The Newspaper Word List: A Specialised Vocabulary for Reading Newspapers

(Teresa) Mihwa Chung

Korea University of Sejong, South Korea

The primary purpose of this study is to identify words that are of special importance for reading newspapers. In the Newspaper Corpus of 579,849 running words, 588 word families are identified as Newspaper Words. These words account for 6.8% of the running words in the corpus. When combined, proper names and 2,521 families of the *General Service List of English Words* (GSL) and the NWL make up 92.5% of the running words in the corpus. This is lower than the 98% ideal coverage required for understanding a text successfully, but very high given the small vocabulary size. Thus, the NWL will give the best return for vocabulary learning to learners of English as a foreign language who wish to read newspapers as soon as possible.

本研究の目的は新聞を読むのに必要な語彙を特定することである。Newspaperコーパスに記載された579,849語の中から6.8％に上る588語をNewspaper　Wordsとして選び出した。固有名詞、General Service List of English Wordsの2,521語、NWLの語彙を総計するとコーパスの92.5％になる。テキストを理解するのに必要とされている98％よりはやや低い数値であるが、NWL全体の総語数を考慮すれば非常に高い値であるということができる。したがってNWLは新聞英語を早く読めるようになりたいと望む英語学習者にとっては最も効率のよいものであるということができる。

JALT Journal, Vol. 31, No. 2, November 2009

Reading is one of the most common and important ways of learning another language and one major type of reading material is newspapers. Newspapers are often used in reading classes in order to develop reading skills and expand vocabulary knowledge (Hwang & Nation, 1989; Klinmanee & Sopprasong, 1997). There are several reasons for this. Firstly, newspapers are easily and cheaply available in hard copy or online. Secondly, newspapers are authentic materials that are commonly read by the native speakers of the language. Thirdly, they provide a wide choice of interesting topics from which teachers or learners can choose reading texts. Finally, reading newspapers is considered to be not only a good way of reviewing old vocabulary learned, but also of learning new vocabulary from context (Hwang & Nation, 1989).

Despite these potential advantages, many learners find it difficult to read unsimplified newspaper texts. There are a number of factors that contribute to difficulty in reading, but vocabulary knowledge has consistently been found to be the most influential factor affecting comprehension (Hirsh & Nation, 1992; Nation & Coady, 1988).

Hu and Nation (2000) suggested that knowledge of at least 98% of the total words (tokens or running words) in a text is the minimum required for adequate reading comprehension. A recent study by Nation (2006) examined the receptive vocabulary size needed for reading newspapers using the reportage category of the parallel LOB [the Lancaster Oslo/Bergen Corpus of British English] (Johansson, 1978), FLOB (Hundt, Sand, & Siemund, 1999), Brown (Francis & Kučera, 1979), and Frown [the Freiburg-Brown Corpus of American English] (Hundt, Sand, & Skandera, 1999) corpora. In that study, Nation used the British National Corpus (BNC) in order to develop word frequency lists and applied them to other corpora, including newspapers. He estimated that knowledge of the 8,000 most frequent word families and proper names is needed to reach 98% lexical coverage. This number represents a fairly large vocabulary, particularly for adult learners of English who want to read newspapers as a means of learning English and for knowing what is occurring inside and outside of the country. An important methodological approach used by Nation (2006) involved using a reportage news category without dividing it into smaller news sections (e.g., international news and business news); as a result, a range criterion was not included, which is important when selecting a wide range of words occurring with high frequency in newspapers. In the present study, four news sections are examined in order to obtain more detailed results from coherent sections.

It is important to remember that particular words common in certain kinds of writing occur frequently in those texts, and therefore provide good coverage for those text types. A good example of a specialised vocabulary is the Academic Word List (Coxhead, 2000), which consists of only 570 word families, and provides coverage of at least 9% of the running words in a wide range of academic texts. As another example, Ward (1999) notes that a vocabulary of only 2000 word families provides 95% coverage of the tokens in many engineering texts, which is sufficient for 1st-year students to read required textbooks. Learning such a specialized vocabulary provides learners with a shortcut to coping with the vocabulary of such texts.

For this reason, if researchers develop specialized vocabulary lists, only a small vocabulary is needed to make certain types of texts easily accessible provided the specialized list of vocabulary is acquired. In addition, when teachers narrow the focus on teaching vocabulary such as for reading newspapers and engineering texts, the vocabulary burden required of learners is lowered and as a result, learners benefit from such instruction. Therefore, well-chosen specialized vocabulary lists can give learners the best results with the least effort.

To date, it is not known how large this specialized newspaper vocabulary might be and what kinds of words would form it. Thus, the primary purpose of this study is to identify the specialized vocabulary of newspapers. Three research questions will be investigated:

1. How many word families make up a specialised vocabulary of newspaper texts (hereinafter, the Newspaper Word List)?
2. What percentage of the tokens in newspaper texts does the Newspaper Word List cover?
3. How often do the words in the Newspaper Word List occur in different newspapers and different news divisions?

