Reducing the Vocabulary Load and Encouraging Vocabulary Learning through Reading Newspapers

Hwang Kyongho and Paul Nation
Victoria University of Wellington, New Zealand

This paper describes how a particular way of selecting stories reduces the vocabulary load and increases the repetitions of new vocabulary for the learners with a limited vocabulary knowledge when they read newspaper articles. It looks at two ways of selecting newspaper stories: selecting *running stories* (i.e. a story and its 'followups') and selecting *unrelated stories* and the effect that they have on the repetitions of words outside the 2,000 most frequent words which we assume to be the vocabulary of EFL high school graduates entering universities.

The analysis of 20 sequences of four *running stories* and 20 groups of four *unrelated stories* shows that ways of selecting stories have a major effect on the repetitions of the words outside the 2,000 most frequent words. *Running stories* provide more repetitions of low frequency words, and therefore reduce the vocabulary load to a greater extent and provide better conditions for the acquisition of words outside the 2,000 most frequent words. This has implications for course design.

Newspapers provide a readily available and interesting source of material for learners of English. However, learners face difficulties in reading them and a major source of difficulty is the vocabulary load.

LEARNERS' VOCABULARY

The limited vocabulary of EFL high school graduates is a major source of difficulty in reading newspapers. Laufer (1986) has shown that in order to achieve successful comprehension, learners need 95% lexical coverage of a text. That is, they need to know sufficient different words (types) to account for 95% of the running words (tokens) in a text. Recent studies suggest that the number of words necessary to cover 95% of an unsimplified text is about 5,000. According to Deville et al (1985), Ostyn and Godin (1985) and Hindmarsh (1980), a lexicon of 5,000 words would give a coverage of 90-95% of the lexis in authentic texts. A study by Laufer (1986) determined 5,000 words as the threshold level beneath which readers are not expected to read an authentic text successfully. This means that when learners know about 5,000 words, they will know about 95 per 100 tokens in any text, thus reaching 95% lexical coverage.

Although there are differences in estimates between studies, these are mainly due to the different definitions of a word.

Most high school graduates in EFL countries have a far smaller vocabulary than

---

Hwang KYONGHO holds a BA in English and Diploma in TESL (Victoria University, New Zealand). He has taught EFL in Korea. This article is based on work done for an MA in Applied Linguistics at Victoria University. His interests include vocabulary teaching and reading in a foreign language.
Paul NATION teaches at the English Language Institute at Victoria University of Wellington. He has taught in Indonesia, Thailand and the United States. His interests are language teaching methodology and vocabulary learning.

this. Laufer (1987:6) provides findings from several studies

> Indonesian graduates are reported to have a passive vocabulary of 900 to 1000 words (Quinn 1970); Chinese – 1,200 (Shui-Chun 1982); Malaysian – 1,000 words (Criper 1981); Nepalese – 500 to 700 (Davies et al 1984); Tanzanian 1,700 (Criper and Dodd 1984).

Goethals et al (1081: 14) discussing the results of a 2,000 word level test suggest,

> the results of the recalling pre-tests taken with these words in the Flemish part of Belgium during the last 3 years showed that students entering universities did not perform all that well — between the positions 1,501 down to 2,000 of this list.

Narayanaswamy (1972) reports that his vocabulary test showed that learners in India at a pre-university school class who had been taught English for a period of six to seven years and had been through Indian language-medium schools knew less that a thousand words of the *General Service List*.

THE VOCABULARY OF NEWSPAPERS

Most EFL high school graduates will experience difficulty with vocabulary in reading newspapers.

According to Hwang's study (1989) involving the analysis of 80 newspaper articles (a total of 26,722 running words), about 80% of the words in newspaper texts are covered by the most frequent 2,000 words and about 10% are proper nouns. This means that if learners know the most frequent 2,000 words and all proper nouns, they will reach about 90% coverage of newspaper texts. Hwang's study suggests that most proper nouns in newspapers may well be regarded as items known to EFL high school graduates for the following reasons.

1. The meanings of most proper nouns can be inferred from the learners' previous knowledge gained through their mother tongue. For example, many geographical names (e.g. Canada, New Zealand, Tokyo, Beijing etc.) and names of famous people (e.g. Margaret Thatcher, Lech Walesa, Pope John Paul II etc.) have been learned in the learners' mother tongue from various sources such as school, television, radio or newspapers or the people around them.

