A Specialised Word List for Reading Newspapers

Mihwa Chung
(Korea University of Seochang)

I. Introduction

Reading is one of the most common and important ways of learning another language and one of the major types of reading material is newspapers. Newspapers are often used in reading classes in order to develop reading skills and expand vocabulary knowledge (Hwang and Nation, 1989; Klinamanee and Sopprasong, 1997). Many learners however find it difficult to read unsimplified newspaper texts. There are a number of factors which contribute to difficulty in reading, but vocabulary knowledge has consistently been found to be the most significant and strongest factor (Nation and Coady, 1988; Hirsh and Nation, 1992), which affects success in understanding. Research by Hu and Nation (2000) shows that knowledge at least 95% of the total words (tokens or running words) in a text is the minimum required for adequate reading comprehension and to guess unknown words from context. Hirsh and Nation (1992) suggests that a vocabulary of around 5000 word families can allow pleasurable reading of a text.

There are high frequency words in certain kinds of text, and they provi good coverage for the same kind of text type. If teachers narrow the focus on teaching vocabulary to areas such as vocabulary for reading newspapers and academic texts, learners would benefit from such a course. A good example of a specialised vocabulary is the Academic Word List (Coxhead, 2000) developed for reading academic texts. The Academic Word List consists of 570 word families, covering about 10% of the running words in a wide range of academic texts - ery high coverage.

Motivated by the Academic Word List, this study is designed to examine the specialised vocabulary of newspapers. To date, it is not known how large this specialised newspaper vocabulary might be and what kinds of words would form this vocabulary.

The research questions are:

1. How many word families make up a specialised vocabulary of newspaper texts (hereafter, the Newspaper Word List)?
2. What percent of the tokens in newspaper texts does the Newspaper Word List cover?

II. Computer programs

Much of the analysis in this study was done using three computer programs. Firstly, the vocabulary analysis program called ¡*Range*¡ which was developed by Paul Nation and Alex Heatley at Victoria University of Wellington was used to count, select, and create a list of Newspaper Words. Secondly, ¡*Excel*¡ was used to sort and calculate the number of tokens in four news divisions and twelve news sections, and keep the details of each text, such as the length of each text. Finally, the ¡*MS-Word*¡ program was used to edit texts to make computer readable data. This involved deleting the date published, names of writers, and apostrophes in names as in O'Brien and Shi'ite in order to avoid counting problems. After that, the texts were saved in the *Plain Text* format in order to make them suitable for the *Range* program.

III. Procedure

1. Compiling the Newspaper Corpus

The news texts for this study were collected from the Internet Public Library (http://www.ipl.org/) drawing on texts published from 23 February to 23 May 2006. All texts were 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 (hereafter, the Newspaper Corpus) had to represent the kinds of English newspapers which native speakers of English would typically read.

The second principle was that the corpus had to be large enough in order to allow the lower frequency candidates for a specialised vocabulary of newspapers to have a reasonable number of occurrences (Leech, 1987; Kennedy, 1998; Sinclair, 1991).

The third principle was that the Newspaper Corpus had to contain roughly equally sized, representative sections of each newspaper in orde ɔ measure the range of occurrences of words. The Newspaper Corpus consisted of twelve sections, namely the four main news divisions (Business, International, National and Sports) from three different newspapers (the Dominion Post, the Independent and the New York Times). Table 1 provides the data about the size of twelve news sections each counted by the *Range* program.

News division	The Dominion Post	The Independent	The New York Times	Total
National	48,270	47,816	48,527	144,613
Business	47,631	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

Table 1: Tokens in each of the twelve news sections

Each of the twelve sections had to be of roughly equal size so that the frequency of the words was not biased by the size of each section.

The fourth principle was that texts in the corpus should be representative of news text types. Three conditions were considered. (1) Texts for the Newspaper Corpus should be selected from a news reportage category rather than from editorials, book and movie reviews or advertisements because reporting news is considered to be a more typical function of a newspaper. (2) News texts should include a large variety of news texts written by lots of reporters. (3) The texts should be whole texts rather than a collection of partial texts. A balance between short and long texts, and a balance in size between different news divisions were achieved where possible as shown in Table 2. Each news division contained 217 texts on average.

