How much vocabulary does a second language learner need?

There are three ways of answering this question. One way is to ask “How many words are there in the target language?” Another way is to ask “How many words do native speakers know?” A third way is to ask “How many words are needed to do the things that a language user needs to do?” We will look at answers to each of these questions.

This discussion looks only at vocabulary and it should not be assumed that if a learner has sufficient vocabulary then all else is easy. Vocabulary knowledge is only one component of language skills such as reading and speaking. It should also not be assumed that vocabulary knowledge is always a prerequisite to the performance of language skills. Vocabulary knowledge enables language use, language use enables the increase of vocabulary knowledge, knowledge of the world enables the increase of vocabulary knowledge and language use and so on (Nation, 1993b). With these cautions in mind let us now look at estimates of vocabulary size and their significance for second language learners.

How many words are there in English?

The most straightforward way to answer this question is to look at the number of words in the largest dictionary. This usually upsets dictionary makers. They see the vocabulary of the language as a continually changing entity with new words and new uses of old words being added and old words falling into disuse. They also see the problems in deciding if walk as a noun is the same word as walk as verb, if compound items like goose grass are counted as separate words, and if names like Vegemite, Agnes, and Nottingham are to be counted as words. These are all real problems, but they are able to be dealt with systematically in a reliable way.

Two separate studies (Dupuy, 1974; Goulden, Nation and Read, 1990) have looked at the vocabulary of Webster’s Third International Dictionary (1963), the largest non-historical dictionary of English when it was published. When compound words, archaic words, abbreviations, proper names, alternative spellings and dialect forms are excluded, and when words are classified into word families consisting of a base word, inflected forms, and transparent derivations, Webster’s 3rd has a vocabulary of around 54,000 word families. This is a learning goal far beyond the reaches of second language learners and, as we shall see, most native speakers.

How many words do native speakers know?

For over 100 years there have been published reports of systematic attempts to measure the vocabulary size of native speakers of English. There have been various motivations for such studies but behind most of them lies the idea that vocabulary size is a reflection of how educated, intelligent, or well read a person is. A large vocabulary size is seen as being something valuable. Unfortunately the measurement of vocabulary size has been bedeviled by serious methodological problems largely centring around the questions of “What should be counted as a word?”, “How can we draw a sample of words from a dictionary to make a vocabulary test?”, and “How do we test to see if a word is known or not?”. Failure to deal adequately with these questions has resulted in several studies of vocabulary size which give very misleading results. For a discussion of these issues see Nation (1993a), Lorge and Chall (1963), and Thorndike (1924).

Teachers of English as a second language may be interested in measures of native speakers’ vocabulary size because these can provide some indication of the size of the learning task facing second language learners, particularly those who need to study and work alongside native speakers in English medium schools and universities or workplaces.

At present the best conservative rule of thumb that we have is that up to a vocabulary size of around 20,000 word families, we should expect that native speakers will add roughly 1000 word families a year to their vocabulary size. That means that a five year old beginning school will have a vocabulary of around 4000 to 5000 word families. A university graduate will have a vocabulary of around 20,000 word families (Goulden, Nation and Read, 1990). These figures are very rough and there is likely to be very large variation between individuals. These figures exclude proper names, compound words, abbreviations, and foreign words. A word family is taken to include a base word, its inflected forms, and a small number of reasonably regular derived forms (Bauer and Nation, 1993). Some researchers suggest vocabulary sizes larger than these (see Nagy, this volume), but in the well conducted studies (for example, D’Anna, Zechmeister nad Hall, 1991) the differences are mainly the result of differences in what items are included in the count and how a word family is defined.

A small study of the vocabulary growth of non-native speakers in an English medium primary school (Jamieson, 1976) suggests that in such a situation non-native speakers’ vocabulary grows at the same rate as native speakers’ but that the initial gap that existed between them is not closed. For adult learners of English as a foreign language, the gap between their vocabulary size and that of native speakers is usually very large, with many adult foreign learners of English having a vocabulary size of much less than 5000 word families in spite of having studied English for several years. Large numbers of second language learners do achieve vocabulary sizes that are like those of educated native speakers, but they are not the norm.
There is some encouraging news however. A study by Milton and Meara (1995) using the Eurocentres Vocabulary Size Test (Meara and Jones, 1988, 1990) shows that significant vocabulary growth can occur if this learning is done in the second language environment. In their study of a study abroad program of 53 European students of advanced proficiency, the average growth in vocabulary per person approached a rate of 2500 words per year over the six months of the programme. This rate of growth is similar to the larger estimates of first language growth in adolescence. Although the goal of native speaker vocabulary size is a possible goal, it is a very ambitious one for most learners of English as a foreign language.