Materials and Methods

Computer Programs

The analyses in this study were performed with the vocabulary analysis program *Range32H* (Heatley & Nation, 2006) in order to count and create a list of proper names and Newspaper Words. The program uses two Baseword lists: Baseword one is the first 1,000 words and Baseword two is the second 1,000 words of the GSL.

A weakness of *Range* is that the program cannot distinguish *ward* (a family name) from *ward* (a section of a hospital). This problem was addressed by looking at the context in which the word was embedded. Similarly, *aid* as a verb is not distinguished from *aid* as a noun; however, this kind of polysemic use was not considered a problem because the learning burden of the noun form is very small if the verb use is known. Many English verbs may commonly be also used as nouns.

Determining a Unit of Counting Word Families

In this study, the word family is used as the unit when counting words. The level of word family used here is composed of a base form together with its inflected forms and derived forms as described in Level 6 of Bauer and Nation's scale (1993). A word family represents a group of words whose forms and meanings are closely related to each other and which can be understood with little or no extra learning when one or more of the members is already well known to a learner. Thus, word types from the same word family are counted as the same word. Both American and British spellings are counted in the same family. For example, *analyse* and *analyze* are counted in the family *analyse*. The main justification for the use of the word family is that it best represents the kind of knowledge needed when meeting words in reading, and the goal of this study is to examine the vocabulary needed for reading newspapers. Table 1 illustrates how large word families can be. Each word in italics is the most frequently occurring form in that family in the Newspaper Corpus.

Compiling the Newspaper Corpus

The news texts used in this study were obtained from the Internet Public Library drawing on texts published from 23 February to 23 May 2006. All texts were obtained in electronic form. The dates of the reports and the names of the reporters and the newspapers were removed.

In making the Newspaper Corpus, four principles were followed. The first principle was that newspapers for the Newspaper Corpus of English (hereinafter, the Newspaper Corpus) had to represent the kinds of English newspapers that native speakers of English would typically read. Three newspapers were chosen: *The Dominion Post* from New Zealand, *The Independent* from the United Kingdom, and *The New York Times* from the United States of America. These are representatives of high quality English newspapers in these three English-speaking countries. Though the sensational tabloids

Table 1. Examples of Three Word Families in the Newspaper Corpus

FINANCE	SECURE	INVEST
FINANCES	SECURES	INVESTS
FINANCED	SECURED	INVESTED
FINANCING	SECURING	INVESTING
FINANCIAL	SECURELY	*INVESTMENT*
FINANCIALLY	*SECURITY*	INVESTMENTS
FINANCIER	SECURITIES	INVESTOR
FINANCIERS	UNSECURED	INVESTORS
	INSECURE	REINVEST
	INSECURITY	REINVESTS
	INSECURITIES	REINVESTED
		REINVESTING
		REINVESTMENT

are widely read, they were not selected because the Newspaper Word List is intended to help learners study English in an intensive course, often with the goal of going to university.

The second principle was that the corpus had to be large enough in order to allow the lower frequency candidates for a specialized vocabulary of newspapers to have a reasonable number of occurrences (Kennedy, 1998; Leech, 1987; Sinclair, 1991). A corpus of 579,849 running words proved to be large enough to obtain a minimum frequency of at least 20 occurrences of each candidate word.

The third principle was that the corpus had to contain approximately equal-sized, representative sections of each newspaper in order to measure the range of occurrence of words. Range is vital because lexical items that will be met when reading different sections of a newspaper and different newspapers should be selected.

The Newspaper Corpus consisted of 12 sections, namely the four main news divisions (Business, International, National, and Sports) from three newspapers (i.e., *The Dominion Post*, *The Independent* and *The New York Times*). Table 2 provides data concerning the size of the 12 news sections each counted by the *Range* program.

Table 2. Tokens in each of the 12 News Sections

News division	*The Dominion Post*	*The Independent*	*The New York Times*	Total
National	48,270	47,816	48,527	144,613
Business	47,361	47,922	48,549	143,832
Sports	48,827	49,020	48,750	146,597
International	48,594	47,848	48,365	144,807
Total	193,052	192,606	194,191	579,849

As shown in Table 2, the National news texts in *The Dominion Post* contained 48,270 tokens and the combined National news texts from the three newspapers totaled 144,613 tokens. The four sections in the Dominion Post contained a total of 193,052 tokens. On average, each of the 12 news sections contained 48,300 tokens, each of the four news divisions 144,900 tokens, and each of the three newspapers a total of 193,000 tokens. Each of the 12 sections was of roughly equal size in order to obtain comparable statistical data from the various sections, and accordingly the frequency of the words was not biased by the size of each section (Leech, 1987; Sinclair, 1991).