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

Table 2: Number of texts in each news division

2. Setting up criteria for identifying specialized words for reading newspapers

There are three criteria to make sure that the words identified really are a specialised vocabulary for reading newspapers (Newspaper Words).

1) *Special purpose vocabulary*: Newspaper Words must be outside the first 2,000 words of English, based on a General Service List of English Words (GSL)

by West (1953) and not include proper names.

2) *Wide range*: The criterion of range has the first priority in selecting a specialised vocabulary because words should occur in a wide range of different news texts. In this study, Newspaper Words must occur in all four news divisions of the corpus: Business, Sports, National and International, and six or more of the twelve smaller news sections.
3) *High frequency*: 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 with a frequency of twenty or more in the corpus.

3. 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 outside the GSL 2,000 words. The list of proper names included personal names (Mary and David), country names (New Zealand and Britain) and organization names (Delta Air Lines and Duke University). Abbreviations were generally included in the list of proper names, such as, NZQA, EU and FIFA

4. Determining what items should be counted as words

In this study, the word family is used as the unit when counting words. The level of word family used here consists 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 is a group of words whose meanings are closely connected with each other and which can be understood with little or no extra learning once the learner already knows one or more of the members. Using word families has the benefit of providing low frequency candidates with a reasonable number of occurrences. For these reasons, word types from the same word family are counted as the same word. Table 3 shows examples of word families. The words in italics are the most frequent type in that family occurring 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

Table 3: Examples of word families in the Newspaper Words

5. Applying the criteria to create a list of specialized words for reading newspapers

1) By running the Range program, select all the word families outside the first 2000 words of the General Service List of English Words (GSL) by West (1953).

2) Decide whether they fit the criteria for selecting specialised words and then put potential candidates on a list of specialised words. To be selected as specialised word families for reading newspapers, two criteria must be met. First, the word families must occur in all four news divisions (Business, National, Sports and International). Second, the word families must occur with a frequency of 20 or more and a range of six or more out of twelve sections.

IV. The Newspaper Word List and Its Text Coverage

From a corpus of 579,849 tokens, 588 word families were classified as specialised 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 (NWL), and how many families in each list occurred in the Newspaper Corpus.

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: Coverage and families by the General Service List and the Newspaper Word List

Table 4 shows that the Newspaper Word List covered 6.8% of the tokens in the Newspaper Corpus, and the second 1,000 GSL covered 5.5% of the corpus. It is

pleasing that the coverage by the Newspaper Word List is higher than the coverage by the second 1,000 words, though the total number of word families in the NWL is smaller, 588 families than the 937 occurring in the second 1,000.

V. Conclusion

In the Newspaper Corpus of 579,849 tokens, 588 word families were classified as Newspaper Words. The list of 588 families is a specialised vocabulary which provides a high coverage of newspaper texts, especially, of the reportage news text type. It accounted for 6.8% of the tokens in newspaper texts.

The vocabulary of the Newspaper Word List (NWL) and GSL 2,000 words consisted of 2,521 word families and covered 86.5% of the total words in newspaper texts. The 2,521 families are much smaller than the 5.000 words needed for adequate reading comprehension of an authentic text and this provides a useful shortcut to reading newspapers. When combining the coverage of the first 2,000 words of the GSL, the NWL and proper names, the coverage of the Newspaper Corpus would come to over 95%. Note that the proper names coverage is estimated around 10% of running words of a text and such words are treated as known words to students (Hwang and Nation, 1989).

The Newspaper Words can add to the number of high frequency words that should 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 ma um benefit, learners should read more related stories than unrelated stories (Hwang and Nation, 1989). The Newspaper Vocabulary 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.

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.

Heatley, A., & Nation, P. (1996). *Range* (computer software). Wellington, New Zealand: Victoria University of Wellington. (http://www.vuw.ac.na/lals/staff/paul-nation/).

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.

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 and New York: Longman.

Nation, P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

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 and 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.

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

Brief biodata

Mihwa Chung is a full-time lecturer at Korea University of Seochang, 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 specialised vocabularies), corpus analysis, reading courses and speed reading courses.

Note: Due to the limited space, the Newspaper Word List is not included here. When needed for educational purposes, please email the author.

Mihwa Chung

Office: 82-41-860-1696

E-mail: teresamihwa@korea.ac.kr