How many words are needed to do the things a language user needs to do?

Although the language makes use of a large number of words, not all of these words are equally useful. One measure of usefulness is word frequency, that is, how often the word occurs in normal use of the language. From the point of view of frequency, the word the is a very useful word in English. It occurs so frequently that about 7% of the words on a page of written English and the same proportion of the words in a conversation are repetitions of the word the. Look back over this paragraph and you will find an occurrence of the in almost every line.

The good news for second language learners and second language teachers is that a small number of the words of English occur very frequently and if a learner knows these words, that learner will know a very large proportion of the running words in a written or spoken text. Most of these words are content words and knowing enough of them allows a good degree of comprehension of a text. Here are some figures showing what proportion of a text is covered by certain numbers of high frequency words.

<table>
<thead>
<tr>
<th>Vocabulary size</th>
<th>Text coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>72.0%</td>
</tr>
<tr>
<td>2000</td>
<td>79.7%</td>
</tr>
<tr>
<td>3000</td>
<td>84.0%</td>
</tr>
<tr>
<td>4000</td>
<td>86.8%</td>
</tr>
<tr>
<td>5000</td>
<td>88.7%</td>
</tr>
<tr>
<td>6000</td>
<td>89.9%</td>
</tr>
<tr>
<td>15,851</td>
<td>97.8%</td>
</tr>
</tbody>
</table>

The figures in Table 1 refer to written texts and are from Francis and Kucera (1982) which is a very diverse corpus of over 1,000,000 running words made up of 500 texts of around 2000 running words long. As we shall see the more diverse the texts in a corpus, the greater the number of different words and the high frequency words cover slightly less of the text, so these figures are a conservative estimate. The figures in the last line of the table are from Kucera (1982). The COBUILD Dictionary claims that 15,000 words cover 95% of the running words of their corpus. The figures in Table 1 are for lemmas and not word families. Word families would give fractionally higher coverage. Table 1 assumes that high frequency words are known before lower frequency words and shows that knowing about 2,000 word families gives near to 80% coverage of written text. The same number of words gives greater coverage of informal spoken text - around 96% (Schonell, Meddleton and Shaw, 1956).

With a vocabulary size of 2,000 words, a learner knows 80% of the words in a text which means that 1 word in every 5 (approximately 2 words in every line) are unknown. Research by Liu Na and Nation (1985) has shown that this ratio of unknown to known words is not sufficient to allow reasonably successful guessing of the meaning of the unknown words. At least 95% coverage is needed for that. Research by Laufer (1989) suggests that 95% coverage is sufficient to allow reasonable comprehension of a text. A larger vocabulary size is clearly better. Table 2 is based on research by Hirsh and Nation (1992) on novels written for teenage or younger readers.

The Hirsh and Nation (1992) study looked at such novels because they might provide the most favourable conditions for second language learners to read unsimplified texts. These conditions could come about because they are aimed at a non-adult audience and thus there may be a tendency for the writer to use simpler vocabulary, and because a continuous novel on one topic by one writer provides opportunity for the repetition of vocabulary. Table 2 shows that under favourable conditions, a vocabulary size of 2000 to 3000 words provides a very good basis for language use.

<table>
<thead>
<tr>
<th>Vocabulary size</th>
<th>% coverage</th>
<th>Density of unknown words</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>99.5%</td>
<td>2000 words</td>
</tr>
<tr>
<td>3000</td>
<td>99.7%</td>
<td>3000 words</td>
</tr>
<tr>
<td>4000</td>
<td>99.9%</td>
<td>4000 words</td>
</tr>
<tr>
<td>5000</td>
<td>99.9%</td>
<td>5000 words</td>
</tr>
<tr>
<td>6000</td>
<td>99.9%</td>
<td>6000 words</td>
</tr>
</tbody>
</table>
To these research based arguments might be added the argument that most serious learners make use of such an approach. They can

strong.