The fourth principle was that texts in the corpus should be representative of news text types. Three conditions were considered. First, texts for the corpus should be selected from a news reportage category rather than from editorials, book and movie reviews, or advertisements because reporting news is a more typical function of a newspaper. Second, a large variety of news texts written by a large number of reporters should be included in the corpus. Third, the texts should be whole texts rather than a collection of partial texts, and relatively long texts need to be chosen in order to obtain specialized words with a higher frequency. Sampling whole texts gives topic-related words more opportunity to occur, though marked differences of individual writing style or topic might appear (Sinclair, 1991). Accordingly, the 868 texts comprising the Newspaper Corpus (see Table 3) were whole texts from reportage and 844 texts (97.2%) were between 200 and 2,000 words long; the shortest text was 131 words long and the longest was 5,054 words. A balance between short and long texts, and a balance in size between different news divisions were achieved where possible as shown in Table 3.

Table 3. Number of Texts in Each News Division

News division	Number of texts
National	221
Business	211
Sports	215
International	221
Total	868

Each news division contained 217 texts on average. Care was taken when compiling the corpus to ensure that texts were not repeated in different newspapers. There is a high risk of this occurring because wire services like Reuters and API provide news to newspapers all over the world. A very large amount of work was involved in collecting the corpus, as each of the 868 texts had to be downloaded one by one, checked to avoid duplication of texts, and edited for misspellings, spelling variations, foreign words, and hyphenated words (for details, see the section *Editing the News Texts* below).

Editing the News Texts

The news texts were edited to make them computer readable and to avoid counting problems. After that, the texts were saved in the *plain text* format in order to make them suitable for analysis by the *Range* program.

Hyphenated Words: For hyphenated words with a deducible meaning from the meaning of each constituent (e.g., *large-scale*, *wide-bodied*, and *anti-war*), a space on each side of the hyphen was inserted using the Find and Replace function on the computer. This is because in terms of counting the occurrences of each word and measuring the vocabulary load of the text, it is better to count each constituent separately, as this avoids inflating the number of different word types. Where it is better to keep a hyphen in order to maintain the meaning of the whole, a hyphenated word was changed into one lexical item without a hyphen (e.g., *preemptive*, *email*, and *hiphop*).

Foreign Letters: When foreign words in *Word* format were saved in *plain text* format, the pre-existing *Word Document* format was lost, and this created problems in counting words. For example, *Löffler* was initially counted wrongly as two items, *L?* and *ffler*. For this reason, foreign letters, for example *ö*, *é* and *á* as in *Löffler*, *René,* and *Chávez,* were replaced with the English letters *o*, *e* and *a*, respectively.

Various Word Forms with the Same Meaning: In the case of varying word forms with the same meaning such as *per cent* and *percent*, *per cent* was replaced with *percent*. Otherwise, *per cent* and *percent* would be counted as three items, *per*, *cent* and *percent*.

Names with an Apostrophe: Words written with an apostrophe, such as *Shi'ite* and *Fa'atau,* were rewritten as *Shiite* and *Faatau* in order to avoid each being counted as two items.

Setting up Criteria for Identifying Specialized Words for Reading Newspapers

Three criteria were set up in order to ensure that the words identified were specialized vocabulary for reading newspapers.

Special Purpose Vocabulary: The first criterion was that newspaper words must be special purpose vocabulary. This meant that they could not be part of the high-frequency 2,000 words of English as defined by West's (1953) *General Service List of English Words* (GSL). In addition, no proper names were included on the list. One reason for choosing the GSL was to make the data comparable with the Academic Word List which also assumes knowledge of the GSL.

Wide Range: The second criterion of range had the highest priority because words should occur in a wide range of different news texts. In this study, range was measured by (a) determining the number of news divisions in which each candidate word occurred and (b) by counting the number of news sections across the three newspapers and the four news divisions in which the word was found (e.g., *The Dominion Post*'s National news section and *The Independent's National*'s news section). Thus, Newspaper Words must occur in all four news divisions of the corpus, and 6 or more of the 12 smaller news sections. Because the primary aim of the study is to create a list of the most useful Newspaper Words rather than create a complete list of Newspaper Words, a range of 6 or more out of 12 was considered sufficient for identifying Newspaper Words.

High frequency: The third criterion was the total frequency with which the candidate words occurred in the Newspaper Corpus. Frequency is important but not foremost because creating a word list based on frequency alone allows a bias towards longer texts and topic-related words. In this study, Newspaper Words must occur 20 times or more in at least 6 out of the 12 sections in the corpus. The frequency cutoff point of 20 occurrences was chosen because in terms of practicality, 20 examples provide enough examples for a useful concordance analysis of an item (Leech, 1987; Sinclair, 1991).