2. The meanings of proper nouns which are not well known and even those that may be well-known are explained in context in newspaper articles. For example, an article about the election in France, 1988 contained the names of three people: Jacques Chirac, Francois Mitterrand and Vahid Gordji. All these names were explained in context as in *Prime Minister* Jacques Chirac,

> *his (Chirac's) socialist rival* Francois Mitterrand and Vahid Gordji, *an Iranian suspected of being implicated in 1986 bomb attacks in Paris.* Moreover, most of the proper nouns in newspaper articles (over 95%) are names of places or people.

This study assumes that the learners know the most frequent 2,000 words and proper nouns and considers whether choosing articles in a certain way will take learners to a 95% coverage of the vocabulary of selected newspaper articles.

HYPOTHESIS

When learners with a vocabulary of 2,000 words read very closely related stories, for example, a sequence of stories about the election in France, they will encounter quite a few unknown words in the first text such as *aptitude, contempt, hostages, implicated, polls, rally.* To comprehend the story successfully, the learners will have to look up these new words, at least some of them, in a dictionary and this will enable them to learn the meanings of the new words. Alternatively they can try to guess the meanings of new words from the context and this will provide the learners with a partial or an approximate understanding of the meanings of the new words. When they read the second text in the series, the learners will encounter words such as *ballot, chasing, polls, rally, rival* and *terrorism.* However, some of the new words such as *polls* and *rally* will have already occurred in the first text. Similarly in the third text, the learners will encounter words such as *ballot, conservatives, contenders, hostages, polls* and *rally.* But many of these words are those which learners have met in the first and the second texts. In this way, proceeding through a sequence of stories will reduce the number of the words unfamiliar to the learners and will help the learners with a limited vocabulary knowedge to read more efficiently without being interrupted too often by unknown words.

On the other hand, if the learners read several stories which have little or no relatedness in topic, words outside the 2,000 words occurring in one story are not likely to recur in another story. In that case, the learners will have to cope with the meanings of many unknown words to comprehend each story.

Therefore it can be postulated that newspaper stories which continue the same topic are likely to provide more repetitions of more words outside the 2,000 most frequent words than stories which are not related.

DATA

In order to investigate the above hypothesis, 20 sequences of 4 *running stories* (accounts of developments in the same event over 4 successive days) and 20 sets of 4 *unrelated stories* were selected from four newspapers: *The Dominion, The Evening Post, The Korea Times* and *The Korea Herald.* A computer was used to analyze

these stories and manual processing followed when necessary. A survey of two weeks' issues of a daily paper revealed that 19% of the stories in the domestic, international and sports sections were *running stories* occurring 4 times or more.

PROCEDURE

It is necessary to look at an example showing the density of new words in a group of stories before discussing the procedure involved in finding the necessary information.

Table 1 shows the proportion of new word families (words outside the first 2,000 words of English which are not proper nouns) that students will encounter in each text in a series of *running stories* (stories about the election in France in 1988) investigated in this research. It also shows how many of the new word families in one or more previous stories occur in a particular story.

Table 1 The density of word families outside the first 2,000 words in a series of running stories

story	length of text (number of tokens)	*total number of new word families	word families which did not occur in any previous stories (density)	word families which occurred in any one or more previous stories
elect 1	255	20	20 (7.8%)	8
2	254	18	13 (5.1%)	5
3	259	30	25 (9.7%)	5
4	336	23	16 (4.4%)	7
total		74		

* total number of word families outside the first 2,000 words

Table 1 shows that learners with a vocabulary knowledge of 2,000 words will encounter 20 new word families in the first text in the series of stories about the election in France. These 20 word families account for 7.8% of the running words in the first text (the method of calculation is explained later). When the learners read the second text, they will encounter 18 word families outside the 2,000 words. 5 of these 18 word families occur in the previous story and are therefore now known by the learners. Thus the remaining 13 word families will be new to the learners. This results in a decrease in the density of new word families to a 5.1% density, in the second text of the series. This means that 94.9% of the words in the second text are familiar to the readers.