This may be so, but the research evidence supporting the use of such an approach as one part of a vocabulary learning program is

teachers as a step back to outdated methods of learning and not in agreement with a communicative approach to language learning.

just to comprehend general texts. So how can we get learners to learn large amounts of vocabulary in a short space of time?

Na and Nation (1985) have shown that we need a vocabulary of about 3000 words which provides coverage of at least 95% of a text

The problem for beginning learners and readers is getting to the threshold where they can start to learn from context. Simply put, if one

is not the main focus.

learning include problem solving group work activities (Joe, Nation and Newton, 1996) and formal classroom activities where vocabulary

the earlier and intermediate levels of language learning, simplified reading books can be of great benefit. Other sources of incidental

1987). Extensive reading is a good way to enhance word knowledge and get a lot of exposure to the most frequent and useful words. At

context is so important that some studies suggest that first language learners learn most of their vocabulary in this way (Sternberg,

A way to manage the learning of huge amounts of vocabulary is through indirect or incidental learning. An example of this is learning

new words (or deepening the knowledge of already known words) in context through extensive listening and reading. Learning from

context is so important that some studies suggest that first language learners learn most of their vocabulary in this way (Sternberg,

efficient to spend class time on the strategies of (1) guessing from context, (2) using word parts and mnemonic techniques to remember

words, and (3) using vocabulary cards to remember foreign language - first language word pairs. Detailed description of these strategies

can be found in Nation (1990). Notice that although the teacher's focus is on helping learners gain control of important strategies, a

major function of these strategies is to help the learners to continue to learn new words and increase their vocabulary size.

The importance of this information is that although there are well over 54,000 word families in English, and although educated adult

native speakers know around 20,000 of these word families, a much smaller number of words, say between 3,000 to 5,000 word families

is needed to provide a basis for comprehension. It is possible to make use of a smaller number, around 2,000 to 3,000 for productive

use in speaking and writing. Hazenburg and Hulstijn (1996) however suggest a figure nearer to 10,000 for Dutch as a second language.

Sutarsyah, Nation and Kennedy (1994) found that a single long Economics text was made up of 5,438 word families and a corpus of

similar length made up of diverse short academic texts contained 12,744 word families. Within narrowly focused areas of interest, such

as in an Economics text, a much smaller vocabulary is needed than if the reader wishes to read a wide range of texts on a variety of

different topics.

How much vocabulary and how should it be learned?

We are now ready to answer the question "How much vocabulary does a second language learner need?" Clearly the learner needs to

know the 3,000 or so high frequency words of the language. These are an immediate high priority and there is little sense in focusing on

other vocabulary until these are well learned. Nation (1990) argues that after these high frequency words are learned, the next focus for

the teacher is on helping the learners develop strategies to comprehend and learn the low frequency words of the language. Because of

the very poor coverage that low frequency words give, it is not worth spending class time on actually teaching these words. It is more

efficient to spend class time on the strategies of (1) guessing from context, (2) using word parts and mnemonic techniques to remember

words, and (3) using vocabulary cards to remember foreign language - first language word pairs. Detailed description of these strategies

can be found in Nation (1990). Notice that although the teacher's focus is on helping learners gain control of important strategies, a

major function of these strategies is to help the learners to continue to learn new words and increase their vocabulary size.

A way to manage the learning of huge amounts of vocabulary is through indirect or incidental learning. An example of this is learning

new words (or deepening the knowledge of already known words) in context through extensive listening and reading. Learning from

context is so important that some studies suggest that first language learners learn most of their vocabulary in this way (Sternberg,

1987). Extensive reading is a good way to enhance word knowledge and get a lot of exposure to the most frequent and useful words. At

the earlier and intermediate levels of language learning, simplified reading books can be of great benefit. Other sources of incidental

learning include problem solving group work activities (Joe, Nation and Newton, 1996) and formal classroom activities where vocabulary

is not the main focus.