Making a List of Proper Names

In order to prevent frequently occurring proper names from being selected as Newspaper Words, a list of proper names was created by examining the words with a frequency of 20 or more occurring outside the GSL 2,000 words. The list of proper names included personal names (e.g., *Mary* and *David*), country names (e.g., *New Zealand* and *Britain*) and organization names (e.g., *Delta Air Lines* and *Duke University*). Abbreviations, such as NZQA, EU, and FIFA, were generally included in the list of proper names.

Certain items, such as *hawk* (as in Black Hawk helicopter and a kind of bird), *ward* (as in *Martin Ward* and a kind of room**),** *mount* (as in *Mount Tambora* and go up), and *range* (as in *Tararua Range* and widespread), were used as both a proper name and an ordinary item in the corpus. Items occurring more frequently as a proper name than an ordinary item in the Newspaper Corpus (e.g., *Hawk* and *Ward*) were placed in the list of proper names.

In order to make a list of proper names, all word types with a frequency of 20 or more that did not occur in the GSL were examined and a decision was made about which would be put into the list of proper names.

Note that after making a list of proper names, there are three Baseword lists to run with the *Range* program: Baseword one and two are the first 1,000 words and the second 1,000 words of the GSL; Baseword three is a list of proper names.

Creating a List of Specialized Words for Reading Newspapers

The following steps were taken in order to identify all the word types outside the 2,000 words of the GSL and the list of proper names, to decide whether they met the criteria for identifying specialized words and thus to select potential candidates for a list of Newspaper Words.

First, word types occurring outside the three Baseword lists were identified by running the four news divisions of the Newspaper Corpus through the *Range* program. Second, 1,012 word types occurring in all four news divisions (Business, National, Sports, and International) and not in the GSL or proper name list were identified. Third, the 1,012 types were organized into 733 word families using the *Copy* function in the *Range* program drawing on Nation's fourteen 1,000-word lists from the British National Corpus. These lists have been carefully created and are a reliable source of word families. Fourth, by running the 12 news sections (see Table 2) of the corpus through the *Range* program, 523 word families with a range of 6 or more out of 12 and a frequency of 20 or more occurrences were identified.

Finally, word types occurring in only two or three news divisions were examined in order to determine whether counting word families rather than word types would allow more words to become candidates for the Newspaper Word List. This resulted in 65 word families (e.g., *adequate*, *bonus*, *consult*, and *score*) being added to the list, giving a total of 588 word families.

Results

The Newspaper Word List and its Text Coverage

From a corpus of 579,849 tokens, 588 word families (Appendix 1) were identified as specialized words for reading newspapers using the criteria of range and frequency. Table 4 shows how much of the Newspaper Corpus was covered by the GSL lists and the Newspaper Word List, and how many families in each list occurred in the corpus.

Table 4. Text Coverage and Number of Families in Each List

Word list (Number of families in the list)	Coverage of the Newspaper Corpus	Number of families occurring in the Newspaper Corpus
Newspaper Word List (588 families)	6.8%	588
Second 1,000 GSL (991 families)	5.5%	937
First 1,000 GSL (998 families)	74.2%	996
Total (2,577 families)	86.5%	2,521

Table 4 shows that the NWL covered 6.8% of the tokens in the corpus. This is higher than the 5.5% coverage of the second 1,000 GSL of the corpus. This contrast is even more striking when we consider that the total number of word families in the NWL (588 families) is much smaller than the 937 families occurring in the second 1,000 of the GSL.

The first 2,000 words of the GSL and the 588 newspaper word families in the corpus provide coverage of 86.5% of the running words in the corpus. This is a high degree of coverage with a relatively small number of words.

The NWL Coverage of National, Business, Sports, and International News Divisions

Table 5 shows a comparison of the coverage of the four news divisions by the NWL, the GSL, and proper names.

Table 5. Text Coverage of the Four News Divisions by Each List

Coverage	National news	Business news	Sports news	International news
NWL	6.7%	8.3%	5.1%	7.1%
Second 1,000 GSL	5.9%	5.5%	5.5%	5.3%
First 1,000 GSL	74.8%	74.7%	74.9%	72.5%
Proper names	4.9%	4.7%	7.4%	7.0%
Total coverage	92.3%	93.2%	92.9%	91.9%

Note: Proper names are treated as known words because proper names are easily understood from the context or are already known to students (Hwang & Nation, 1989).

The NWL coverage of the Business news division is the highest (8.3%) and the coverage of the Sports news division is the lowest (5.1%). A factor contributing to the high coverage by the NWL of the Business news division is that some word families occurred extremely frequently in the Business news division, but were of much lower frequency in the Sports news as seen in Table 6.

