This is an offprint from:

Mohsen Ghadessy, Alex Henry and Robert L. Roseberry (eds)
Small Corpus Studies and ELT
Theory and practice
John Benjamins Publishing Company
Amsterdam/Philadelphia
2001
(Published as Vol. 5 of the series
STUDIES IN CORPUS LINGUISTICS, ISSN 1388-0373)

ISBN: 90 272 2275 4 (Eur.) / 1 58811 035 4 (US)

CHAPTER 2

Using small corpora to investigate learner needs
Two vocabulary research tools

Paul Nation
LALS, Victoria University of Wellington, New Zealand

Abstract

Corpus research has three essential requirements — a set of good research questions that can be answered by study of a corpus, a corpus to provide a source of data, and the computer programmes that can facilitate the task of organising the data from the corpus. This paper looks at such programmes and in particular describes two computer programmes that have been specially developed to address the needs of learners of English as a second or foreign language. It also looks at how teachers can use corpora based on the texts that learners will have to read, or based on writing produced by learners to investigate learners' vocabulary needs. This paper will also suggest several research questions that teachers could use to guide their investigation. It describes published and unpublished research that has addressed similar questions.

Introduction: Research questions and background

The two computer programmes that will be described in this paper, *VocabProfile* and *RANGE*, have been used to investigate the amount of low frequency and other types of vocabulary in different kinds of written input that second or foreign language learners may have to cope with, and to investigate the amount of different kinds of vocabulary in the written output produced by language learners.

When analysing reading texts, it is possible to use *VocabProfile* and *RANGE* to answer questions like the following.

- How large a vocabulary do you need to read newspapers?
- How large a vocabulary do you need to read novels?
- Do graded readers provide good conditions for vocabulary learning?
- Do you have to know the vocabulary introduced at a level in a graded reader scheme before you begin reading books at that level?
- What is the vocabulary load of an economics textbook?
- Is there a special purposes academic vocabulary which is important for reading academic texts?

When looking at learners' writing, it is possible to use *VocabProfile* and *RANGE* to answer questions like the following.

- Are learners using an appropriate variety of vocabulary in their written work?
- Is there a significant gap between learners' receptive and productive vocabulary?
- Is the vocabulary learners meet in their reading carried over into their written production?

The main idea lying behind some of the uses of these programmes is that the words in the vocabulary of a language are not created equal. Some words are much more important than others. The typical way of determining the importance of a word is by looking at its frequency and range of occurrence. That is, words that occur often in a wide range of language uses are much more generally useful for a language user to know than words which occur rarely and in a limited range of areas, particularly if these areas such as biology, computing, geography etc are not of immediate interest to the language user.

The words that occur often in a range of uses of the language are called high frequency words or general service words (West 1953). Typically the dividing line between high frequency and low frequency words in English is drawn at the 2000 word level. That is, it is generally considered there are around 2000 high frequency words. This dividing line is an arbitrary one but it is supported by a substantial amount of research. Table 1 shows how the coverage of vocabulary drops substantially beyond the most common 2000 word families in a corpus of 5,000,000 running words.

Table 1: Vocabulary size and coverage (Carroll, Davies and Richman 1971)

Number of words%	Text coverage
86,741	100
43,831	99
12,448	95
5,000	89.4
4,000	87.6
3,000	85.2
→ 2,000	81.3
1,000	74.1
100	49
10	23.7

When we say that the high frequency vocabulary is important, its importance comes from its high probability of being met in a wide range of language uses. If a learner does not know it, then the learner will face the same vocabulary difficulties in all uses of the language. There are other ways of defining *important*, such as important for understanding the meaning of a particular text, but the way used here takes account of learners' long term use of the language.

There is plenty of evidence that generally learners learn high frequency vocabulary before they learn low frequency vocabulary (Read 1988). This is not surprising because language courses sensibly focus on high frequency vocabulary and there are many more opportunities in normal language use to meet high frequency vocabulary. For learners at an intermediate level and beyond, it will generally be the low frequency vocabulary that causes them difficulty. It is thus useful to be able to quickly analyze texts to see how much low frequency vocabulary they contain and what their low frequency vocabulary is. This can help a teacher decide if the text needs to be simplified, if it needs to be discarded and an easier one found, or if it is feasible to pre-teach some of the vocabulary. The computer programmes described in this article allow this analysis to be done.

