How Many High Frequency Words are There in English?

I .S. P. Nation
Victoria University of Wellington

The distinction between high frequency and low frequency words is critical in language teaching and course design, because teachers should deal with high frequency words in different ways from low frequency words. High frequency words deserve direct teaching and classroom time. Low frequency words do not. Teachers should concentrate on strategy development for dealing with low frequency words. This study looks at five different ways of deciding the boundary between high and low frequency words. These include text coverage, cost/benefit analysis, overlap of different lists, total number of words, and criteria based on meaning and use. These different viewpoints roughly confirm that the group of high frequency words should consist of about 2,000 word families.

In countries like Thailand, Finland and Indonesia, where English is taught as a foreign language, learners have limited contact with English outside the classroom and have limited time at school to study English. It is thus critically important that learners get the best return from what they focus on and that classroom time is used well.

The distinction between high frequency and low frequency words is made to distinguish the words that teachers should spend class time on from the words that do not deserve the teacher's attention. Teachers need to monitor and assess the learning of the high frequency words, because these words account for such a large proportion of language use.

The group classified as high frequency words needs to be small enough for most of them to get some attention from the teacher over the course of a long term English programme. The words need to be frequent enough to justify the class time spent on them. They also

M. GILL, A. W. JOHNSON, L. M. KOSKI, R. D. SELL, B. WÅRVIK (EDS.). 2001.
LANGUAGE, LEARNING, LITERATURE: STUDIES PRESENTED TO HÅKAN RINGBOM
(ENGLISH DEPARTMENT PUBLICATIONS 4). TURKU: ÅBO AKADEMI UNIVERSITY, 167-181.

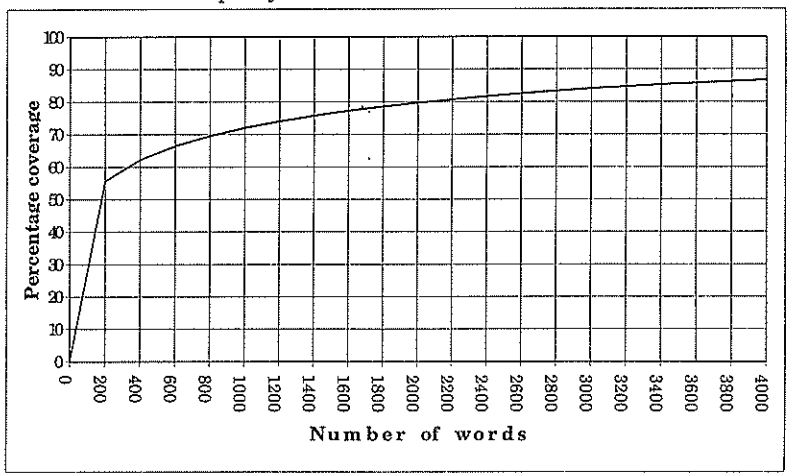
need to be of wide general use so that they are of value to learners with a wide range of eventual uses of the language.

A list of high frequency words, like West's (1953) *General Service List of English Words* (GSL), provides an important basis for deciding what goes into language courses and supplementary materials, for designing graded reader schemes, and for providing a trusted source for direct vocabulary learning.

Frequency studies show that there is no clear dividing line between high and low frequency words. Figure 1 is a graph of the frequency of lemmas in the *Brown Corpus* (Francis and Kučera 1982). Although there is a very sharp drop in frequency initially, the curve does not immediately suggest an obvious point where high frequency words end and low frequency words begin. Kilgarriff (1997: 149) points out that it is easier to draw the dividing line between the most frequent 500 words and the second most frequent 500 words than it is to draw the line, for example, between the most frequent 2,000 words and the next most frequent 500 words. This is because there are not many words around the frequency level of the dividing line between the first 500 and second 500 words because the words are of very high and quite different frequencies. Around the 2,000 word dividing line there are many words which have similar frequencies. Figure 1 shows this based on data from Table 3 using figures from the Brown corpus.

There are several ways of determining the dividing line. Each way is based on a particular kind of evidence which sometimes reflects a particular purpose for making the high frequency/low frequency distinction. In this paper, we will look at five different ways of deciding on the dividing line. The reason for doing this is to clearly delimit a high frequency vocabulary that should be systematically covered in course books, graded readers, classroom activities, and independent learning. At the same time the discussion will also underline the importance and purpose of a high frequency list.

Figure 1 Cumulative percentage text coverage by decreasing frequency levels



Reaching 95% Coverage of Unsimplified Text

Table 1 shows the importance of the 95% coverage point. When the learners have a vocabulary which is large enough to reach 95% coverage of the running words in a text, then there is, on average, one unknown word in every 20 running words. The 19 known words provide enough context for largely successful guessing from context and comprehension. Laufer (1989) and Liu and Nation (1985) provide some evidence to support this 95% threshold.