The following procedures were used to compare the density of new word families in *running stories* and *unrelated stories*.

1. 20 series of *running stories* and 20 groups of *unrelated stories* were selected. To compare the effect of reading these two different combinations of stories on the density of new words, only four parts in each series of *running stories* were selected so that each part of a series of *running stories* could be matched with each part of a group of *unrelated stories* in text length (i.e. number of tokens). On the average each part contained 335 tokens.

2. In order to find the density of words outside the 2,000 most frequent words using a computer, a word list containing the 2,000 most frequent word families was prepared. This list consists of 1,000 words with a frequency of 332 or above in the *General Service List* (West 1953) and another 1,000 words chosen from Francis and Kucera's rank list (1982). The selection of the second 1,000 words from Francis and Kucera's rank list was to overcome the weaknesses of the second 1,000 words of the *General Service List* which show a small lexical coverage of texts in general (See Hwang 1989).

   The concept of a 'word family' is used to represent a group of words whose meanings can be inferred when the meaning of the base form in the group is known to a learner. A word family includes a base form, its inflected forms and most of the derivatives whose meaning can be inferred from the base form. For example, a word family includes a base form *agree,* its inflected forms *agreed, agrees* and *agreeing,* and its derivatives, *agreement, disagree* (and the inflected forms of *disagree: disagreed,* and *disagreeing*) and *disagreement.* This concept is useful when we are concerned with reading vocabulary. The 2,000 most frequent words include 2,000 base forms, their inflected forms and derivatives, making a total of 8418 items.

3. A specially written computer program called "VORDS" was used to find the length of each text and also to obtain the frequency list of the words in the base list (i.e. the 2,000 most frequent words) and that of the words outside the first 2,000 word list.

4. These words were grouped into word families. For example, *candidate* and *candidates* were grouped together into a single word family, *candidate.* Then the number of word families outside the first 2,000 word list was counted manually.

5. These lists were used to find the word families which did not occur in any of the previous stories. In order to calculate the density of new words in a text, there are two possibilities (see Figure 1). Method (1) uses words families while method (2) uses tokens.

   As it is important that new items to be learned are distinguished from previously met items, method (1) is used in this study: once a new word has

been met, then on the next occurrence in the same story it is not counted as a new word any more. If method (2) is used, new words and repetitions of the same words will not be distinguished.

$$(1)\quad \frac{\text{number of new words (word families)}}{\text{total number of tokens}} \times 100$$

$$(1)\quad \frac{\text{number of new words (tokens)}}{\text{total number of tokens}} \times 100$$

Figure 1: Possible Ways of Calculating the Density of New Words

RESULTS AND DISCUSSION

The above procedure provided the information in Table 2.

Table 2: The average density of word families outside the first 2,000 words in running stories and unrelated stories

running stories (20 series of 4 running stories)				
story	length of text (number of tokens)	total number of new word families	word families which did not occur in any previous stories (density)	word families which occurred in any one or more previous stories
1	309.6	18.0	18.0 (5.8%)	0
2	345.1	19.9	15.9 (4.6%)	4.0
3	342.8	20.0	14.9 (4.3%)	5.1
4	342.3	20.3	12.8 (3.7%)	7.5
total		61.6		
unrelated stories (20 sets of unrelated stories)				
story	length of text (number of tokens)	total number of new word families	word families which did not occur in any previous stories (density)	word families which occurred in any one or more previous stories
1	309.7	16.9	16.9 (5.4%)	0
2	343.9	18.4	18.1 (5.3%)	0.3
3	344.2	20.8	20.1 (5.8%)	0.7
4	342.1	19.1	18.1 (5.3%)	1.0
total		73.2		