One of the ways of making low frequency vocabulary more manageable is to increase the number of high frequency words. This can be done by taking a special purposes approach. That is, the language use goals of the learners are examined and then some research is done to see if there is a specialised vocabulary that is not in the most frequent 2000 words of the language but which is frequent and of wide range within the limited area of specialisation that the learners are interested in. The programmes described in this article are a useful means of doing such research.

One such specialised vocabulary that has been developed (using the *RANGE* programme) is an academic word list (Coxhead 2000). This is described in a little more detail later in this paper. Such vocabulary typically covers around 8.5% of the running words in academic texts (see Table 3). It is thus a very important learning goal for learners with academic purposes.

The most frequent 2000 words of English and the academic word list are available as word lists with the programmes described in this article.

The word lists which come with the programmes consist of base words and their closely related inflected and derived forms. They do not include collocations or lexical phrases, and the programmes are not capable of distinguishing different meanings of the same forms such as *row* (a lot of noise) and *row* (a line of objects or people), or *date* (the fruit) and *date* (a time). These are problems which could probably only be satisfactorily solved by marking up texts to tag such items. Only a small number of idiomatic collocations are frequent enough to get into the most frequent 2000 words of English (Nation 1999: 293). This is because the frequency of any collocation will be much less than the frequency of the individual items that make it up. The lists used in the programmes however do not include collocations.

Let us now look in detail at the two programmes.

VocabProfile and *RANGE*

VocabProfile and *RANGE* are freeware available at http://www.vuw.ac.nz/lals/. Detailed instructions about how to use the programmes are in a file called *instruct.wp* (or *instruct.dos*). They only run on PCs.

VocabProfile can be used to compare a text against vocabulary lists to see what words in the text are and are not in the lists, and to see what percentage of the items in the text are covered by the lists. It can also be used to compare the vocabulary of two texts to see how much of the same vocabulary they use and what vocabulary differences exist between them.

Specifically, *VocabProfile* shows which words in a text are covered by each of three user-created or ready-made word lists and which words are not

covered by any of these lists. The available lists are described in detail below. It does this by comparing the words in a text with the words in the three lists, marks words in the text according to which word list or lists contain them, and produces various lists of the words from the texts.

In addition, *VocabProfile* provides a table which shows how much coverage of a text each of the three lists provides. For each word list, the programme shows how many word tokens (total running words), how many word types (different word forms), and how many word families, or *lemmas*, (groups containing different forms of a word) the text contains. (For information on word families see Bauer and Nation (1993).) It also provides similar statistics for words that are not contained in any of the lists. Both totals and percentages are given, as shown in Table 2, below. (In the 'word list' column, 'one', 'two', and 'three' refer to the three word lists.)

Table 2: ***A word list table produced by VocabProfile***

Word list	Tokens/%	Types/%	Families
One	54/72.0	34/69.4	33
two	2/ 2.7	2/ 4.1	2
three	14/18.7	9/18.4	9
not in the lists	5/ 6.7	4/ 8.2	?????
Total	75	49	44

In the programme, word list one is called BASEWRD1.DAT, word list two is called BASEWRD2.DAT, and word list three BASEWRD3.DAT. The word count data in Table 2 refer to a short text which contained 75 running words (tokens). Fifty-four words in the text are in word list one and these 54 words make up 72% of the running words in the text. These 54 tokens are made up of 34 different words (types) from 33 word families.

In the three lists, *VocabProfile* also keeps a cumulative record of the frequency of items it meets in all the texts it is run on. Figure 1 shows the first items in word list one (BASEWRD1.DAT), showing word frequencies for a short text. Note that family members are indented under the head word for that word family.

```
A            14
  AN         11
ABLE          0
   ABLER      0
   ABLEST     0
   ABLY       0
   UNABLE     0
ABOUT         6
ABOVE         3
```

Figure 1: Frequencies of the first few words in BASEWRD1.DAT

The programmes thus count the vocabulary used in a text and compare it with accompanying word lists which allow the vocabulary in the lists to be counted as word families.