Note that where there is only 90% coverage of a text, then one word in every 10 or one word per line will be unfamiliar. Gaining 95% doubles the amount of context available. Each 1% increase in coverage after that makes very significant changes.

Table 1 The Number of Unfamiliar Tokens per 100 Tokens and the Number of Lines of Text Containing One Unfamiliar Word

% text coverage	Ratio of unfamiliar to familiar tokens	Number of text lines per 1 unfamiliar word
99	1 in 100	10.0
98	1 in 50	5.0
97	1 in 33	3.3
96	1 in 25	2.5
95	1 in 20	2.0
90	1 in 10	1.0
80	1 in 5	0.5

One way of deciding on the high frequency words is to define them as the group of words that provide 95% coverage of text. This is the approach taken in the second edition of the *COBUILD Dictionary* (Sinclair 1995). Table 2 shows the cumulative coverage of a wide variety of pieces of written text.

Table 2 Percentage Coverage of Tokens by each Successive Frequency Ranked 1000 Words in the Lemmatised Count of the Brown Corpus (Francis and Kučera 1982)

1,000 word (lemma) level	Percentage coverage of text (tokens)
1,000	72.0%
2,000	79.7%
3,000	84.0%
4,000	86.7%
5,000	88.6%
6,000	89.9%

Table 2 shows that over 6,000 lemmas are needed to cover 90%. Evidence from the *Brown* corpus indicates that around 14,000 word

types are needed to cover 95% of the tokens. This figure is far too large for a group of words that could feasibly be dealt with in a long-term English programme. Clearly, learners who wish to read very diverse types of text need to have a large vocabulary size to do this with ease. Hazenburg and Hulstijn (1996) found roughly comparable figures for written Dutch.

These figures overestimate the vocabulary size needed because they are based on a very diverse corpus of short texts and do not represent an individual's typical reading. A much smaller vocabulary size is needed to read in a more limited range of areas (Sutarsyah, Nation and Kennedy 1994).

The most frequent 2,000 words of English plus an academic word list provide coverage of about 87% of general academic text and 91% of an economics text (Sutarsyah, Nation and Kennedy 1994). For learners of English with academic purposes, the most frequent 2,000 words of English plus the academic word list represent the high frequency words of the language. These combined with proper nouns would provide close to 95% coverage.

We have however only looked at one of the most difficult kinds of language use, reading academic text. When we look at less formal uses of the language, such as informal conversation, reading newspapers, reading novels, we find that fewer words are needed to get good coverage. Schonell, Meddleton and Shaw (1956: 73) found that the most frequent 1,007 word families in their study of the oral vocabulary of the Australian worker covered 94% of the tokens. Hirsh and Nation (1992) found that in novels written for teen-age native speakers of English, the most frequent 2,000 word families covered 90% of the running words.

Total coverage with a 95% or reduced goal is a useful way to decide on the size of the high frequency words. It has strong validity as a procedure because the purpose of vocabulary study is to make language use, such as academic reading, easier. However, it seems

necessary for academic texts to use both general service and specialised lists plus proper nouns to reach the desired coverage.

Using the Cost / Benefit of Increase in Coverage

Another way to determine the end point of high frequency vocabulary is to perform a kind of cost/benefit analysis. The cost is the teaching effort required to deal with a certain number of words, for example a group of 50 or 100 words. The benefit is the text coverage provided by each group of words. For example, the most frequent 100 words of English provide about 50% text coverage. The 21st most frequent 100 words (words with a frequency rank of 2,001 – 2,100) provide approximately 0.5% of text coverage. Clearly in terms of cost/benefit the most frequent 100 words provide greater benefit than the 21st most frequent 100 words.

For the sake of simplicity, let us assume that each group of words of similar size, say each group of 100 words, involves the same teaching cost. For the very high frequency words this is clearly not true - there is much more to know about the very high frequency words than the less frequent words. However, the assumption is probably true enough for groups of words at the high frequency/low frequency boundary. The critical factor in making a decision about the high frequency/low frequency boundary will be the amount of coverage provided. Table 3 provides percentage coverage figures for successive groups of 100 words.

