What Vocabulary Size is Needed to Read Unsimplified Texts for Pleasure?

David Hirsh and Paul Nation
Victoria University of Wellington

The types of vocabulary in three short novels were analyzed to determine the text coverage of the most frequent 2,000 words of English, and the vocabulary needed to gain 97-98% coverage of the running words in each text. It was found that the most frequent 2,000 words do not provide adequate coverage for pleasurable reading and that a vocabulary size of around 5,000 word families would be needed to do this. The study also showed a need for graded readers at the 2,600- and 5,000-word level and unsimplified texts. The feasibility of preteaching vocabulary and intensive reading of unsimplified texts were also examined.

INTRODUCTION

This study looks at the vocabulary demands of short unsimplified novels which were written mainly for young native speakers of English. This genre was chosen because, for several reasons, it offers the most favourable conditions for a light vocabulary load. Because novels tend to be less formal than most non-fiction texts, the most frequent words of English provide a greater coverage of the text (Hwang 1989). Novels are continuous texts and thus provide opportunities for words to be repeated. In this way the vocabulary load may be lightened. The novels chosen, *Alice in Wonderland* (Carroll 1865), *The Pearl* (Steinbeck 1949) and *The Haunting* (Mahy 1982), were intended for younger readers and thus there is likely to be a tendency to use common rather than uncommon vocabulary. Before looking at the vocabulary of these novels, let us first look at the effect of vocabulary coverage on reading.

HOW MUCH OF THE VOCABULARY IN A TEXT NEEDS TO BE KNOWN?

Knowledge of the vocabulary in a text is one of the many factors that affect reading. In this article important factors like interest in the story, the relationship of the story to past experience, and skill at reading are ignored in order to concentrate on the role that the reader's vocabulary size may have in making reading more pleasurable.

Having to struggle with reading because many words are unknown will take a lot of the pleasure out of reading. If a reader knows 90% of the running words (the tokens) in a text, then there will be 10 unknown words in every 100. If each line in the text

David HIRSH is a teacher of English as a foreign language in Thailand. The work reported here was done for his Master of Arts in Applied Linguistics at Victoria University of Wellington. Paul NATION is Reader in Applied Linguistics at the English Language Institute at Victoria University of Wellington. His research interests are in vocabulary teaching and learning and language teaching methodology. The authors can be contacted at the English Language Institute, Victoria University of Wellington, PO Box 600, Wellington, New Zealand.

contains about 10 words, then there will be one unknown word in every line. Reading will certainly be a struggle. Table 1 shows how the burden changes with increasing text coverage.

Table 1: The number of unfamiliar tokens per 100 tokens and the number of lines of text containing one unfamiliar word

% Text Coverage	Number of Unfamiliar Tokens per 100 Tokens	Number of Text Lines per 1 Unfamiliar Word
99	1	10
98	2	5
97	3	3.3
96	4	2.5
95	5	2
94	6	1.6
93	7	1.4
92	8	1.25
91	9	1.1
90	10	1

There are two important thresholds in Table 1. The first threshold is at the 95% coverage level where there is one unknown word in every two lines. Research by Laufer (1989) and Liu and Nation (1985) shows that 95% coverage of text is needed to gain adequate comprehension and to guess unknown words from context.

The second threshold is around the 97%-98% coverage level, where the density of known to unknown words becomes significantly less.

It is around this point that a 1% increase in coverage has a very big effect on the number of lines per unknown word. In practical terms this means that, from the aspect of vocabulary, reading becomes more pleasurable because there are fewer unknown words. In the next two sections of this article we will look at the number of words a learner needs to know in order to reach this level of coverage.

IS A 2,000-WORD GENERAL SERVICE VOCABULARY ENOUGH ?

Speakers and writers do not use all the words of English with equal frequency. The word *the,* for example, is the most frequently used English word, occurring around seven times in every one hundred words. We use words like *find, place* and *difficult* much more frequently than we use words like *pandemonium, hearty* and *abrogate.* Michael West (1953) had this in mind when, in collaboration with others, he developed a list of around 2,000 high frequency words that would be most useful for a learner of English as a second or foreign language. He called his list 'a general

service list' because the words it contained would be needed (of service) in a wide range (general) of situations, genres and uses. This 2,000-word general service vocabulary is the indispensable basis of all our uses of English. Is this vocabulary sufficient for reading unsimplified short novels? Table 2 shows the vocabulary size needed to reach various levels of coverage in the three texts, *The Pearl, Alice in Wonderland* and *The Haunting.*

Table 2: Percentage text coverage of various vocabulary sizes in three short novels

	2,000	2,000 + proper nouns	2,600	5,000	7,000
The Pearl	89.7	92.5	95	97	98
Alice	91.9	95.0	97	98	99
The Haunting	90.2	94.9	97	98	99

Table 2 shows that a learner with a vocabulary of the 2,000 words of the *General Service List* (West 1953) will be familiar with around 90-92% of the words on any page in the novels. This means that 8-10% of the words, or about one word in every line will be unknown (see Table 1). If, however, proper nouns are included in the learner's known vocabulary, then the coverage rises to around 95% for two of the novels. There are strong reasons for considering proper nouns as words that do not require previous learning. First, the text reveals what we need to know about them as the story progresses. Who *Alice,* the *Dormouse* and the *Mad Hatter* are is revealed by the story. We are not expected to know this before coming to the story. Second, their form (an initial capital letter) and their function clearly signal they are proper nouns.