Table 2 shows that learners with a vocabulary text knowledge of 2,000 words will encounter about 18 new word families when reading a text in a series containing about 310 tokens. This means that the number of new word families in the first text make up 5.8% of the tokens in the text. When these learners read the following stories in the same series, the density of new word families gradually decreases: 4.6% in the second, 4.3% in the third and 3.7% in the fourth text, This is due to the repeated occurrence of word families outside the first 2,000 words in different stories of the same series. In terms of actual words rather than percentages, when reading the second text, learners will encounter about 20 (19.9 on the average) word families outside the 2,000 words. 4 out of these 20 word families occur in the first text and we assume that these 4 words will be learned then by the learners. The remaining 16 word families will be unknown to the learners. Similarly, learners will encounter 20 word families in the third text. About 5 out of these 20 word families occur in the first or the second text and these will be known by the learners, so the remaining 15 word families will be unknown to the learners. Finally, when learners read the fourth text, they will encounter only 12 to 13 unknown word families among 20 word families outside the first 2,000 words. In this way learners will encounter a decreasing proportion of unknown word families as they proceed through a series of *running stories.*

On the other hand Table 2 shows that reading *unrelated stories has a much smaller effect on reducing the density of new word families. When proceeding through four unrelated stories,* the density of new word families remains consistent or fluctuates within a narrow range (5.3% – 5.8%). The density of new word families does not decrease to a significant degree as it did in the *running stories.* In general a word family outside the 2,000 words in one newspaper story rarely recurs in another unrelated story in any group of four stories.

Table 3: ANOVA for the density of new word families in 20 series of 4 running stories and 20 sets of 4 unrelated stories

running stories				
source of variance	SS	d.f.	MS	F
Between groups	56.55	3	18.85	5.18*
Within groups	276.49	76	3.64	
*P<.01				
unrelated stories				
source of variance	SS	d.f.	MS	F
Between groups	2.61	3	0.87	0.30
Within groups	222.79	76	2.93	(n.s.)

The analysis of variance in Table 3 clearly shows the difference in the change of the density of new word families between *running stories* and *unrelated stories.* The significant F value in the first analysis of variance indicates that there is a real decline in the density of new word families in a series of *running stories.* However, the fact that the second F value is not significant shows that there is no meaningful change in the density of new word families when the stories are unrelated.

Another way of analyzing the difference between the two types of story is to look at the difference in density between the two types of stories at the four steps in the sequence, in other words, to compare the density of new word families in the first *running story* with that in the first *unrelated story,* the second *running story* with the second *unrelated story* and so on.

Table 4 shows that if there is a series of two stories, there is no significant difference in the density of new words between running and unrelated stories. But with a series of three stories, a significant difference is found between the two. In other words, the density of new word families is much higher in *unrelated stories* than in *running stories.* An even bigger difference is found when the series is extended to 4 stories ($P<.001$). This provides evidence for the hypothesis that selecting *running stories* results in the reduction of the vocabulary load in reading newspaper stories which is not likely to be achieved by random selection of stories.

Table 4: The significance of difference between running stories and unrelated stories

	mean % of new word families		t	significance
	running stories	unrelated stories		
1	5.8%	5.4%	0.68	n.s.
2	4.6%	5.3%	-0.01	n.s.
3	4.3%	5.8%	-2.62	$p<.02$
4	3.7%	5.3%	-3.71	$p<.001$

REPETITIONS

So far we have looked at how, in general, word families outside the first 2,000 words occur in a sequence of running stories and in a group of unrelated stories. Now we will look at the total repetitions of the word families outside the first 2,000 words and the possible effect of these repetitions on vocabulary acquisition.

Previous research shows that word learning is a gradual process and it often proceeds by small increments. Therefore repeated encounters with an unfamiliar word play an important role in establishing the complete meaning of the word.

Nagy and Anderson (1985:237) state that

> Although a single encounter with a word would seldom lead to a full knowledge of its meaning, substantial, if incomplete, knowledge about a word

can be gained on the basis of even a single encounter. Therefore, if coupled with a sufficiently large volume of exposure to written language, incidental learning from context should be able to account for a substantial amount of vocabulary growth.

If a single encounter with a word is not enough to gain a full knowledge of its meaning, how many times need a learner encounter the word? There is no clear answer to this particular question since the effort needed to learn the full meaning of a word largely depends on individual words. Some words may carry more various meanings or a more complex meaning than others (Higa 1965; Nagy, Anderson & Herman 1987). Some contexts may be more helpful than others (Beck, 1983). However there are studies which have examined readers' knowledge of a word gained while reading in relation to its frequency of occurrence. These studies provide information on the approximate number of repetitions necessary for a word to be learned. Kachroo's study (1962) shows that words occurring seven times or more in a text book were known to most learners. More recently Saragi, Nation and Meister (1978) suggest that the minimum repetitions for a word to be learned in a reader should be about 10. Therefore for this study, we will regard between seven and ten as the necessary repetitions for learning to occur. Using this information, we will look at whether reading *running stories* provides better conditions for learning vocabulary than reading *unrelated stories.*

Table 5: The number of word families which are expected to be learned when N is regarded as the minimum repetitions needed.