What is needed to run VocabProfile*?*

To run *VocabProfile* on a text, three word lists are needed: They must be called BASEWRD1.DAT, BASEWRD2.DAT, BASEWRD3.DAT. They must be ASCII files. Any words can be put into these three lists as described below. If you want to use only one or two lists, you can just make one or two other files with the correct name (e.g. BASEWRD2.DAT) but with no words or a nonsense word in them in order to have the three lists which must be there for *VocabProfile* to run.

VocabProfile will usually only run accurately on an ASCII (DOS) text file so it may be necessary to convert word processor files before using them. Word processing programmes are capable of doing this using the "save as" option.

The lists available for VocabProfile *and* RANGE

Three ready made lists are available. The first (BASEWRD1.DAT) includes the most frequent 1000 words of English. The second (BASEWRD2.DAT) includes the 2nd 1000 most frequent words, and the third (BASEWRD3.DAT) includes words not in the first 2,000 words of English but which are frequent in university texts from a wide range of subjects. All of these lists include the base forms of words and derived forms. The first 1000 words thus consist of around 4,000 forms. The sources of these lists are *A General Service List of English Words* (West 1953) for the first 2000 words, and *An Academic Word List* (Coxhead 1998). The first thousand words of *A General Service List of English Words* are usually those in the list with a frequency higher than 332 occurrences per 5 million words, plus months, days of the week, numbers, titles (Mr, Mrs, Miss, Ms, Mister), and frequent greetings (Hello, Hi etc.).

The lists include both American and British spellings. Apostrophes are treated as spaces, so "I've" is counted as two items, as is "Jane's".

The word forms in the lists are grouped into word families under a headword. For example, the headword *aid* has the following family members *aided*, *aiding*, *aids*, and *unaided*. In the lists the family members have a Tab (indent) in front of them, as shown in Figure 1. The headword occurs just before the family members and has no Tab.

Preparing your own lists

You do not need to use the lists that are provided with the programmes. If, for example, you wish to look at the overlap between two texts, because you want to see how well one text prepares learners for a subsequent text, you can turn one of the texts into a word list by running the programme *WORD*, *VocabProfile*, or *RANGE*, edit it to make word families, and give it the name BASEWRD1.DAT, so that it becomes a base list that *VocabProfile* will use. You then make two other lists named BASEWRD2.DAT and BASEWRD3.DAT each with only a nonsense word in them, and run the programme with the other text as the input text.

To run *VocabProfile*, you type "VP". The computer will ask you to give the name of the data file you want to run *VocabProfile* over. Give the name of the data file (e.g. TEXT1.DOS) and hit the enter key. The data file is the file name of the text that you want to analyse. The computer will next ask you to give the name of the output file you want to send the results to. When you analyse your text with *VocabProfile*, you need a separate file for the results so that your original text remains unchanged. You can use any name for the output file, and you then can choose the options to get the kind of output you want.

You can choose to create a marked text if you want to see where the words in the lists occur in the text. The output, shown in Figure 2, will include a full copy of the text with the words marked according to the lists they occur in.

There are two numbers, separated by a vertical slash, following each headword in the family lists in the output. For example,

```
LEARN    4  |  1
```

The first figure is the total family members including the headword that occurred. In the example with *learn*, *learner* occurred twice in the text, *learning* once and *learn* once. The second figure is the number of times the actual form of the headword *learn* occurred.

The information produced by *VocabProfile* allows teachers to see how many and what words are not in the word lists and thus which might be unknown to their learners. Sometimes the table of tokens, types, and families is enough information to decide if a text will be too difficult, but it is wiser to look at the words not in the list as well, because many of these may be proper nouns or familiar words that could be considered as not adding to the vocabulary difficulty of the text.

RANGE

Another programme, *RANGE,* is used to compare the vocabulary of up to 32 different texts at the same time. This programme is useful when looking at a series of texts in a course book or a set of graded readers to see how much the vocabulary is repeated in different texts. For each word in the texts, the programme provides a range or distribution figure (how many texts the word occurs in), a headword frequency figure (the total number of times the actual headword type appears in all the texts), a family frequency figure (the total number of times the word and its family members occur in all the texts), and a frequency figure for each of the texts the word occurs in. It can be used to create word lists based on frequency and range, and can be used to discover shared and unique vocabulary in several pieces of writing.