The problem for beginning learners and readers is getting to the threshold where they can start to learn from context. Simply put, if one
does not know enough of the words on a page and have comprehension of what is being read, one cannot easily learn from context. Liu

Na and Nation (1985) have shown that we need a vocabulary of about 3000 words which provides coverage of at least 95% of a text
before we can efficiently learn from context with unsimplified text. This is a large amount of startup vocabulary a learner needs, and this
just to comprehend general texts. So how can we get learners to learn large amounts of vocabulary in a short space of time?

The suggestion that learners should directly learn vocabulary from cards, to a large degree out of context, may be seen by some

teachers as a step back to outdated methods of learning and not in agreement with a communicative approach to language learning.

This may be so, but the research evidence supporting the use of such an approach as one part of a vocabulary learning program is

strong.

1 There is a very large number of studies showing the effectiveness of such learning in terms of amount and speed of learning. See Nation (1982), Paivio and Desrochers (1981) and Pressley et al. (1982) for a review of these studies.

2 Research on learning from context shows that such learning does occur but that it requires learners to engage in large amounts of reading and listening because the learning is small and cumulative (Nagy, Herman and Anderson, 1985). This should not be seen as an argument that learning from context is not worthwhile. It is by far the most important vocabulary learning strategy and an essential part of any vocabulary learning program. For fast vocabulary expansion, however, it is not sufficient by itself. There is no research that shows that learning from context provides better results than learning from word cards (Nation, 1982).

3 Research on the learning of grammar shows that form focused instruction is a valuable component of a language learning course (Ellis, 1990; Long, 1988). Courses with a form focused component achieve better results than courses without such a component. The important issue is to achieve a balance between meaning focused activities, form focused activities, and fluency development activities (Nation, forthcoming). Direct learning of vocabulary from cards is a kind of form focused instruction which can have the same benefits, perhaps even more markedly so, as form focused grammar instruction.

To these research based arguments might be added the argument that most serious learners make use of such an approach. They can
be helped to do it more effectively. There are other advantages for using word cards. They can give a sense of progress, and a sense of achievement, particularly if numerical targets are set and met. They are readily portable and can be used in idle moments in or out of class either for learning new words or revising old ones. They are specifically made to suit particular learners and their needs and are thus self motivating.

It should not be assumed that learning from word lists or word cards means that the words are learned forever, nor does it mean that all knowledge of a word has been learned. Learning from lists or word cards is only an initial stage of learning a particular word (see Schmitt and Schmitt, 1995 for further information). It is however a learning tool for use at any level of vocabulary proficiency. There will always be a need to have extra exposure to the words through reading, listening and speaking as well as extra formal study of the words, their collocates, associations, different meanings, grammar and so on. This shows a complementary relationship between contextualized learning of new words and the decontextualized learning from word cards.

What vocabulary does a language learner need?

The previous sections of this paper have suggested that second language learners need first to concentrate on the high frequency words of the language. In this section we look at some useful vocabulary lists based on frequency and review the research on the adequacy of the General Service List (West, 1953). Most counts also consider range, that is the occurrence of a word across several subsections of a corpus (McIntosh, Halliday and Strevens, 1961).

The practice of counting words has a long history dating as far back as Hellenic times (DeRocher, 1973). Several early word counts are mentioned in Fries and Traver (1960). There are many lists of the most frequently occurring words in English and a few of the most well known are described here.

The General Service List (West, 1953) The GSL contains 2000 headwords and was developed in the 1940s. The frequency figures for most items are based on a 5,000,000 word written corpus. Percentage figures are given for different meanings and parts of speech of the headword. In spite of its age, some errors, and its solely written base, it still remains the best of the available lists because of its information about the frequency of meanings, and West's careful application of criteria other than frequency and range.

The Teachers Word Book of 30,000 words (Thorndike and Lorge, 1944) This list of 30,000 lemmas (or about 13,000 word families (Goulden, Nation and Read, 1990)) is based on a count of an 18,000,000 word written corpus. Its value lies in its size. It is based on a large corpus and contains a large number of words. However, it is old, based on counts done over sixty years ago.

The American Heritage Word Frequency Book (Carroll, Davies and Richman, 1971) This comprehensive list is based on a corpus of 5,000,000 running words drawn from written texts used in United States schools over a range of grades and over a range of subject areas. The main values of the list are its focus on school texts and its listing of range figures, namely the frequency of each word in each of the school grade levels and in each of the subject areas.