N	the number of word families to be learned in	
	a series of **running stories**	a group of **unrelated stories**
2	17	14
7	2 (2.1)	1 (0.8)
10	1	0
total word families in a group of 4 stories	61.6	73.2

( ) : number of word families when one decimal point was calculated.

Table 5 shows that in a series of 4 *running stories,* the learners encounter 62 (61.6) word families outside the first 2,000 words (excluding proper nouns). Similarly, in a group of 4 *unrelated stories* they will encounter 73 (73.2) word families outside the first 2,000 words. It shows that when learners read a sequence of 4 *running stories,* 45 of the 62 new word families which they will encounter will be onetimers (words which appear only once). About 17 of the 62 word families occur twice or more and 2 (2.1) of the 62 seven or more times. Only about one of the 62 occurs ten or more times.

On the other hand, when learners read a group of *unrelated stories,* about 59 of the word families outside the first 2,000 words will be onetimers. About 14 of the 73 word families occur twice or more. A smaller number, about one (0.8) out of the 73 word families occurs seven or more times. The learners will rarely encounter a word occurring ten or more times in a group of *unrelated stories.*

CONCLUSIONS

The analysis in this article shows that a higher proportion of word families outside the 2,000 words will recur in stories from the same series, thus reading *running stories* reduces the vocabulary load to a greater extent than reading *unrelated stories.* It also shows that *running stories* provide more repetitions of more words outside the first 2,000 words than *unrelated stories,* and thus provide more favourable conditions for learning vocabulary than *unrelated stories.*

The method of calculating the density of new word families in this study (see Figure 1) assumed that the learners learn new words on their first occurrences. According to this way of counting, learners can reach about 94-95% lexical coverage of separate newspaper articles and more than 95% lexical coverage from the second story on of a series of *running stories.*

This method has been very useful in distinguishing new items from those items that readers would have met in one of the previous stories if they were reading (n)th text in a group of stories.

But Laufer's finding that 95% lexical coverage is needed to read authentic texts (1986) seems to be based on another assumption that repeated occurrences of new words are still new items. A similar assumption is made in Liu Na and Nation (1985). When method (2) (See Figure 1) was used to calculate the absolute proportion of new words (tokens) to total tokens, it was found that learners with a knowledge of the first 2,000 words and all proper nouns will reach only about 90% lexical coverage of ordinary newspaper articles. If they learn all the new words in the first text in a series, the learners will reach 91.8% lexical coverage of the second text of the same series. Similarly, if the learners learn all the new words in the first two texts, they will reach 92.3% lexical coverage of the third text and if they learn all the new words in the third text as well, they will reach 93.4% of the fourth.

This means that when the learners with a vocabulary knowledge of 2,000 words (and all proper nouns) read a newspaper article without any particular preparation, they will encounter about 10 unknown words per 100 words, about 30-35 unknown words (about 20 word families) in the text.

With so many unknown words in each newspaper article, the learners would not be able to read newspaper articles successfully. They would need to look up at least about 10 words to be able to read each newspaper article successfully.

But by reading a sequence of *running stories,* the learners can reduce the vocabulary load as they proceed. There are ways of providing additional help. A series of 4 *running stories* tends to involve five to ten *key words* (excluding proper nouns) which occur three or more times in the series (Hwang, 1989). For example, a series of 4 stories about the election in France, 1988, contained six *key words:* clash, hostage, rally, poll, cheat and terrorism. Preteaching these words would enable the learners to read each story in a series of *running stories* with only the occasional need to consult a dictionary. EFL high school graduates with a limited vocabulary should also be encouraged to guess from context (Nation, 1988). In addition, it will be useful for learners to gain some background knowledge of the stories to be read through the learners' own language.

