Pergamon

System, Vol. 23, No. 1, pp. 35–41, 1995

0346-251X/95 $9.50+0.00

0346–251X(94)00050–6

WHERE WOULD GENERAL SERVICE VOCABULARY STOP AND SPECIAL PURPOSES VOCABULARY BEGIN?

PAUL NATION and HWANG KYONGHO

Victoria University of Wellington, Wellington, New Zealand

Using frequency, text coverage, and range as criteria, this study looks at the dividing line between a general service vocabulary and a special purposes vocabulary. A general service vocabulary gives a good return for learning up to the 2000 word level and after that a special purposes vocabulary gives a better return for learning effort for those learners going on with special interests.

GENERAL SERVICE, SPECIAL PURPOSES AND LOW FREQUENCY VOCABULARY

In non-fiction texts vocabulary can be divided into high frequency (or general service) vocabulary, subtechnical or academic vocabulary, technical vocabulary, and low frequency vocabulary (Nation, 1990: p. 19). There is research evidence to support such a division and it is possible to typify each kind of vocabulary in a particular text or group of texts according to the criteria of frequency, coverage and range.

General service vocabulary

General service vocabulary consists of words that are of high frequency in most uses of the language. It is the essential common core. It includes the most useful function words, like *the*, *of*, *be*, *because*, and *could*, and content words like *stop*, *agree*, *person*, *wide*, and *hardly*. General service words occur frequently across a wide range of texts. The most well-known general service vocabulary is Michael West's (1953) *A General Service List of English Words* which contains around 2000 word families. This list has been the basis of many series of graded readers. Typically the coverage of the *General Service List* is around 75% of the running words in non-fiction texts (Hwang, 1989) and around 90% of the running words in fiction (Hirsh, 1993). "Coverage" refers to the percentage of the running words in a text or corpus that are also in, or covered by, a particular word list. So, if we examined a page of a novel, we might find that there were 300 running words on the page (each repetition of a word already counted would be counted as a new running word), and that around 90% of these words were in the *General Service List*.

There has been criticism of West's list for its size (Engels, 1968) and its age (Richards, 1974). The size criticisms question whether the second 1000 words of the *General Service List* (GSL) are necessary as they usually cover only 4–5% of the running words in non-fiction texts compared to the 70% plus coverage of the first 1000. The age criticisms point out that the report on which

the GSL is based was prepared in the 1930s and thus, because of changes in the language and changes in views of what is the appropriate content for ESL courses, the GSL contains many non-essential words (e.g. *crown* and *canal*) and does not contain words of current high frequency (e.g. *computer*).

Whatever the criticisms of the General Service List, a general service vocabulary is essential for all learners no matter whether they are using English as a foreign or second language, for spoken or written use, or for general or special purposes. The evidence for this comes from the very high coverage that the General Service List provides of a wide range of texts (Hwang and Nation, 1989; Hirsh and Nation, 1992), and the considerable overlap (approximately 80%) of the General Service List and other lists of high frequency words (Table 1).

Learners with a good control of a general service vocabulary need to consider their intended use of English. If they intend to use English for a wide range of informal purposes, they need to continue learning high frequency general service vocabulary. If teachers wanted to plan courses around this, they would look at frequency lists like Thorndike and Lorge (1944), Carroll *et al.* (1971), or Francis and Kučera (1982) for the continuation of the initial general service vocabulary. All of these frequency counts are based on a wide range of material and consider both frequency and range when making their ranked frequency lists.

Special purposes vocabulary

If, however, the learners intend to go on with academic study, their vocabulary learning should go in a different direction. There have been several studies that have investigated the vocabulary needed for academic study. Two of them (Campion and Elley, 1971; Praninskas, 1972) assumed that learners already know a general service vocabulary and these studies looked at academic texts to see what words not in the general service vocabulary occur frequently across a range of academic disciplines. Two other studies (Lynn, 1973; Ghadessy, 1979) looked at the words that learners of English wrote translations above in their academic texts. There were considerable overlaps between these four lists and they were combined into one list, the University Word List (UWL), by Xue and Nation (1984). This combined list of academic vocabulary was designed so that it consists of words not in the GSL but which occur frequently over a range of academic texts. The academic vocabulary, which contains over 800 word families, gives an 8.5% coverage of academic texts, 3.9% coverage of newspapers, and 1.7% coverage of fiction (Hwang, 1989). Some writers have called this academic vocabulary "subtechnical vocabulary". The value of such a vocabulary is shown by its impressive 8.5% coverage of academic texts. Its low coverage of non-academic texts shows its specialized nature. Words from the UWL (Xue and Nation, 1984) include *acquire, complex, devise, fallacy, goal, imply, intelligent, phase, status*. In the very first paragraph of this paper, the words *texts, academic, research, evidence, technical*, and *range* are all in the UWL. All the other words except for *non-fiction, typify*, and *coverage* are in West's GSL. The UWL can also be found in Nation (1990).