The NWL coverage of the National and International news divisions is similar (6.7% and 7.1%, respectively). Within the most frequent top 10 words, three word families: *percent*, *issue*, and *secure* (*security* is the most frequent type) were included in both National and International news. The other 7 families out of the top 10 included *labour*, *fund*, *drug*, *job*, *sex*, *investigation*, and *port* in the National news; and *military*, *protest*, *bomb*, *prime*, *terror*, *major*, and *region* in the International news.

Table 6. A Comparison of the Number of Occurrences of 18 Word Families in the Business News and Sports News Sections

Word families	Number of Occurrences in Business News	Number of Occurrences in Sports News
PERCENT	707	29
INVEST	456	9
FINANCE	195	18
EXECUTIVE	184	33
FUND	178	13
ENERGY	132	4
CONSUME	124	1
REGULATE	122	8
BID	112	32
COMMISSION	106	12
ISSUE	95	33
PENSION	93	1
ANALYSE	91	13
CORPORATE	84	3
SHAREHOLDER	81	1
INCOME	80	2
EXPORT	76	2
REVENUE	75	6

The second 1,000 words of the GSL had very similar coverage (about 5.5%) in all four news divisions, but the first 1,000 words of the International news division had slightly lower coverage (72.5%) than the others (around 74.8%).

The proper name coverage of the Sports news division was the highest (7.4%), while the Business news coverage was the lowest (4.7%). Because players' skills, team performances, and new players' names are mentioned frequently in the Sports News section, proper names occurred more frequently in this section than in any other news division. For this reason, if the Sports news division is excluded from the Newspaper Corpus, the NWL coverage of the remaining combined texts rises to 7.4%. The smallest coverage by the NWL of the Sports news division was balanced by the biggest coverage of proper names. Thus, the total coverage of each of the four news divisions by the four combined lists was very similar, between 91.9% and 93.2%.

The NWL Coverage of each Newspaper of the Three Countries

The coverage of the three newspaper corpora provided by the three word lists is shown in Table 7.

Table 7. Coverage of Each Newspaper by the Three Word Lists

	The Dominion Post	*The Independent*	*The New York Times*
NWL coverage	7.3%	6.5%	6.6%
Second 1,000 GSL coverage	5.7%	5.6%	5.2%
First 1,000 GSL coverage	73.3%	74.8%	74.5%
Total	86.3%	86.9%	86.3%

As shown in Table 7, there is little difference in coverage provided by all three lists for the three newspapers. The NWL is clearly an international list and it could be expected to work well with other similar newspapers. Eight word families: *executive*, *final*, *invest*, *issue*, *major*, *percent*, *secure*, and *team*, were within the most frequent 20 words in the three newspapers.

Range of the 588 Families of the NWL

Table 8 shows how many of the 588 families of the NWL occurred in 6 or more of the 12 sections of the Newspaper Corpus.

One hundred thirty-eight (24%) of the 588 word families occurred in all 12 news sections, and 567 families (96%) occurred in 7 or more sections. The wide range of the 588 families indicates that the list is likely to apply well to other similar quality newspapers.

Table 8. Cumulative Number and Their Percent of 588 Word Families in Sections 6 to 12

Number of news sections	Number of NWL families	Cumulative number	Cumulative percent of families
12	138	138	24%
11	127	265	45%
10	110	375	64%
9	94	469	80%
8	60	529	90%
7	38	567	96%
6	21	588	100%

Evaluation of the Newspaper Word List

A frequency-based word list made from a particular corpus will provide reasonably high coverage of that corpus. In order to test whether the NWL provides good coverage of a different newspaper corpus, the newspaper sections of the Frown Corpus and the FLOB Corpus, both containing material written in the early 1990s, were chosen. These are relatively new compared with similarly structured but older Brown and LOB corpora compiled over 30 years earlier.

From the Frown and FLOB corpora, three categories (reportage, editorials, and reviews) were selected for making three corpora to test the NWL: (a) a reportage news corpus, (b) a reportage and editorials combined corpus, and (c) a reportage, editorials, and reviews combined corpus. The reportage news corpus contains 88 texts, amounting to 180,170 tokens; the reportage and editorials combined corpus, 142 texts (292,048 tokens); the reportage, editorials, and reviews combined corpus, 176 texts (362,584 tokens). Note that all texts of the Newspaper Corpus used in this study would be classified as reportage in the Frown and FLOB corpora. Table 9 shows a comparison of the three news corpora from the Frown and FLOB.