What is needed to run RANGE?

To run the programme you need

1. the programme *Range.exe*,
2. the three base word lists (BASEWRD1.DAT, BASEWRD2.DAT, BASEWRD3.DAT),

RANGE can be used with the same word lists used by *VocabProfile*. This allows it to classify some of the words in the input files into word families. The programme will give different figures depending on whether the word lists are used or not. If the word lists are used, the figures will represent a mixture of families and types. All the words in the word lists are counted as families and the remainder are counted as types. If the word lists are not used, then all the words are counted as types, because it is the word lists that are used to make families.

Figure 3 shows some sample output from *RANGE*. In this example, the programme was run on four files (indo1.dos, indo2.dos, indo3.dos, indo4.dos). The output file was called 'results'. Notice the choices available in the options. Comments are included in brackets.

Comments included in brackets are not produced by the program but are added here to help understand the example.

```
WORD LIST               TOKENS/%        TYPES/%         FAMILIES

one                     6464/85.1       650/74.1        432 (the first 1000 words)
two                      420/ 5.5       158/18.0        120 (the second 1000 words)
three                      8/ 0.1         3/ 0.3          3 (the Academic Word List)
not in the lists         706/ 9.3        66/ 7.5        ?????

Total                   7598            877             555

(The following figures show how many words each list contains e.g. the AWL has 570 families and
these are made up of 3110 types)

 Number of BASEWRD1.DAT types: 4125  Number of BASEWRD1.DAT families: 999
 Number of BASEWRD2.DAT types: 3707  Number of BASEWRD2.DAT families: 986
 Number of BASEWRD3.DAT types: 3110  Number of BASEWRD3.DAT families: 570

(Only a very small part of the output is provided here. The output has been sorted by frequency.
All the words shown here appear in all four input files and so have a range of 4. The is a single
member word family, so its family frequency and type frequency are the same - 333. The type
frequency of the word type A is 128. Its family frequency is 135 which means that the other
member of the family, an, occurred 7 times. F1, F2 etc refer to each of the four texts and the
family frequencies are given for each text.)

LIST OF FAMILY GROUPS

BASE ONE FAMILIES                    RANGE  TYFREQ  FAFREQ    F1     F2     F3     F4
THE                                      4     333     333    96     79     94     64
AND                                      4     261     261    74     57     87     43
TO                                       4     201     201    39     51     44     67
YOU                                      4     154     177    33     39     44     61
HE                                       4     151     247    98     51     54     44
SHE                                      4     147     253    62     84     46     61
I                                        4     142     215    51     45     36     83
A                                        4     128     135    36     28     34     37
BUT                                      4      83      83    16     28     21     18
IT                                       4      83      83    28     13     17     25
...

(The following data is for some of the low frequency words)

TYPE                                 RANGE    FREQ    F1     F2     F3     F4
SITA                                     1       1     1      0      0      0
TANK                                     1       1     0      0      1      0
TRUCKS                                   1       1     0      1      0      0
TUTI                                     1       1     1      0      0      0
VEGETABLE                                1       1     1      0      0      0
WELL-KNOWN                               1       1     0      0      0      1

(This is a list of types not families. Each type occurs only once in one of the four texts.)
```

Figure 3: Sample output from the *RANGE* program.

Comments included in brackets are not produced by the program but are added here to help understand the example.

This data on occurrences across texts, especially related texts, is valuable when making lists of useful words to focus on and when looking at what spaced repetitions are provided to help the incidental learning of words, as we shall see in some of the studies described below.

```
range.exe        (the programme)
basewrd1.dat     (the word lists)
basewrd2.dat
basewrd3.dat
indo1.dos
indo2.dos        (the files to be processed)
indo3.dos
indo4.dos
indo5.dos
indo6.dos
```

Figure 4: A sample directory for the RANGE programme

Research using *VocabProfile* and *RANGE*

There are now several studies which have used *VocabProfile* or *RANGE* to look at the vocabulary load of texts that learners may need to read.