At the 2,000-word level, 100 lemmas cover just over 0.5% (.5659) of the tokens. On a 300-running-word page of a book, learning these 100 words would on average give access to an additional 1.5 words. Note also that the change in frequency of each lemma around that level is greater than 1 (see Column 2) – 56, then 53, then 50 and so on. The change in percentage coverage between the groups of 100 lemmas (see Column 5) is greater than 0.02%. At the 3,000-word level 100 lemmas cover just over 0.33% of the tokens (one word per 300-word page), the frequency change in Column 2 is 1, and the change in

Table 3 Coverage Data from the Brown Corpus (Francis and Kučera 1982)

Rank	No. of occurrences of the particular lemma at each rank given in column 1	Cumulative coverage	% cumulative coverage	Text coverage per 100 lemmas, i.e. % increase
200	464	564,422	55.6520	5.8186
300	330	602,762	59.4323	3.7803
400	260	631,985	62.3137	2.8814
500	210	655,219	64.6046	2.2909
600	176	674,361	66.4920	1.8874
700	155	690,713	68.1043	1.6123
800	138	705,307	69.5433	1.4390
900	124	718,348	70.8291	1.2858
1,000	113	730,263	72.0039	1.1748
1,100	103	741,044	73.0669	1.0630
1,200	94	750,838	74.0326	0.9657
1,300	87	759,906	74.9267	0.8941
1,400	80	768,267	75.7511	0.8244
1,500	75	776,042	76.5177	0.7666
1,600	70	783,258	77.2292	0.7115
1,700	66	790,044	77.8983	0.6691
1,800	63	796,487	78.5336	0.6353
1,900	59	802,604	79.1368	0.6032
2,000	56	808,344	79.7027	0.5659
2,100	53	813,773	80.2380	0.5353
2,200	50	818,896	80.7431	0.5051
2,300	47	823,745	81.2213	0.4782
2,400	45	828,364	81.6767	0.4554
2,500	43	832,792	82.5278	0.4366
2,600	41	836,996	82.5278	0.4145
2,700	39	841,013	82.9239	0.3961
2,800	37	844,847	83.3019	0.3780
2,900	36	848,504	83.6625	0.3606
3,000	34	851,985	84.0057	0.3432
3,100	33	855,318	84.3344	0.3287
3,200	31	858,513	84.6494	0.3150
3,300	30	861,593	84.9531	0.3037
3,400	29	864,549	85.2445	0.2914
3,500	28	867,378	85.5235	0.2790
3,600	26	870,086	85.7905	0.2670
3,700	26	872,686	86.0468	0.2533
3,800	25	875,187	86.2934	0.2466
3,900	24	877,610	86.5323	0.2389
4,000	23	879,952	86.7633	0.2310

percentage coverage is less than 0.02%. This small frequency change indicates that there is little frequency difference between the words in adjoining 100-word levels.

These arguments suggest a high frequency cut-off point at the 3,000-word level. The weaknesses of this argument are that the number of words per page is based on a 300-word page (why not a 200-word or 400-word page?), and that the actual frequency figures in Column 2 are based on a 1,014,000-word corpus. If the corpus was larger then the frequency figures would be higher and the change of actual frequency at each level would be greater. The percentage change figures (Column 5) however should be substantially the same with a different-sized corpus and thus should carry the greatest weight.

Choosing the Overlap of Competing Lists

A third way of deciding on the group of high frequency words is to look at several frequency counts and see where they agree or overlap with each other. This approach involves a kind of triangulation, that is looking at the same thing from different viewpoints. Where the different viewpoints give the same result we can feel a high degree of certainty about it. This is in effect using *range* as a major criterion for the inclusion or exclusion of items. A truly general service list would be made by (1) having several lists to compare, (2) having the corpora that the lists are based on be quite different from each other, and (3) having the variety of the corpora represent the variety of common uses of the language. The more various the corpora, the more likely that the resulting overlapping list will be quite small - probably around 1000 words.

Nation and Hwang (1995) examined the overlap between three word lists based on written texts – the 2,000 headword *GSL* (West 1953), the 1,810 words occurring in 10 or more of the 15 sections of the *LOB Corpus*, and 2,410 words occurring in 10 or more of the 15 sections of the *Brown Corpus*. Note that using the range figure of 10

sections means that each word had to occur in at least one of the "imaginative" sections (K-R) of the corpus. Table 4 shows the overlap between the three lists.

Table 4 Overlap of the Words in the *GSL, LOB* and *Brown* Lists

In all three lists	1,331
In only two lists	
Brown/LOB	250
Brown/GSL	226
LOB/GSL	138
Total	1,945

There were 452 words that occurred only in the *GSL*, 91 only in *LOB*, and 333 only in *Brown*. The *Brown* and *LOB* figures only refer to items with a range of 10 or more. Table 4 shows that it is possible to decide on a group of very high frequency, wide range words. This consists in the Nation and Hwang study of 1,331 words. It is also possible with less agreement to add another 614 items to this list. These items are of narrower range but they are still frequent enough and of wide enough range to justify their inclusion in a high frequency, wide range word list. This results in a list of 1,945 words – a very similar size to the 2,147 headword *GSL*. Table 5 gives the coverage by the *GSL* and the combined 1,945-word *LOB/Brown/GSL* list of the 1,014,000-running-word *LOB Corpus*.