Technical vocabulary and low frequency vocabulary

Technical vocabulary occurs with very high or moderate frequency within a very limited range of texts or just within one text (Bečka, 1972; Carroll and Roeloffs, 1969). In economics texts, for example, we find words like *isocost, utility*, and *duopoly* occurring frequently. They are unlikely to occur at all or with high frequency in other kinds of texts with the same meaning. Research on novels (Hirsh, 1993) suggests that technical vocabulary may also have a limited range *within*

a text. That is, each item occurs quite frequently within one chapter or subsection of the text. This seems to be the case with many of the proper nouns in novels (which seem to act like technical vocabulary) and may also occur with technical words in non-fiction texts. It is not clear whether there is a distinction to be made between technical vocabulary, and words with no technical meaning but closely related to the topic which occur with such high frequency in a particular text that several are among the most frequent 100 words in the text. Bramki and Williams (1984) indicate that a large amount of technical vocabulary is explained when it first appears in a text. This may be less likely to happen with topic vocabulary.

Low frequency vocabulary consists of words that occur with low frequency over a range of texts, that are so rare that low frequency is inevitably related to narrow range, or that are the technical vocabulary of other subjects (one person's technical vocabulary is another person's low frequency vocabulary!).

These four kinds of vocabulary, high frequency, academic, technical, and low frequency, have been described as if they are clearly separable but that is not true. Any division is based on an arbitrary decision on what numbers represent high, moderate, or low frequency, or wide or narrow range, because vocabulary frequency, coverage and range figures for any text or group of texts occur along a continuum. Nevertheless, it is important for teaching and learning to make the divisions. Different kinds of vocabulary require different teaching and learning responses because different kinds of vocabulary provide different benefits for the cost of teaching and learning. The division between high and low frequency vocabulary is the most important of these. For the same reason it is also important to course designers, teachers and learners to know when specialization in vocabulary learning should begin. In this way learners can continue to gain maximum benefits in terms of text coverage from their vocabulary learning. An informed decision on the boundary between general service and academic (special purposes) vocabulary would also help designers of general service courses to set clear vocabulary learning goals for their courses.

The rest of this article looks at how and where the arbitrary dividing line between general service vocabulary and academic vocabulary can be most sensibly drawn.

HOW LARGE IS A GENERAL SERVICE VOCABULARY?

The criteria used to help answer this question are frequency, coverage, and range. Coverage and frequency are directly related as coverage simply consists of the frequency figures for a group of words added up and converted to a percentage of the total number of running words in a particular text or collection of texts. When West designed the GSL he used other criteria as well, because he wanted the list to cover things like the classroom use of language, and the definition of vocabulary outside the list. He also wanted to exclude words whose function could be covered by other words in the list (West, 1953: pp. 9–10). This resulted in the inclusion of some low frequency words and the exclusion of some high frequency words with a resulting reduction in text coverage. The aim of this article is to determine the size of a general service vocabulary that gives the highest possible coverage of written text as a prerequisite to moving on to the learning of more specialized vocabulary. This aim does not exclude the development of additional vocabularies of words that are common in spoken English, words that are not frequent but essential in daily life (such as *toilet*!), and words that are needed in the classroom.

The data

A general service vocabulary should consist of high frequency vocabulary which occurs across a wide range of different kinds of texts. The LOB (Johansson, 1978) and Brown (Francis and Kučera, 1982) corpora along with the GSL were used to select such words. The LOB and Brown corpora each consist of approximately 1,000,000 running words and are divided into 15 different sections including press reportage, religion, learned and scientific, and general fiction. The LOB contains samples of British English and the Brown contains samples of American English. All the high frequency items which occurred in 10 or more of the 15 sections were selected from each corpus: 1810 from the LOB and 2410 from Brown. The choice of 10 or more sections as a measure of wide range was arbitrary, but ensured that each word occurred both in fiction and non-fiction texts. From here on, the terms LOB list and Brown list are used to refer to the high frequency words from those lists with a range of 10 sections or more.

The first step was to look at the overlap between the LOB, Brown and GSL lists. Next, the text coverage of the various overlapping and non-overlapping parts of the lists was examined. This was done by using the LOB corpus as the text and seeing what percentage of LOB was covered by the various sublists. Using LOB as the text favoured the coverage of those sublists where LOB is involved, for example the sublist of the LOB and GSL overlap. This, however, was unavoidable because, except for LOB and Brown, there was no other large corpus available at the time. Although this slightly overestimates coverage it does not effect the figures needed to draw the dividing line between general service and academic vocabulary. Finally, the text coverage of each 100 words in the various sublists was calculated (% text coverage divided by the number of words times 100 over 1). It was important to get this average figure for each sublist because this made it possible to compare the coverage of each sublist even though each sublist contained a different number of words. Table 1 contains the results. For comparison the text coverage of a specialized word list, the UWL (Xue and Nation, 1984), was also included. Note that the coverage for the UWL is 4.5% for a corpus of various texts including fiction and non-fiction, and 8.5% for a specialised corpus of non-fiction academic texts.

