from Paul Nation's web site, but the goal is to have at least fifty, each at three different levels so that the mid-frequency vocabulary is well covered with plenty of repetitions. Table 7 is adapted from Nation (in press), and shows how much reading would have to be done to meet most of the words at each of the mid-frequency 1000 word-family levels.

Column 2 of Table 7 shows that 534,697 words need to be met before 10 repeats of the 4th 1000 word families are encountered. Although this may seem a lot of reading, it only represents six average length novels, and at a reading speed of around 200 words per minute (a moderate reading speed), it would require only 1 hour 5 minutes a week of reading for forty weeks, or 13 minutes a day, five days a week for forty weeks. Such an amount of reading is possible if the material is at the right level for the learners.

Making *mid-frequency readers*

Each *mid-frequency reader* is in a controlled vocabulary and is adapted from the *Project Gutenberg* version of the text. Most of the words beyond the specified 1000 word-family levels have been removed, largely through replacement by higher frequency words, but very occasionally adjectives and adverbs and the occasional short sentence have simply been omitted where this does not affect the story line. This is rarely done. Almost all substitutions are single word substitutions. Where mis-spelling is used to represent accented speech (orl = all, nuff = enough), this is changed to regular spelling. No other changes are made to the books.

The aim of the changes is to make the books more accessible for non-native-speaking learners of English by removing the large number of low-frequency words which are way beyond their vocabulary level. The goal is to adapt the book so that in each book only a relatively small number of word families which make up much less than 2% of the running words (tokens) are unfamiliar words. The number of word families considered to be unknown which are left in the text, differs according to the length of the text, but should only be a few hundred words. These unknown words can be dealt with by guessing from context, or looking them up in a dictionary. The number of words beyond the adapted word-family level is given at the beginning of each book. Here is an example from the 4,000 word-family level version of *Alice's Adventures in Wonderland* adapted by Sonia Millett.

> This book, Alice in Wonderland, is a Mid-Frequency Reader and has been adapted to suit readers with a vocabulary of 4,000 words. It is about 27,500 words in length. It is available in three versions of different difficulty. This version is adapted from the Project Gutenberg E-text prepared by the Project Gutenberg Online Distributed Proofreading Team (http://www.pgdp.net/). In this book, the adaptation involved replacing over 152 word families. There are 82 different word families at the 5th 1000 level and 62 word families beyond that, totalling a target vocabulary of 144 words. It was adapted by Sonia Millett.

Note in the description above that the number of word families at the 5th 1000 word-family level is given (82 in *Alice*). These are the target word families that the reader is expected to deal with and begin to learn while reading. The word families remaining in the book beyond that (62) largely contain words that are repeated several times in the book and so have a good chance of being learned. So, there is a total of 144 target word families in *Alice* beyond the 4th 1000 word-family level, and 152 word families from the 6th 1000 word-family level on are replaced by higher frequency words. *Alice* is a short book at 27,500 tokens long. The average novel is four times this length at over 100,000 tokens long.

Each book is adapted to three different levels:

- The 8,000 word-family level is for learners with a vocabulary size of 8,000 words. The words at the 9th 1000 and 10th 1000 word-family levels are the target words, and the words beyond the 10th 1000 word-family level have been replaced.

- The 6,000 word-family level is for learners with a vocabulary size of 6,000 words. The words at the 7th 1000 and 8th 1000 word-family levels are the target words, and the words beyond the 8th 1000 word-family level have been replaced.
- The 4,000 word-family level is for learners with a vocabulary size of 4,000 words. The words at the 5th 1000 word-family level are the target words, and the words beyond the 5th 1000 word-family level have been replaced.

Most words beyond the target level which occur around ten times or more in a particular book are not removed. The repetition of these words provides a good chance that they might be learned. A few words that are important for the message of the text and that are very difficult to replace are also left in the text.

Examples of adaptation

Here is an example comparing the adapted and original (unadapted) text. The purpose of the comparison is to show how little the text is changed from the original. Once again, the example is from *Alice's Adventures in Wonderland*.

Original

> Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had **peeped** into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'
>
> So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a **daisy**-chain would be worth the trouble of getting up and picking the **daisies**, when suddenly a White Rabbit with pink eyes ran close by her.
>
> There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge.

Adapted to 4,000 word-family level

> Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had **looked** into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'
>
> So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a **flower**-chain would be worth the trouble of getting up and picking the **flowers**, when suddenly a White Rabbit with pink eyes ran close by her.
>
> There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge.

Notice that only two words needed to be changed in these three paragraphs from the beginning of the story.