The Brown (Francis and Kucera, 1982), LOB and related corpora There are now several 1,000,000 word written corpora each representing a different dialect of English. Some of these have published lemmatized word lists ranked according to frequency.

The classic list of high frequency words is Michael West's General Service List (1953). The 2000 word GSL is of practical use to teachers and curriculum planners as it contains words within the word family each with its own frequency. For example, excited, excites, exciting and excitement came under the headword excite. The GSL was written so that it could be used as a resource for compiling simplified reading texts into stages or steps. West and his colleagues produced vast numbers of simplified readers using this vocabulary. This is actually a very old list being based on frequency studies done in the early decades of this century. Doubts have been cast on its adequacy because of its age (Richards, 1974) and the relatively poor coverage provided by the words not in the first 1000 words of the list (Engels, 1968).

Engels makes two major points. Even if a limited vocabulary covers 95% of a text, a much larger vocabulary is still needed to cover the remaining 5% (p. 215). However Engels overestimates the size of this vocabulary. He suggests 497,000 words. His second point is that the limited vocabulary chosen by West (1953) is not the best selection. Engels examined 10 texts of 1000 words each. He found that West's GSL plus numbers covered 81.8% of the running words (This did not include proper nouns which covered 4.13%). Engels' definition of what should be included in a word family did not agree with West's and so Engels considered that West's GSL contained 3,372 words. This is because Engels considered flat and flatten, and police and policeman to be different word families. West gives separate figures for such items but indicates through the format of the GSL that they are in the same family. This difference however does not influence results. Engels considered the first 1000 of the GSL to be a good choice because the words were of high frequency and wide range (p. 221).

Engels correctly points out that the GSL does not provide 95% coverage of texts. He also says that the words outside the first 1000 of the GSL are "fallacious ... [because] they cannot be called general service words". Engels considers that the range and frequency of these words are too low to be included in the list. He suggests that for the lower frequency words in the GSL "the work should be done all over again" (p. 226), giving more attention to topic and genre divisions. Hwang and Nation (1995) report on such a study. The results only partly support Engels' ideas. It is possible to replace 452 of the words in the GSL with 250 words of higher frequency across a range of genres, but the change in total text coverage is small - from 82.3% to 83.4%. Even adjusting for the difference in size of the GSL, 2,147 words, and the new list, 1,945 words, still leaves the percentage difference in coverage at 1.68%. Thus although the GSL is
in need of replacement because of its age, errors it contains, and its written focus, it is still the best available list, given the range of information it contains about the relative frequency of the meanings of the words. In a variety of studies (Hwang, 1989; Hirsh and Nation, 1992; Sutarsyah, Nation and Kennedy, 1994) the GSL has provided coverage of 78% to 92% of various kinds of written text, averaging around 82% coverage.

Engels (1968) criticized the low coverage of the words not in the first 1000 words of the list. He found that whereas the first 1000 words covered 73.1% of the running words in the ten one thousand word texts he looked at, the words in the GSL outside the first 1000 covered only 7.7% of the running words. Other researchers have found a similar contrast.

Table 3: Coverage of first and second 1000 words of the GSL

<table>
<thead>
<tr>
<th>Researchers</th>
<th>1st 1000</th>
<th>2nd 1000</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sutarsyah (1993)</td>
<td>74.1%</td>
<td>4.3%</td>
<td>78.4%</td>
</tr>
<tr>
<td>academic texts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a long economics</td>
<td>77.7%</td>
<td>4.8%</td>
<td>82.5%</td>
</tr>
<tr>
<td>text</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hwang (1989)</td>
<td>77.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a range of texts</td>
<td></td>
<td>4.9%</td>
<td>82.1%</td>
</tr>
<tr>
<td>Hirsh (1992)</td>
<td></td>
<td>84.8%</td>
<td></td>
</tr>
<tr>
<td>short novels</td>
<td></td>
<td>5.8%</td>
<td>90.6%</td>
</tr>
</tbody>
</table>

What is also interesting is the number of different words (word types) from the second 1000 that actually occurred in a mixture of different kinds of texts compared with more homogeneous texts. In any one text, such as a novel or a textbook, around 400 to 550 of the second 1000 words from the GSL actually occurred. When a mixture of texts was looked at however around 700 to 800 of the second 1000 words occurred (Hirsh and Nation, 1992; Sutarsyah, Nation and Kennedy, 1994).