IMPLICATIONS

This study has looked at different ways of selecting stories in newspapers and the effect that these ways have on the occurrence of low frequency words. It has been found that reading *running stories* provides better conditions for reducing the learners' vocabulary load and also for the acquisition of vocabulary.

This study also has implications for the use of texts other than newspapers. In EFL countries textbooks in common use consist of a dozen or so lessons based on separate texts often with little relatedness in topic. The lack of relatedness makes it unlikely that learners will meet the vocabulary of the texts already studied in the following lessons except in the case of a small number of high frequency words. As it has been shown that texts closely related to each other in topic provide more repetitions of vocabulary, single continuous stories such as novels and graded readers are likely to provide learners with more favourable conditions for learning vocabulary (Nation and Wodinsky 1988). Therefore future studies of these two types of texts will help language teachers and learners to gain a much clearer understanding of how to choose texts and also how the conditions for vocabulary learning and reading comprehension can be improved.

REFERENCES

Anderson, V. Billie and John G. Barnitz (1984) "Cross-cultural schemata and reading comprehension instruction". *Journal of Reading,* 28,2, 102-108.

Criper, C. (1981) "The role and function of English in the 1980s". A report for the government of Malaysia.

Criper, C. and W.A. Dodd (1984) "Report on the teaching of the English language and its use as a medium of instruction in Tanzania".

Davies, A., Glendenning, E.H. and McLean, A.C. (1984) "The English language teaching survey of Nepal". A report by a British Council / ODA Survey team and addressed to the Ministry of Education and Culture.

Deville, G., M. Vandecasteele, P. Ostyn and P. Kelly (1985) "Measuring the F.L. learner's lexical needs". Paper presented at the 5th European LSP Symposium in Leuven, Belgium.

Francis, W.N. and Kucera, H. (1982) *Frequency Analysis of English Usage.* Boston: Houghton Mifflin.

Goethals, M., L.K. Engels and T. Leenders (1987) "Automated analysis of the vocabulary of English texts". Paper delivered at the AILA conference in Sydney.

Higa, Masanori (1965) "The psycholinguistic concept of 'difficulty' and the teaching of foreign language vocabulary". *Language Learning,* 15, 3 & 4, 167-179.

Hindmarsh, R. (1980) *Cambridge English Lexicon.* Cambridge: Cambridge University Press.

Laufer, Batia (1987) "The lexical perspective of reading comprehension". *English Teachers' Journal* (Israel), 35, 58-67.

Laufer, Batia (1986) "What percentage of text-lexis is essential for comprehension?" LSP symposium, Vaasa.

Nagy, W.E. and R.C. Anderson (1984) "How many words are there in printed school English?" *Reading Research Quarterly,* 19, 3, 304-330.

Nagy, W.E., R.C. Anderson and P.A. Herman (1987) "Learning words meanings from context during normal reading". *American Educational Research Journal,* 24, 2, 237-270.

Narayanaswamy, K.R. (1972) "An experiment in reading comprehension at the college level". *ELT Journal,* 26, 3, 300-309.

Nation, I.S.P. (1988) *Teaching and Learning Vocabulary.* Wellington: English Language Institute of Victoria University of Wellington.

Nation, I.S.P. and J. Coady (1988) "Vocabulary and Reading". In Carter, R. and M. McCarthy (Ed.) *Vocabulary and Language Teaching.* New York: Longman.

Ostyn, P. and P. Godin (1985) "RALEX: An alternative approach to language teaching". *The Modern Language Journal,* 69, 346-353.

Palmer, H. (1931) *The Second Interim Report on Vocabulary.* Tokyo: I.R.E.T.

Quinn, G. (1970) "The English vocabulary of some English university entrants". Department monograph. IKIP Kristen Satya Wacana Salatiga, Indonesia.

Saragi, T., I.S.P. Nation and G.F. Meister (1987) "Vocabulary learning and reading". *System,* 6, 2, 72-78.

Shui-Chun, G. (1982) "A Survey of the size of vocabulary of Chinese students". *Language Learning and Communication,* 1 (2), 163-178.

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

Wodinsky, M. and I.S.P. Nation (1988) "Learning from graded readers". *Reading in a Foreign Language,* 5, 1, 155-161.