Table 9. Number of NWL Families and Their Coverage in Various Corpora

Corpus (Tokens)	NWL Families	NWL Coverage
Reportage news (180,170 tokens)	577	6.0%
Reportage and editorial combined (292,048 tokens)	582	6.0%
Reportage, editorial, and reviews combined (362,584 tokens)	582	5.7%

As shown in Table 9, the coverage by the NWL of each of the three news corpora was similar at 6.0%, 6.0%, and 5.7%. This indicates that the NWL also works well with editorials and reviews sections. The 6.0% is slightly lower than the 6.8% coverage of the Newspaper Corpus compiled for this study. Five hundred seventy-seven out of the 588 newspaper families occurred in the reportage news texts, quite a lot given the smaller corpus size. The 11 NWL families which did not occur in the reportage corpus were *cellphone*, *detention*, *email*, *enrich* (*enrichment* is the most frequent type), *enroll*, *flu*, *immigrate* (*immigration* is the most frequent type), *internet*, *refine* (*refinery* is the most frequent word type), *virus*, and *website*. Such items are likely to be affected by the age of the corpus because there is more than a 15 year difference in age between the Newspaper corpus and the Frown and FLOB combined corpus. The items occurring frequently are also affected by new or emerging topics such as email and bird flu. Six families out of the 11 did not occur in the reportage and editorial combined news: *cellphone*, *email*, *enrich*, *flu*, *internet*, and *website*.

Newspaper texts from the Brown and LOB corpora were also examined to determine how much of the text the NWL covered. Because the Brown and LOB corpora were compiled in the 1960s (almost 50 years ago), the coverage of the NWL was around 5.1%, suggesting that the NWL is affected by current issues and needs to be updated periodically.

Conclusions

In the Newspaper Corpus of 579,849 tokens, 588 word families were classified as Newspaper Words. The list of 588 families is a specialized vocabulary which provides a high coverage of newspaper texts. It accounted for 6.8% of the tokens in the corpus. One strength of the Newspaper Word List is that the 588 families are a much smaller group than 937 families occurring in the second 1,000 GSL, but the coverage of the NWL is 1.3% better than the coverage by the second 1,000 GSL.

When combining the coverage of the NWL, GSL, and proper names, the coverage of the corpus comes to 93% coverage. Though this is lower than the 98% target coverage criterion specified by Hu and Nation (2000), this is very high and thus the NWL can provide second language learners who want to read English newspapers with a way to focus their vocabulary studies.

The NWL can add to the number of high frequency words that could be directly taught in class time and that deserve deliberate study by learners. It is important to remember that vocabulary learning should take place in a balance of activities, covering not only meaning-focused activities but also language-focused and fluency development activities. For maximum benefit, learners should read more related stories than unrelated stories. Following the same story through the several issues of the newspaper is an effective way of helping learners review the vocabulary learned previously (Hwang & Nation, 1989; Schmitt & Carter, 2000). The NWL would be useful for teachers of English for specific purposes (ESP) who are interested in designing a vocabulary course for foreign language learners who wish to read newspapers as soon as possible.

In future studies, firstly, it may be desirable to collect data for more than 3 months and compile a bigger corpus covering a wider range of sections so that the NWL could be more widely applied in each newspaper. Secondly, the NWL may need to be updated every 5 to 7 years as it is partly influenced by current world events.

Acknowledgements

I would like to thank, Professor Paul Nation of the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand for his helpful comments and suggestions in preparing this article.

Mihwa Chung teaches at Korea University of Sejong in Korea. Her doctoral thesis examined a range of ways of distinguishing technical terms from other words in English for Specific Purposes. Her current teaching and research interests include teaching and learning vocabulary (in particular, technical terms and specialized vocabularies), corpus analysis, reading courses and speed reading courses.

References

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253-279.

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

Francis, W. N., & Kučera, H. (1979). *Manual of information to accompany a standard corpus of present day edited American English, for use with digital computers*. Providence, RI: Brown University.

Heatley, A., & Nation, P. (2006). Range32H (computer software). Wellington, New Zealand: Victoria University of Wellington.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8*(2), 689-696.

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

Hundt, M., Sand, A., & Siemund, R. (1999). *Manual of information to accompany the Freiburg-LOB Corpus of British English (FLOB).* Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg.

Hundt, M., Sand, A., & Skandera, P. (1999). *Manual of information to accompany the Freiburg-Brown Corpus of American English (Frown).* Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg.

Hwang K., & Nation, P. (1989). Reducing the vocabulary load and encouraging vocabulary learning through reading newspapers. *Reading in a Foreign Language*, *6*(1), 323-335.

Johansson, S. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers.* Oslo: University of Oslo.

Kennedy, G. (1998). *An introduction to corpus linguistics.* London: Longman.

Klinmanee, N., & Sopprasong, L. (1997). Bridging the vocabulary gap between secondary school and university: A Thai case study. *Guidelines*, *19*(1), 1-10.

Leech, G. (1987). General introduction. In R. Garside, G. Leech, & G. Sampson (Eds.), *The computational analysis of English: A corpus-based approach* (pp. 1-15)*.* London: Longman.

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

Nation, P., & Coady, J. (1988). Vocabulary and reading. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 97-110). London: Longman.