The second 1000 words behave in this way because they are lower frequency words than the first 1000 words and have a narrower range of occurrence. That is their occurrence is more closely related to the topic or subject area of a text than the wide ranging more general purpose words in the first 1000. But given a range of topics and genres, and enough texts, the second 1000 words are more generally useful than other lists of words.

After the 2000 high frequency words of the GSL, what vocabulary does a second language learner need? The answer to this question depends on what the language learner intends to use English for. If the learner has no special academic purpose then the learner should work on the strategies for dealing with low frequency words. If however the learner intends to go on to academic study in upper high school or at university, then there is a clear need for general academic vocabulary. This can be found in the 836 word list called the University Word List (UWL) (Xue and Nation, 1984; Nation, 1990).

The UWL consists of words that are not in the first 2000 words of the GSL but which are frequent and of wide range in academic texts. Wide range means that the words occur not just in one or two disciplines like economics or mathematics, but occur across a wide range of disciplines. The word frustrate for example which is in the UWL can be found in many different disciplines. The UWL is really a compilation from four separate studies, Lynn (1973), Ghadessy (1979), Campion and Elley (1971), and Praninskas (1972). Here are some items from it.

- accompany
- formulate
- index
- major
- objective
- biology
- genuine
- indicate
- maintain
- occur
- comply
- hemisphere
- individual
- maximum
- passive
- deficient
- homogeneous
- job
- modify
- persist
- edit
- identify
- labour
- negative
- quote
- feasible
- ignore
- locate
- notion
- random

(Nation, 1990)

The value of the UWL can be seen when we look at the coverage of academic text that it provides.

Table 4: Coverage by first 2000 of the GSL and the University Word List

<table>
<thead>
<tr>
<th>Researchers</th>
<th>1st 2000</th>
<th>UWL</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4 shows that for academic text, knowing the UWL makes the difference between approximately 80% coverage of a text (1 unknown word in every 5 words) and 90% coverage (1 unknown word in every 10 words).

Table 5 derived from Hwang (1989) shows the somewhat specialized nature of the UWL.

<table>
<thead>
<tr>
<th>Source</th>
<th>1st 2000 (GSL)</th>
<th>UWL</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic</td>
<td>78.1%</td>
<td>8.5%</td>
<td>86.6%</td>
</tr>
<tr>
<td>Newspapers</td>
<td>80.3%</td>
<td>3.9%</td>
<td>84.2%</td>
</tr>
<tr>
<td>Popular magazines etc.</td>
<td>82.9%</td>
<td>4.0%</td>
<td>86.9%</td>
</tr>
<tr>
<td>Fiction</td>
<td>87.4%</td>
<td>1.7%</td>
<td>89.1%</td>
</tr>
</tbody>
</table>

Note the low coverage the UWL has of fiction. Newspapers and magazines which are more formal make use of more of the UWL. Very formal academic text makes the greatest use of the UWL. The UWL is thus a word list for learners with specific purposes namely academic reading. The purpose behind the setting up of the UWL is to create a list of high frequency words for learners with academic purposes, so that these words can be taught and directly studied in the same way as the words from the GSL can.

Word frequency lists

The major theme of this paper has been that we need to have clear sensible goals for vocabulary learning. Frequency information provides a rational basis for making sure that learners get the best return for their vocabulary learning effort. Vocabulary frequency lists which take account of range have an important role to play in curriculum design and in setting learning goals.

This does not necessarily mean that learners must be provided with large vocabulary lists as the major source of their vocabulary learning. It does mean however that course designers should have lists to refer to when they consider the vocabulary component of a language course, and teachers need to have reference lists to judge whether a particular word deserves attention or not, and whether a text is suitable for a class.

The availability of powerful computers and very large corpora now make the development of such lists a much easier job than it was when Thorndike and Lorge (1944) and their colleagues manually counted 18,000,000 running words. The making of a frequency list however is not simply a mechanical task, and judgements based on well established criteria need to be made. The following list suggests several of the factors that would need to be considered in the development of a resource list of high frequency words.