As Table 3 shows, each of the novels contains a small number of proper nouns. Most of these occur very frequently in the novels. Table 3 also shows the length of each novel (the running words or tokens), the vocabulary of the novel (the different word families), and the number and percentage of the different words that are not proper nouns and that are not in the general service vocabulary.

Table 3: The vocabulary of the novels

	Total running words	Total different word families	Proper nouns	Different words not proper nouns and not in *GSL*
The Pearl	26,479	2,232	18	972 (43.5%)
Alice	27,331	1,743	30	591 (33.9%)
The Haunting	34,909	2,144	22	864 (40.2%)

At first glance it would seem that a vocabulary of around 2,000 words would be enough to read any of these short novels. Column 5 in Table 3, however, shows

that between 35% and 45% of this vocabulary consists of words that are not in the most frequent 2,000 words of English. Moreover, three-quarters or more of these words outside the most frequent 2,000 occur only once or twice in the novels.

There are thus several reasons why the 2,000-word general service vocabulary is not enough for reading short novels with ease. First, there are still too many unknown words in the novels (about one in every two lines). Each novel contains several hundred words that would not be known to a learner whose vocabulary consisted of the *General Service List.* Second, the majority of these unknown words do not occur more than once or twice in the novels. Analysis of the repetitions of all the words in each novel outside the *General Service List* shows that if only three repetitions of the unknown words were required for learning (and this is a very optimistic assumption) there would still not be a significant increase in coverage in the later parts of the novel to make reading more pleasurable.

HOW MANY WORDS ARE NEEDED FOR PLEASURABLE READING?

In an earlier section of this article we looked at how much vocabulary in a text needed to be known. This was done by relating vocabulary coverage to ease of reading. Table 2 gives coverage figures for vocabulary sizes of 2,600, 5,000 and 7,000 words. These vocabulary sizes are based on the assumption that vocabulary is learned in relation to its frequency of occurrence, with higher frequency words being learned before low frequency words. The vocabulary size figures of 2,600, 5,000 and 7,000 were reached by checking the vocabulary in the novels against frequency figures in *A Teacher's Word Book of 30,000 Words* (Thorndike & Lorge 1944). The frequency levels in Thorndike and Lorge were adjusted to match the definition of a word family used in this study. This resulted in a lowering of the Thorndike and Lorge figures. For example, the 12,000-word level in Thorndike and Lorge actually consists of 7,000 word families because Thorndike and Lorge do not include derived forms in word families. Table 2 shows, for example, that 97% of the different word families in *Alice in Wonderland* and *The Haunting* occur within the most frequent 2,600 words in the Thorndike and Lorge list. Table 2 shows that to reach the 97-98% threshold of coverage a vocabulary size of around 5,000 words would be needed.

In this article the term 'word family' has been used several times. For this piece of research a word family was defined as the base form of a word plus its inflected forms (third person *-s, -ed, -ing,* plural *-s,* possessive *-s,* comparative *-er* and superlative *-est*) plus derived forms made from certain uses of the following affixes (*-able, -er, -ish, -less, -ly, -ness, -th, -y, non-, un-, -al, -ation, -ess, -ful, -ism, -ist, -ity, -ize, -ment, in-*). The idea behind a word family is that inflected and regularly derived forms of a known base word can also be considered as known words if the learners are familiar with the affixes. Bauer and Nation (in preparation) look at this in detail.

HOW CAN LEARNERS BE HELPED TO READ UNSIMPLIFIED TEXTS?

This article has shown that a learner would need to have a vocabulary size of around 5,000 word families in order to read a short unsimplified novel with reasonable ease. If learners do not have this vocabulary size, what can be done to reduce the vocabulary burden? Three approaches will be examined – simplification, pre-teaching and learning of needed vocabulary, and intensive study of unsimplified novels.

1. Simplification

Most series of simplified graded readers stop at the 2,000-word level. This is not adequate if learners are to move from graded readers to unsimplified novels. This was apparent many years ago to the designers of the Bridge Series of adapted texts. These were books which were 'moderately simplified in vocabulary and often slightly reduced in length, but with little change in syntax'.

> in the Bridge Series words outside the commonest 7,000 (in Thorndike & 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 1,000 words.
>
> 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.
>
> [Introduction to each Bridge Series reader]

The present research confirms the wisdom of the decision made by the designers of the Bridge Series. Using data from the present study of short novels, it is possible to suggest levels for graded readers beyond the 2,000 level which would make the move to unsimplified novels easier. Table 2 shows three levels – 2,600, 5,000 and 7,000 – which in most cases correspond to a 1% increase in coverage.