Table 5 Coverage by the GSL and Nation and Hwang (1995) List of the *LOB Corpus*

List	Number of items	Percentage of text coverage
GSL	2,147	82.3%
Nation and Hwang list	1,945	83.4%

The *GSL* contains 200 more items and gives 1% less coverage. This is a significant but small difference. It suggests that beyond the 1,331 word level it is possible to make changes to the words in a high

frequency list, but that these changes have only a small return for text coverage.

It would be foolish to regard this 1,945 base list as a suitable list for learners with survival learning goals, and learners who want to learn the language predominantly for informal spoken use. The corpora on which the list is based do not match these purposes. However, for adult learners with reading as their learning goal, it is a very suitable base list.

Using the overlap between lists is a useful way of determining the high frequency, wide range words. It seems to result in a list of around 2,000 word families. The Carroll, Davies and Richman (1971: xxix) frequency count showed that there was a very high correlation of .8538 between word frequency and range. High frequency words are likely to be of wide range, while low frequency words have a narrow range. It may be that beyond the most frequent 2,000 words of English, words are of much narrower range and so it is difficult to develop a third thousand, and fourth thousand list of words that several different frequency counts agree on.

Using the Total Number of Words

Another way of deciding on the cut-off point between high and low frequency words is simply to consider what number of words could be sensibly dealt with in a language programme. Where English is taught as a foreign language, as in Japan or Indonesia, learners typically study English for about five hours a week for about 40 weeks a year for about five years – very approximately 1,000 hours. In such a programme you would expect learners to have become familiar with the high frequency words of the language and to have worked on strategies for dealing with low frequency words. The number of words in the high frequency group should be small enough to represent a feasible goal in such a programme. If learners leave school without control of the high frequency words then their learning will have been for little purpose.

This criterion for deciding the cut-off point does not have high validity but it has high practicality. It is thus not a strong criterion, but one that could be used in association with other criteria.

Studies of the amount of vocabulary learned over programmes of 1,000 hours or more indicate that it tends to be less than 2,000 words (Barnard 1961; Quinn 1968). Laufer (1998) found that over six years of instruction learners' receptive vocabulary size reached 1,900 words. By the end of the following year however it rocketed to 3,500 words. Similarly, productive vocabulary averaged 1,700 words after six years and rose to 2,550 at the end of the following year. The gap between receptive and productive vocabulary size increased as learners' proficiency developed. Waring (1997) found similar results.

These studies indicate that a productive vocabulary size of around 2,000 words is a possible though ambitious goal after 1,000 hours of study of English as a foreign language.

Using Criteria Based on Meaning and Use

Another way to decide on the equivalent of a high frequency vocabulary is to take an approach that looks more directly at reasons why a word may be useful. Frequency of occurrence, range and evenness of frequency of occurrence over a range of sub-corpora are measurements that look at the effects or symptoms of usefulness. More direct measures of the usefulness of words look at what words can be used to replace or define other words, what words can readily combine with other words, what words can be used to express a wide range of related meanings, and so on.

The most famous study of this type was Ogden and Richards's (Richards 1943) development of Basic English. This resulted in a list of 850 words, containing a deliberately small number of verbs.

Although Basic English received a lot of criticism (Carter 1987) and was not successful in being adopted as an international language of communication, the principles on which it was based are very sensible. There have been more recent attempts to outline and

operationalize the criteria for setting up a *core* vocabulary. Carter (1986, 1987) and Stubbs (1986) in very closely related articles describe a wide range of criteria which they suggest should all be applied as a way of more reliably distinguishing core (or *nuclear* or *basic*) vocabulary from other vocabulary. Carter makes the point that

> no single test will on its own be a sufficiently systematic measure, and [...] core vocabulary itself has no unambiguously clear boundaries.
>
> (Carter 1987: 186)

The tests for core vocabulary can be classified into two main types, those that are based on syntactic and semantic usefulness of core words (has many meanings, can define other words, form opposites, have many collocates), and those that are based on the neutrality of core words (are non-formal, non-topic related, non-culture specific). In general, core vocabulary is neutral, generic and non-restricted.