Deciding on the general service vocabulary cut-off point

There are four questions that can help decide where the cut-off point for a general service vocabulary should be. The first question is, is the vocabulary truly "general service"? That is, does it occur in a range of texts and a range of corpora? Table 1 shows that 1331 words occurred in all three lists. These words give the highest coverage of the texts in the LOB corpus with each 100 words covering 5.9% of the text. They are clearly part of a general service vocabulary.

A general service vocabulary could also include the 614 words that occur in two lists. Two hundred and fifty occur in both Brown and LOB but not in the GSL. Two hundred and twenty-six occur in both Brown and the GSL but not in words with a range of 10 or more in LOB, and 138 occur in both LOB and the GSL but not in the words with a range of 10 or more in Brown, 1945 words (1331 + 614) occur in two or three lists.

The average coverage of the lists

The second question is, do the items included in the general service vocabulary give a better average coverage of text than the items that are excluded? The text coverage by each 100 words of the words that occur in two lists (0.96, 0.58, 1.01) is higher than that of the words that occur in only one of the lists (0.54, 0.55, 0.29). There are thus clear enough dividing lines based on

coverage by each 100 words between the words occurring in all three lists, the words occurring in two lists, and the words occurring in only one list.

Table 1. The amount of overlap, text coverage, and text coverage by each 100 words of the LOB, Brown and GSL lists

Subgroup	Number of words	% text coverage	Text coverage by each 100 words
In all three lists			
Brown/LOB/GSL	1331	78.3	5.90
In only two lists			
Brown/LOB	250	2.4	0.96
Brown/GSL	226	1.3	0.58
LOB/GSL	138	1.4	1.01
Total	614	5.1	0.83
In only one list			
Brown only	333	1.8	0.54
LOB only	91	0.5	0.55
GSL only	452	1.3	0.29
Total	876	3.6	0.41
UWL (all of LOB)	833	4.5	0.54
(academic texts)	833	8.5	1.02

Seventy per cent (174) of the words in the Brown/LOB subgroup are also in the University Word List. It is thus possible to exclude these 174 words from the general service vocabulary and make a general service vocabulary of 1771 (1945 minus 174) words, or to include them in the general service vocabulary and make an academic vocabulary of 659 (833 minus 174) words. The general service vocabulary can be seen as a group of words selected from a continuum. We have identified three points at which a cut-off point could be made, namely the overlap of the three lists (1331 words), the overlaps of two of the lists minus the UWL items in the Brown/LOB overlap (1771 words), and the overlaps of two of the Brown/LOB overlap (1771 words), and the overlaps of two of the lists (1945 words).

The coverage by specialized vocabulary

The third question to help decide the cut-off point is, is the average coverage of specialized texts by the UWL higher than the average coverage of the words just outside the general service vocabulary? That is, for learners intending to go on to academic study, is the UWL the next most useful vocabulary to learn in terms of the value for learning that it provides through text coverage? The text coverage by each 100 words of the UWL on the whole unspecialized LOB corpus (0.54) is similar to the coverage by each 100 words of the words that occur in only one of the three lists (0.54, 0.55, 0.29). This indicates that it is a suitable point for learners with special purposes to move to learning specialized vocabulary. The coverage figure for the UWL is 8.5% for academic (learned and scientific) texts. This gives a coverage by each 100 words of 1.02. This is similar to the average coverage per word of the Brown/LOB and LOB/GSL subgroups (0.96, 1.01). This means that the value in learning the Brown/LOB and LOB/GSL sublists is the same as learning the specialized UWL lists. So all of these should be learned before going on to the items with less coverage. This supports the cut-off point at the 1771 or 1945 word levels,

because it shows that learning the UWL should come immediately after the overlapping sublists, rather than later.

Cumulative coverage

The fourth question is, does the sequence of learning provide the best cumulative coverage of text? The overlap of the Brown/LOB/GSL lists gives a learner 78.3% coverage of the LOB corpus. The Brown/LOB, Brown/GSL, and LOB/GSL overlaps add another 5.1% coverage, bringing the total coverage to 83.4%. For learners with academic goals, the UWL contains the next vocabulary to learn. The 8.5% coverage of the UWL added to the 83.4% coverage of the general service vocabulary brings the total coverage to 91.9%. Note that although the coverage per 100 words of the UWL is a fraction higher than the coverage per 100 words of the Brown/LOB and LOB/GSL overlaps, it is best to learn the Brown/LOB and LOB/GSL overlaps first to add the coverage that they provide to the learner's cumulative coverage. To put it another way, it is best to get the best possible coverage of general service vocabulary before moving to more specialized vocabulary.