Schmitt, N., & Carter, R. (2000). The lexical advantages of narrow reading for second language learners. *TESOL Journal*, *9*(1), 4-9.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Ward, J. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language, 12*(2), 309-323.

West, M. (1953). *A general service list of English words.* London: Longman.

Appendix 1: Newspaper Words in 10 Sublists

The NWL is grouped into 10 sublists, and each sublist includes 60 words, except for Sublist 10 which includes 48. Groups of 60 break the learning task into units of a manageable size for a short-term course. The grouping of the sublists is based on range and frequency, and range is given precedence over frequency. Sublist 1 contains words which are of the widest range and the words in Sublist 10 are of the narrowest range among the 10 sublists.

The most frequently occurring member of each word family in the NWL is displayed in the list. Figures indicate the sublist of the NWL. For example, *abandoned* is the most frequent type of the members of the family *abandon*, and this family occurs in Sublist 8 of the NWL. Note that both American and British spellings are included in the word families (e.g., both *rumors* and *rumours* are included in the family *rumor*). Three prefixes (*pre*-, *ex*-, *pro*-) are included in the list because they are frequently used to make words, are predictable in meanings as in *pre-season, pre trial, pre-match, ex-adviser, ex-offenders, ex-employee, pro-democracy campaign, pro-Palestinian,* and *pro-life groups*, and are within the Level 6 affixes (Bauer & Nation, 1993) used for making a list of Newspaper words in this study.

Headwords of the Newspaper Word List in 10 Sublists

abandoned	8
abuse	5
academic	3
access	3
accompanied	6
accurate	6
achieve	1
acknowledged	2
acquired	2
adequate	9

adjusted	3
administration	3
affected	2
agenda	3
aggressive	2
aid	3
airline	8
airport	4
alarm	5
alcohol	8

allegations	5
alliance	5
allies	4
alongside	5
alternative	4
amazing	8
amid	3
analysts	1
announced	1
annual	3

anticipated	5
apartment	7
apparently	2
appeal	3
approach	1
appropriate	3
area	1
aspects	7
assembly	6
asserts	9
assessment	2
assets	7
assistant	2
assume	3
assured	2
athletic	10
attached	3
attitude	7
attorney	9
authority	1
available	1
awaited	4
award	4
aware	2
bail	10
ban	3
beach	4
behalf	5
benefits	3
bet	9
bid	1
bomb	8
bond	9
bonus	9
boom	4
boost	7
boss	3
bounce	7
brand	5
breach	4
brewers	8
brief	5
broker	9
budget	3
bullet	8
burden	9
bureau	10
cabinet	7
campaign	1
cancer	10
candidate	3
capable	3
capacity	7
captured	5
career	3
cash	3
cast	6
casualties	10
category	6
celebrated	6
cellphone	10
challenge	1
champion	5
chancellor	10
channel	7
chaos	7
chase	8
chip	8
circuit	8
circumstances	9
cited	4
clash	7
classic	10
climate	6
clinical	7
collapse	4
colleagues	7
column	9
combat	9
comment	1
commission	3
committed	1
communications	6
community	1
compensation	6
complex	6
comply	5
compound	10
computer	6
conceded	7
concentrate	8
concert	10
concluded	6
conclusion	8
condemned	9
conducted	3
conference	1
confirmed	3
conflict	8
confrontation	7
congress	10
consecutive	10
consent	9
consequences	3
conservative	6

considerable	6
consistent	4
construction	2
consultant	3
consumers	7
contact	7
contend	6
contest	9
contract	1
contrast	6
contributed	2
controversy	2
convention	8
converted	10
convicted	8
convinced	6
cooperation	9
cope	10
corporate	3
counter	8
counterparts	8
county	10
couple	2
create	1
credit	2
crew	6
crisis	3
criticised	2
criticism	6
crucial	3
culture	4
curb	9
deadline	4
debate	2
debut	9
decade	1
declined	2
defence	1
defendants	9
deficit	5
definitely	4
definition	5
deliberately	6
demonstrations	4
denied	1
departure	8
depressed	8
deputy	5
designed	2
desperate	8
despite	1
detention	10
disabled	9
disaster	5
disclose	7
discount	7
discrimination	10
display	6
dispute	2
distinctive	7
distribution	6
dividend	10
document	5
domestic	4
dominated	4
donations	10
draft	7
dramatic	6
drug	3
echoed	7
editor	5
element	8
eliminate	8
email	10
embarrassed	6
embrace	5
emerged	2
emotional	4
emphasis	5
enable	5
endorsed	9
energy	5
enforcement	4
enormous	8
enrichment	9
enrolled	10
ensure	4
entitled	8
environment	3
equipment	5
era	8
errors	6
erupted	9
established	2
estimated	2
evaluate	9
eventually	2
evidence	1
ex-	5
exceed	8
exclude	9
executed	6
executive	1
expand	2
experts	4