Table 8 compares the spread of word families in the original and the 4,000 word-family level adaptation of *Glimpses of Unfamiliar Japan* by Lafcadio Hearn. Note how up to the 5,000 word-family level, the adaptation results in an increase in the number of word families. The 5th 1000 word-family level is the target word family level for this book. That is, most of the new words to learn are at 5th 1000 word-family level. From the 6th 1000 word-family level on, the number of

families drops dramatically in the 4,000 word-family level adaptation. The words that remain from the 6th 1000 word-family level on are words that are repeated at least ten times in the book. The original had 1,330 words in the 6th to 25th 1000 word-families. The adaptation has 253 word families.

Table 8: Effects of changes in word families at various frequency levels in the original and 4,000 version of *Glimpses of Unfamiliar Japan*

1000 word-family level	Original	4,000 level adaptation
1	908	917
2	757	784
3	555	574
4	444	455
5	359	376
6	280	56
7	215	42
8	160	26
9	144	17
10	129	16
11	78	19
12	77	11
13	50	10
14	35	7
15	39	8
16	18	5
17	18	3
18	13	4
19	11	3
20	12	4
21	11	2
22	13	5
23	11	6
24	10	7
25	6	2

BNC/COCA lists and *AntWordProfiler*

The BNC/COCA word lists used for the adaptation are based on the BNC and COCA corpora and currently go up to the 25th 1000 word-family level. They are accompanied by lists of proper nouns, exclamations (*gee, ow, um*), transparent compounds and abbreviations. These lists are available free from Paul Nation's web site and Laurence Anthony's web site (http://www.antlab.sci.waseda.ac.jp/)

The BNC/COCA word lists were developed for two main purposes.

1. The lists are designed to be used in the *Ant WordProfiler* program to analyse the vocabulary load of texts.

2. The lists are designed to be as complete as possible lists of the high-frequency (1000-2,000), and mid-frequency (3,000-9,000) words. They also cover 16,000 low-frequency words which include the more common low-frequency words.

The unit of counting in the BNC/COCA word lists is the word family (Bauer & Nation, 1993). The word family was chosen, rather than the word type or lemma, firstly because research has shown that word families are psychologically real (Nagy, Anderson, Schommer, Scott, & Stallman, 1989; Bertram, Laine & Virkkala, 2,000). Secondly, when reading, knowing one member of the family and having control of the most common and regular word-building processes makes it possible to work out the meaning of previously unmet members of the family. Here are some examples of low-frequency family members of high-frequency word families – *burningly, helplessness, snappishly, curiouser*.

The unit of counting makes a substantial difference to the number of words in the language. For example, the first 1000 word families contain 6,857 word types, which means each word family at this level has an average of just under seven words per family. The average number of types per family decreases as you move down the frequency levels. When measuring vocabulary knowledge for productive purposes, the word type or the lemma is a more sensible unit of counting. If the lemma was the unit of counting for receptive purposes, the following items

would be counted as different words – *walk* (verb), *walk* (noun), *walking* (adjective), *walker*. These are all members of the same word family.

The *AntWordProfiler* program (Anthony, 2012) was developed to assist with the adaptation process with the generous support of Compass Publishing in Seoul. *AntWordProfiler* can be used to analyse the word family levels of texts using word-family lists, and also can be used for editing texts according to the word-list levels. The General Service List (West, 1953) and the Academic Word List (Coxhead, 2,000) are the default lists in the program, but the BNC/COCA lists can also be easily loaded into the program and used. *AntWordProfiler* is available free from Laurence Anthony's web site.

Using *mid-frequency readers*

Learners can sit the Vocabulary Size Test at http://my.vocabularysize.com/select/test or www.lextutor.ca to find their vocabulary size, so that they can choose the appropriate level of *mid-frequency readers*.

Ideally, each book should be read electronically on a tablet, e-reader, or computer. This would allow easy electronic look up of unknown words. If the book is read in *Microsoft Word*, learners can right click on unknown words to access Look_up, Synonyms, or Translate. It is also possible to read each book accompanied by a spoken version using a text-to-speech program such as *Dragon Naturally Speaking*.

Because the vocabulary in *mid-frequency readers* is controlled, every unknown word that is met in the readers is worth learning. Such learning includes of course incidental learning through reading, but also can include deliberate study using word cards.