This study, then, shows that a general service vocabulary of close to 2000 words is appropriate for learners who are going to move on to special purposes study. After the 2000-word level, greater text coverage is gained by moving to the study of specialized vocabulary than is gained by continuing to learn the next words on a frequency list. Table 1 shows that the text coverage gained from this 1945-word vocabulary will be around 83.4%. When the 8.5% text coverage that the UWL provides for academic texts is added to this, learners are approaching the 95% text coverage which is the minimum required for adequate comprehension (Laufer, 1989) and successful guessing from context (Liu and Nation, 1985).

Of all the sublists, the one containing the 452 words that are only in the GSL gives by far the poorest coverage by each 100 words. These are words in the GSL that need replacing if frequency and range are to be the criteria for selection. In defence of the GSL it should be noted that as well as being designed for a different purpose from the general service vocabulary investigated here, the GSL contains 2147 word families and gives 82.3% text coverage. The general service vocabulary made from the overlaps of the LOB, Brown, and GSL contains 1945 word families and gives 83.4% text coverage. The difference is significant but not great.

The aim of this study has not been to suggest that learners should be learning from word lists, although there is ample research evidence to show that properly managed this is a useful contribution to vocabulary learning (Nation, 1982). Rather the aim is to contribute to the knowledge required to guide the planning of the vocabulary component when setting goals as a part of special purposes course design.

REFERENCES

BEČKA, J. V. (1972) The lexical composition of specialized texts and its quantative aspect. *Prague Studies in Mathematical Linguistics* **4**, 47–64.

BRAMKI, D. and WILLIAMS, R. C. (1984) Lexical familiarisation in economics text and its pedagogic implications in reading comprehension. *Reading in a Foreign Language* **2**, 169–181.

CAMPION, M. E. and ELLEY, W. B. (1971) *An Academic Vocabulary List*. Wellington: NZCER.

CARROLL, J. B. DAVIES, P. and RICHMAN, B. (1971) *The American Heritage Word Frequency Book*. New York: Houghton Mifflin.

CARROLL, J. M. and ROELOFFS, R. (1969) Computer selection of keywords using word-frequency analysis. *American Documentation* **20**, 227–233.

ENGELS, L. K. (1968) The fallacy of word counts. *IRAL* **6**, 213–231.

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

GHADESSY, M. (1979) Frequency counts, word lists, and materials preparation: a new approach. *English Teaching Forum* **17**, 24–27.

HIRSH, D. (1993) The vocabulary demands and vocabulary learning opportunities in short novels. Unpublished M.A. thesis, Victoria University, Wellington.

HIRSH, D. and NATION, I. S. P. (1992) What vocabulary is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language* **8**, 689–696.

HWANG KYONGHO (1989) Reading newspapers for the improvement of vocabulary and reading skills. Unpublished M. A. thesis, Victoria University, Wellington.

HWANG KYONGHO and NATION I. S. P. (1989) Reducing the vocabulary load and encouraging vocabulary learning through reading newspapers. *Reading in a Foreign Language* **6**, 323–335.

JOHANSSON, S. (1978) Manual of Information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers. Department of English, University of Oslo.

LAUFER, B. (1989) What percentage of text-lexis is essential for comprehension? In Lauren, C. and Nordman, M. (eds), *Special Language: From Humans Thinking to Thinking Machines*. Clevedon: Multilingual Matters.

LIU, N. A. and NATION, I. S. P. (1985) Factors affecting guessing vocabulary in context. *RELC Journal* **16**, 33–42.

LYNN, R. W. (1973) Preparing word lists: a suggested method. *RELC Journal* **4**, 25–32.

NATION, I. S. P. (1982) Beginning to learn foreign vocabulary: a review of the research. *RELC Journal* **13**, 14–36.

NATION, I. S. P. (1990) *Teaching and Learning Vocabulary*. New York: Heinle and Heinle.

PRANINSKAS, J. (1972) *American University Word List*. London: Longman.

RICHARDS, J. C. (1974) Word Lists: Problems and prospects. *RELC Journal* **5**, 69–84.

THORNDIKE, E. L. and Lorge, I. (1944) *The Teacher's Word Book of 30,000 Words*. Teachers College: Columbia University.

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

XUE GUOYI and NATION I. S. P. (1984) A University word list. *Language Learning and Communication* **3**, 215–229.