Academic text and academic vocabulary

Studies of the vocabulary load of academic texts typically show that the most frequent 2,000 words of English cover around 80% of the running words, and the University Word List covers around 8.5% of the running words, making a total of 88.5% coverage by these two groups of words (Sutarsyah, Nation and Kennedy 1994).

Sutarsyah, Nation and Kennedy (1994) used *VocabProfile* to compare the vocabulary needed to read one complete economics text book with a collection of short academic texts on a variety of topics totaling the same length as the book. The texts were from a wide range of academic disciplines. Table 3 shows the results. In this table the word coverage is based in part on the *General Service List* (GSL) and the *University Word List* (UWL), both of which were mentioned earlier. The "1st 1,000" and "2nd 1,000" words are from the *General Service List.*

In Table 3, note that the economics text has a total vocabulary of 5,438 word families, while the collection of short academic texts contains 12,744 word families. Clearly, reading a continuous text on the same topic by the same author requires a much less diverse vocabulary. It is thus useful to focus on learners' specific disciplines as soon as they have control of the general academic vocabulary in order to keep a focus on the most useful vocabulary.

Table 3: ***Number of word families and percentage of coverage of the economics text and the general academic corpus by GSL and UWL***

Word level	Families in the econom-ics text	Coverage of the econom-ics text	Families in the general academic corpus	Coverage of the general academic corpus
1st 1,000 (GSL)	1,029	77.72%	1,095	74.11%
2nd 1,000 (GSL)	548	4.78%	796	4.32%
UWL	636	8.74%	811	8.40%
Others	3,225	8.77%	10,042	13.16%
Total	5,438	100%	12,744	100%

In a subsequent study, Coxhead (1998) used the *RANGE* programme to create an academic word list consisting of vocabulary that was frequent and wide ranging in academic texts, but which was not contained in the most frequent 2,000 words of English. Coxhead's corpus of academic texts consisted of four divisions of Arts, Commerce, Science, and Law. Each of these four divisions was divided into seven subject areas such as education, history, linguistics, and so on. *RANGE* was used to find the range and frequency of occurrence of words that were not among the 2,000 most frequent words. The resulting list is now available as BASEWRD3.DAT. Table 4 below shows the coverage of the sublists of this list.

This 570 word vocabulary provides very substantial coverage of academic text and is an important learning goal for learners with academic purposes.

Books written for teenagers

Hirsh and Nation (1992) used the *VocabProfile* programme[1] to look at the accessibility for second language learners of novels written for young native speakers. They found that if proper nouns are considered to be known, a vocabulary of the 2,000 most frequent words brought learners very close to the 95% coverage which is the minimum needed for adequate comprehension (Laufer 1989). Hirsh and Nation however considered that for reading for plea-

Table 4: Coverage of Coxhead's academic corpus by sublists of her Academic Word List (1998)

AWL sublist	Coverage of the Academic Corpus (%)
1 (60 families)	3.6%
2 (60 families)	1.8%
3 (60 families)	1.2%
4 (60 families)	0.9%
5 (60 families)	0.8%
6 (60 families)	0.6%
7 (60 families)	0.5%
8 (60 families)	0.3%
9 (60 families)	0.2%
10 (30 families)	0.1%
570 families	10.0%

sure, 98% coverage of the running words (1 unknown word in every 50 running words) was necessary. To achieve this coverage, learners would need a vocabulary size of around 5,000 word families.

Graded readers

Nation and Wang (1999) looked at forty-two books in a scheme of graded readers to see what conditions they provide for reading and vocabulary learning. The word lists were based on the vocabulary levels of the grading scheme. *VocabProfile* was used to measure how well the words from the preceding levels and current level covered the text. *RANGE* was used to investigate how often the words were repeated in individual books and groups of books. For most levels of the scheme studied, learners needed to know the vocabulary introduced at each level before they could read the books at that level comfortably. If they did not know these words, then they would know less than 95% of the running words in the books at that level and would have difficulty reading for pleasure. In terms of repetition, graded readers provide very good conditions for learning the high frequency vocabulary of English. The study also determined that to gain the best effects of repetition, learners needed to be

reading about one graded reader per week, and should move through all the levels in the scheme, reading from three to five books at each level, preferably more at the later levels. This was calculated by relating the average gap between repetitions with the length of time that memory for a meeting with a word might remain in the learner's mind.