1. **Representativeness**
   The corpora that the list is based on should adequately represent the wide range of uses of language. In the past, most word lists have been based on written corpora. There needs to be a substantial spoken corpus involved in the development of a general service list. The spoken and written corpora used should also cover a range of representative text types. Biber's (1990) corpus studies have shown how particular language features cluster in particular text types. The corpora used should contain a wide range of useful types so that the biases of a particular text type do not unduly influence the resulting list.

2. **Frequency and range**
   Most frequency studies have given recognition to the importance of range of occurrence. A word should not become part of a general service list because it occurs frequently. It should occur frequently across a wide range of texts. This does not mean that its frequency has to be roughly the same across the different texts, but means that it should occur in some form or other in most of the different texts or groupings of texts.

3. **Word families**
   The development of a general service list needs to make use of a sensible set of criteria regarding what forms and uses are counted as being members of the same family. Should *governor* be counted as part of the word family represented by *govern*? When making this decision, the purposes of the list and the learners for which it is intended need to be considered. As well as basing the decision on features such as regularity, productivity, and frequency (Bauer and Nation, 1993), the likelihood of learners seeing these relationships needs to be considered (Nagy and Anderson, 1984).

4. **Idioms and set expressions**
   Some items larger than a word behave like high frequency words. That is, they occur frequently as a unit (*Good morning*, *Never mind*), and their meaning is not clear from the meaning of the parts (*at once*, *set out*). If the frequency of such items is high enough to get them into a general service list in direct competition with single words, then perhaps they should be there. Certainly the arguments for idioms are strong, whereas set expressions could be included under one of their constituent words (but see

---

Hwang (1989)
academic texts  78.1%  8.5%  86.6%

Sutarsyah (1993)
an economics text  82.5%  8.7%  91.2%
5 **Range of information** To be of full use in course design, a list of high frequency words would need to include the following information for each word - the forms and parts of speech included in a word family, frequency, the underlying meaning of the word, variations of meaning and collocations and the relative frequency of these meanings and uses, and restrictions on the use of the word with regard to politeness, geographical distribution etc. Some dictionaries, notably the revised edition of the COBUILD dictionary, include much of this information, but still do not go far enough. This variety of information needs to be set out in a way that is readily accessible to teachers and learners.

6 **Other criteria** West (1953: ix) found that frequency and range alone were not sufficient criteria for deciding what goes into a word list designed for teaching purposes. West made use of ease or difficulty of learning (it is easier to learn another related meaning for a known word than to learn another word), necessity (words that express ideas that cannot be expressed through other words), cover (it is not efficient to be able to express the same idea in different ways. It is more efficient to learn a word that covers a quite different idea), stylistic level and emotional words (West saw second language learners as initially needing neutral vocabulary). One of the many interesting findings of the COBUILD project was that different forms of a word often behave in different ways, taking their own set of collocates and expressing different shades of meaning (Sinclair, 1991). Careful consideration would need to be given to these and other criteria in the final stages of making a general service list.

With a continuing emphasis on communication in language teaching there is a tendency to give less attention to the selection and checking of language forms in course design. Now that the benefits of form focused instruction are being positively reassessed, we may see a change in attitude towards vocabulary lists and frequency studies. The benefits of giving attention to principles of selection and gradation in teaching however remain important no matter what approach to teaching is being used. A goal of this review of the findings of research on vocabulary size and frequency is to show that this information can result in considerable benefits for both teachers and learners.

**References**


University of Wellington, New Zealand.


Nation, I. S. P. forthcoming. Teaching Listening and Speaking.


Contact Info:
Rob Waring
Notre Dame Seishin University, 2-16-9 Ifuku-cho, Okayama, Japan 700
Tel 086 252 1155 Fax 255 7663 Home 086 223 0341
Email: Rob Waring

Return to Main menu of papers
The “size” of a particular language is to calculate the total number of words in its largest dictionary. From that, determine the number of words in current usage by measuring (somehow) the words most frequently used correctly in samples of users at various ages and educational levels. (Credit: Eika Dopludo via Vocabulary Size, New York Times Magazine). Muscle ranking (number of words, largest dictionary) In this connection, I encourage teachers to teach kids words in context rather than learning parrot-fashion from wordlists. The more students passing exams, the better will be the teacher’s career prospects — no kidding! (via What's your magic number? - Rockwealth LLP).