exports	7
exposed	4
extract	10
facilities	4
factor	4
feature	4
federal	5
federation	4
feeding	3
fees	7
final	1
financial	1
fines	6
fled	9
flexible	8
flu	9
focus	1
forecast	7
founder	5
franchise	10
frustrated	3
fuel	1
fund	3
fundamental	5
generation	1
giant	2
global	5
goal	3
golf	10
goods	6
grab	8
grade	8
graduate	10
grant	2
guarantee	4

guidelines	9
guys	7
hail	9
halt	4
haul	8
headlines	7
headquarters	3
height	8
highlighted	7
huge	1
identified	3
ignore	4
image	2
immigration	9
impact	2
implications	7
import	5
imposed	2
impression	4
incident	5
income	7
incorporated	8
indicated	2
individual	2
inevitable	7
infection	9
inflation	8
infrastructure	9
initial	2
initiative	6
injury	3
insisted	3
inspector	6
inspired	9
installed	10

instance	6
institute	1
intelligence	7
intense	2
interim	10
internal	4
internet	3
interview	3
investigation	1
investment	1
involved	1
isolated	6
issue	1
items	9
jail	5
job	1
journal	10
journalists	2
junior	9
jury	10
justify	5
keen	10
kids	6
label	8
labour	5
lane	8
lap	9
launched	3
league	7
leaking	9
legal	3
licence	8
link	5
lobby	9
location	4

Ltd.	10
magazine	7
maintaining	2
major	1
margin	5
massive	6
maximum	3
media	3
medical	2
mental	7
military	3
minimum	7
ministry	8
minor	2
mirror	7
mission	7
mobile	6
monitors	2
mood	8
motivated	4
mount	4
mutual	8
negative	4
negotiations	1
network	3
nomination	9
normal	2
obligations	8
obstacle	8
obviously	4
occupation	9
occupied	4
occurred	6
odd	6
olympic	5
opponent	4
optimistic	5
option	1
outcome	3
overall	2
overnight	6
pace	4
panel	6
panic	5
participation	6
partner	1
passion	7
patients	10
peak	9
penalty	3
pension	9
percent	1
period	1
personality	10
personnel	6
physical	6
pit	10
pledged	7
plunged	9
plus	2
PM (Prime Minister)	8
policy	3
polls	8
port	3
posed	7
positive	3
potential	1
pre-	5
predicted	2
premier	7
previous	1
primary	6
prime	1
prince	10
principal	7
principle	8
priority	7
pro-	8
proceedings	2
process	1
professional	2
profile	4
project	1
prominent	9
promote	4
prop	10
prosecution	5
prospects	2
protests	3
province	7
provoke	8
publisher	5
purchase	10
pursue	5
quit	7
quoted	7
raid	7
rally	9
range	1
reaction	3
recalled	7
recovery	2
recruiting	2
refinery	10

regain	8
regime	7
region	3
register	6
regulation	7
rejected	3
release	1
reluctant	5
relying	2
remote	9
removed	4
required	1
research	5
residents	5
resolve	9
resort	9
resource	5
respond	2
response	4
restrictions	4
resume	9
retain	3
revealed	2
revenue	5
reverse	2
revised	8
revolution	9
riot	8
role	1
route	4
routine	3
rumors	6
sanctions	7
scandal	7
scared	6

scheduled	4
scored	1
section	4
security	1
seeking	1
select	2
senior	1
series	1
session	4
sex	8
shareholders	9
shift	4
significant	1
similar	1
site	1
slim	10
slumped	10
smart	6
soared	7
sole	5
source	2
sparked	10
specific	4
speculation	4
spokesman	1
spokeswoman	6
sponsored	4
spurs	9
stability	3
stake	5
statistics	4
status	4
strain	6
strategy	1
stress	4

structure	2
stunned	5
style	4
suburb	9
successor	10
sufficient	9
sum	9
super	7
supreme	9
surge	8
surgery	7
survive	2
suspended	6
sustained	6
switch	6
symbol	6
tackle	5
tactics	6
tank	9
tape	8
target	3
task	7
team	1
technical	4
technology	3
teenagers	6
tensions	4
territory	7
terrorist	5
testified	9
testimony	8
text	10
theme	8
tiny	7
toll	9

traditional	2
traffic	8
transfer	2
transformed	5
transmission	9
transport	5
trend	7
tribunal	10
triple	10
ultimately	2
undermine	6
unique	8
urgent	8
utility	8
variable	5
vast	5
venture	6
verdict	10
version	4
veteran	6
veto	10
vice	6
victim	5
video	5
virus	9
vital	4
volume	6
volunteers	10
vulnerable	8
watchdog	8
website	5
widespread	5
withdrawal	2
zone	2