Vocabulary richness in learners' written production

The studies of the vocabulary richness of learners' writing (Laufer and Nation 1995; Laufer, 1994) have looked at the percentage of word types at various frequency levels. The most common measure used in these studies is the Lexical Frequency Profile, developed by Laufer and Nation (1995), which looks at the percentage of words beyond the most frequent 2,000 words (proper nouns and lexical errors are excluded from the compositions before they are processed by the *VocabProfile* programme using the existing 2,000 and academic word lists). The Lexical Frequency Profile has been shown to be a reliable and valid measure, and to change over time with instruction. Nation (2001) also suggests that teachers can use the learners' compositions with the words marked up according to their frequency level as a way of commenting on learners' vocabulary use in their writing.

Summary

The *VocabProfile* and *RANGE* programmes are useful ways of gathering information about learners' vocabulary needs. They allow texts to be compared with word lists based on word families and thus provide a way of measuring the vocabulary load or richness of texts. The results of these measures can then be compared with learners' vocabulary sizes to see what they need to focus on in their learning. The programmes put this kind of analysis well within the reach of teachers.

The most immediate use a teacher might make of the programmes is to look at texts that the learners are working with to see what low frequency words they contain and if there are too many of them. Learners' vocabulary size can be checked using the Vocabulary Levels Test (Nation 2001). This tests knowledge of the second 1000 high frequency words, the academic word list, and three low frequency word levels.

A teacher may also wish to check that the learners' course material is providing sufficient spaced repetition of target vocabulary. If it is not, then additional activities may need to be devised to reinforce such vocabulary.

If learners do some of their writing on the computer, it becomes very easy to see, using *VocabProfile*, if learners' receptive vocabulary knowledge (as measured by the Vocabulary Levels Test) is moving into productive use in their writing. If it is not, then it may be useful to encourage this move to productive use through the use of discussion activities before writing, or the use of outline sheets while writing.

Finally, the programmes can help in the preparation of material using a controlled vocabulary. In spite of the large numbers of graded readers available, there are still only a few with settings outside English speaking countries. Teachers are probably the best people to write such readers and the *VocabProfile* programme can be an easy way of checking on the control of vocabulary.

Note

[1] When running the programme, the proper nouns in the texts were made into a word list used by the programme.

References

Bauer, L. and Nation, I. S. P. 1993. "Word families". *International Journal of Lexicography* 6 (3):1–27.

Carroll, J. B., Davies, P. and Richman, B. 1971. *The American Heritage Word Frequency Book.* New York: Houghton Mifflin, Boston American Heritage.

Coxhead, A. 2000. "A new academic word list". *Tesol Quarterly* 34 (2):213–238.

Coxhead, A. 1998. "An academic word list". *Occasional Publication Number 18*, LALS, Wellington, New Zealand: Victoria University of Wellington.

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

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

Laufer, B. 1994. "The lexical profile of second language writing: Does it change over time?" *RELC Journal* 25(2):21–33.

Laufer, B. and Nation, P. 1995. "Vocabulary size and use: Lexical richness in L2 written production". *Applied Linguistics* 16(3):307–322.

Nation, I. S. P. 1990. *Teaching and Learning Vocabulary.* Rowley MA: Newbury House.

Nation, I. S. P. 2001. *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Nation, I. S. P. and Wang, K. 1999. "Graded readers and vocabulary". *Reading in a Foreign Language* 12(2):355–380.

Read, J. 1988. "Measuring the vocabulary knowledge of second language learners". *RELC Journal* 19(2):12–25.

Sutarsyah, C., Nation, P. and Kennedy, G. 1994. "How useful is EAP vocabulary for ESP? A corpus based study". *RELC Journal* 25(2):34–50.

West, M. 1953. *A General Service List of English Words.* London: Longman, Green & Co.