The 7,000-word level is not necessary, as the 5,000 level already provides learners with 98-99% coverage. Table 4 shows the changes that would have to be made to the three short novels to bring them within the 2,600- and 5,000-word levels.

Table 4: The maximum changes necessary in terms of running words and different word families to bring the three short novels within two proposed graded reader levels

	2,600 level		5,000 level	
	Tokens to change	Types to change	Tokens to change	Types to change
The Pearl	1,323	901	794	702
Alice	819	524	546	440
The Haunting	1,047	765	698	618

Table 5 shows the opportunities for meeting new vocabulary if the three novels were simplified to the two levels. The 2,600 level assumes learners know 2,000 words. The 5,000 level assumes learners know 2,600 words. So, if *The Pearl* was simplified to the 2,600 level, learners would meet 71 words that were not in the 2,000 word vocabulary, but which were in the 2,600 word vocabulary. This figure affects the density of unknown to known words and thus reading difficulty, and also affects how many words they might learn by reading the simplification.

Table 5: Minimum words at a particular vocabulary level in proposed simplifications of three novels

	2,600 level	5,000 level
The Pearl	71	199
Alice	68	84
The Haunting	99	147

The rules governing simplification to these levels would be :

1 Words above the 2,000 level should be repeated in the same text wherever possible.

2 Words outside the derived level (2,600 or 5,000) which occur five times or more should not be changed.

Computer programmes such as the *Oxford Concordance Programme* can be used to help apply these rules.

2. Teaching and learning new vocabulary before reading

One way of helping learners read a text which contains too many unknown words is to get them to learn many of the words before they begin to read the text. Is this feasible for the three novels in this study ?

Table 6 shows the frequency of the words outside the 2,000 level which are not proper nouns.

Table 6: Frequency of occurrence of words outside the 2,000 level in the three novels

	Total words outside 2,000 level	Words occurring 1 or 2 times	Words occurring 5 times or more	Words occurring 7 times or more	Words occurring 10 times or more
The Pearl	972	776	75	42	19
Alice	591	341	63	35	16
The Haunting	864	712	52	22	13

Note the large numbers of words occurring only one or twice in the novels. There is little point in studying these beforehand as there are too many of them and each one will not be repeated enough times in a novel to enrich or establish it. It would be

feasible to pre-teach or get the learners to learn the words occurring 5 times or more. This would bring text coverage of *The Pearl* just over the 95% threshold, and *Alice* and *The Haunting* to near the 97% threshold. The remaining words outside the 2,000 level could be put into a glossary. Such glossaries for the three texts in this study would have to contain well over 500 words. Glossaries for the Bridge Series contain up to 1,000 words.

Computer analysis of the text would be needed to choose the words. If a novel was to be a set text for a large number of learners or for several years, the small amount of cost and effort of scanning the text into the computer and running a simple word frequency programme over it would be well repaid.

3. Intensive study of unsimplified novels

Is it worth studying a novel intensively and learning every unknown word? The answer to this question is probably no. First, there would be many unknown words to learn (see Table 3). Second, the learner would have to read 3-5 novels of a similar size in order to meet about half of the words again. The other half would not be met again in those novels.

It was found in this study that the more times a word was repeated in one novel, the more likely it was to occur in other texts. Therefore, instead of learning every unknown word, it is better to learn words that were met before in the novel. Computer analysis of a novel could easily provide a list of such useful words.

CONCLUSIONS

This study examined unsimplified texts which provide the most favourable conditions for reducing the vocabulary burden in reading. It was found that even when proper nouns are added, a 2,000-word general service vocabulary is not sufficient to allow pleasurable reading of a text. To achieve such reading it is necessary for learners to have a vocabulary of around 5,000 words. Because the gap between the end of most series of graded readers and the vocabulary size needed to read unsimplified texts is so large, there is a need for two more levels of graded readers beyond the 2,000-word level. These could usefully be at the 2,600- or 3,000-word level and the 5,000-word level. If unsimplified texts are used with learners with a 2,000-word vocabulary, it is worthwhile considering computer analysis of such texts to guide vocabulary learning before or during reading. The most useful rule of thumb is pre-teach or learn repeated words.

BIBLIOGRAPHY

Carrol, L. (1865) *Alice's Adventures in Wonderland.* New York: Random House.

Hwang Kyongho (1989) *Reading newspapers for the improvement of vocabulary and reading skills.* Unpublished M.A. thesis. Wellington: Victoria University.

Laufer, B. (1989) What percentage of text-lexis is essential for comprehension? In C. Lauren and M. Nordman (Eds.) *Special language: From humans thinking to thinking machines.* Clevedon: Multilingual Matters.

Liu Na and Nation, I.S.P. (1985) Factors affecting guessing vocabulary in context. *RELC Journal,* 16, 33-42.

Mahy, M. (1982) *The Haunting.* London: Dent.

Nation, I.S.P. (1990) *Teaching and Learning Vocabulary.* New York: Newbury House.

Steinbeck, J. (1954) *The Pearl.* Oxford: Heinemann.

Thorndike, E. and Lorge, I. (1944) *The Teachers' Word Book of 30,000 Words.* Columbia: Teachers College.

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