Researching *mid-frequency readers*

Mid-frequency readers are a new concept and have not been researched beyond that described in this article, which focuses on text coverage and the number of unknown words. There are thus many issues that would be worth investigating:

1. Recent research on comprehension and vocabulary knowledge has focused on coverage of the word tokens in a text, generally examining whether 98% coverage is needed for adequate text comprehension. The texts looked at have been rather short and the focus has been on comprehension rather than vocabulary learning. Working on the creation of *mid-frequency readers* has made it clear that from a vocabulary learning perspective, it is the actual number of unknown words that needs to be considered. If only 2% of the words are unknown words, this works out as 6 words per 300 word page, and in a two hundred page book, it equates to well over 1000 word families, even allowing for repetitions. This seems too many to deal with and have a chance of retaining some memory for the meetings. It would be useful to investigate the effect of various vocabulary loads on learners and see what vocabulary load guidelines are best followed when creating *mid-frequency readers*.

2. Nation (forthcoming) used unsimplified texts to examine the occurrence of mid-frequency words and their repetitions. When there are enough *mid-frequency readers*, it would be worth examining the occurrence and repetition conditions they provide to assist learning. This could include looking at how many word families actually occur at each word frequency level, how often these word families are repeated, and the occurrence and repetition of higher frequency vocabulary at later frequency levels.

3. *Mid-frequency readers* do not involve any grammatical adaptation. It is assumed that learners with a vocabulary size of 4,000 word families and beyond can cope with difficulties caused by grammar. This could be seen as a strength as well as a weak-

ness of the books. The strength is that the changes made to the books are minimal, so readers can get close to the original. The weakness is that many of the adaptations involve classic novels and thus the writing style may cause grammatical difficulties. It is important to get feedback on the books to see if some grammatical adaptation is needed.

4. Derivational knowledge of words plays a large part in the adaptations. The word lists used are based on word families, and the assumption is that knowing one or two members of the family allows other members to be understood in context with little difficulty. Further research on this is needed.

5. *Mid-frequency readers* very closely resemble the original texts in both vocabulary demand and actual appearance (e.g. chapter headings, number of pages, word length, etc.). As a result, some learners may feel more motivated to read the original text than the modified one. Again, feedback on the books is necessary to evaluate this motivational effect, and it would also be interesting to investigate if learners actually enjoy the *mid-frequency readers* and find them easier to read.

6. *Mid-frequency readers* provide ideal conditions for researching the nature of simplification and its effect on repetition. *Mid-frequency readers* require replacing only relatively small amounts of vocabulary. Thus, the original and adapted versions of *mid-frequency readers* are usually very close to each other in length. This makes comparison of the original and adapted versions easier and fairer. Questions to look at could include the following:

   What is the effect of simplification on word family repetition?
   What does simplification do to the ideas content of a text?
   Is the amount of simplification required predictable using Zipf's law?
   How predictable is the repetition of words beyond the target level?
   Do learners notice the relative ease of the adapted version?
   How many unknown words can learners tolerate in a 100,000-token novel?
   Do learners read the adapted version more quickly?
   Is vocabulary learning greater from the adapted version?
   Is the vocabulary learned what the word frequency analysis would predict?
   What is the look-up behaviour of learners reading *mid-frequency readers*?
   How can look-up behaviour be explained?

Conclusions

In this article, we have proposed a new concept of *mid-frequency readers*. These books are designed for learners whose vocabulary size is beyond that of traditional graded readers but is still not large enough to deal comfortably with unsimplified text. In addition, we have prepared a set of free, online *mid-frequency readers* that demonstrate the concept. There remain various issues and uncertainties with *mid-frequency readers* that require further research. As a result, we recommend that teachers should approach them with a healthy degree of scepticism, and should seek feedback from the learners if they are adopted. However, we anticipate that *mid-frequency readers* can provide enjoyable and manageable reading conditions for learners and thus become a valuable and much used part of extensive reading programs.

Note: We welcome volunteers to adapt texts for *mid-frequency readers*. On Laurence Anthony's website there is a set of instructions for using the *AntWordProfiler* program and those interested should contact Paul Nation at Paul.Nation@vuw.ac.nz.

References

Anthony, L. (2012). *AntWordProfiler* (Version 1.3.1) [Computer Software]. Tokyo, Japan: 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*(4), 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*(4), 287-296.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

Nagy, W. E., Anderson, R., Schommer, M., Scott, J. A., & Stallman, A. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24*(3), 263-282.

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

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

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

Nation, I. S. P. (2012). Notes on the BNC/COCA lists. Wellington: New Zealand: Victoria University of Wellington. Available from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Nation, I. S. P. (in press). *Learning Vocabulary in Another Language.* (Second edition). Cambridge: Cambridge University Press.

Nation, I. S. P. (forthcoming). How much input do you need to learn the most frequent 10,000 words?

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

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26-43.

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

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 130-163.

West, M. (1953). *A General Service List of English Words*. London: Longman, Green & Co.