There has been no research on core vocabulary that indicates how large a list may be. A related kind of vocabulary list however is contained in the *Longman Language Activator* (Summers 1993). The *Longman Language Activator* is largely organised around 1,052 concepts or key words. This list does not include concrete nouns. It seems to have been arrived at largely through trial and error classification and grouping of words with checks against frequency counts. From this point of view it is a well tested list. It seems to satisfy several of the requirements of a core vocabulary, although its purpose is a little different. West's (1965) defining vocabulary has also been thoroughly tested in the preparation of dictionaries and contains 1,490 words.

Where should the High Frequency/Low Frequency Cut-off Point Be?

We have looked at several ways of deciding how many high frequency words there are in English, and Table 6 summarizes these. The available evidence indicates then that Michael West (1953) and his

colleagues made a very sensible decision when they set the *General Service List of English Words* at 2,000 word families.

Table 6 Ways, Purposes and Results of Five Approaches to Deciding on the High Frequency/Low Frequency Cut-off

Way of determining the cut off point	Particular purpose	Number of high frequency words
95% coverage (coverage)	Make reading text accessible	2,000+UWL+Proper nouns
Cost/benefit (frequency)	Make the best use of teaching time	2,000-3,000 words
Overlap of competing lists (range)	Choose the most widely used words for learning	1,300 or 1,945 words
Total number of words (quantity)	Feasible to cover in the time available	1,000 – 2,000 words
Meaning and use (quality)	Cover most useful concepts	850 words 1,052 words

Although it seems sensible to have a high frequency word list of 2,000 words, it is unlikely that a list can be made that will not be open to criticism regarding what is included and excluded. There are two main reasons for this. Firstly, and most importantly, the words in the list depend on the corpus on which it is based. Even if the corpus is very large, its nature still influences the resulting word list. For a general purpose, high frequency list, range is a very important criterion, and once again the different kinds of corpora that are used to represent the range of uses the learner will make of the language will influence what is included in the list. A different range of corpora would result in a slightly different list.

As a best guess, around 20% of the word families in a very well designed high frequency word list would change if another very well designed high frequency word list was made. In terms of text coverage, the changes would make a negligible difference.

Secondly, there may well be many items around the dividing line between high frequency and low frequency and the division would ultimately be an arbitrary one.

This means that teachers should limit the words that they spend teaching time on to about 2,000, but they should consider the goals and needs of their learners when they are determining what should be in the list. The flexibility of the list should not be taken as a reason for spending teaching time on more than 2,000 words. The return for effort is not justified.

References

Barnard, H. 1961. "A Test of P.U.C. Students' Vocabulary in Chotanagpur". *Bulletin of the Central Institute of English* 1: 90-100.

Bauer, L. and I. S. P. Nation. 1993. "Word Families". *International Journal of Lexicography* 6: 253-279.

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

Carter, R. 1986. "Core Vocabulary and Discourse in the Curriculum : A Question of the Subject". *RELC Journal* 17: 52-70.

Carter, R. 1987. "Is There a Core Vocabulary?" *Applied Linguistics* 8: 178-193.

Coxhead, A. 1998. *An Academic Word List* (Occasional Publication Number 18). Wellington: LALS, Victoria University of Wellington, New Zealand.

Francis, W. N. & H. Kučera, H. 1982. *Frequency Analysis of English Usage.* Boston: Houghton Mifflin Company.

Hazenberg, S. & J. H. Hulstijn. 1996. "Defining a Minimal Receptive Second-Language Vocabulary for Non-Native University Students: an Empirical Investigation". *Applied Linguistics* 17: 145-163.

Hirsh, D. & P. Nation. 1992. "What Vocabulary Size is Needed to Read Unsimplified Texts for Pleasure?" *Reading in a Foreign Language* 8: 689-696.

Kilgarriff, A. 1997. "Putting Frequencies in the Dictionary". *International Journal of Lexicography* 10: 135-155.

Laufer, B. 1989. "What Percentage of Text-Lexis is Essential for Comprehension?" In Lauren, C. & M. Nordman (eds). *Special Language: From Humans Thinking to Thinking Machines.* Clevedon: Multilingual Matters.

Laufer, B. 1998. "The development of Passive and Active Vocabulary: Same or Different?" *Applied Linguistics* 19: 255-271.

Liu Na & I. S. P. Nation. 1985. "Factors Affecting Guessing Vocabulary in Context". *RELC Journal* 16: 33-42.

Nation, I. S. P. & K. Hwang. 1995. "Where Would General service Vocabulary Stop and Special Purposes Vocabulary Begin?" *System* 23: 35-41.

Quinn, G. 1968. *The English Vocabulary of Some Indonesian University Entrants.* Salatiga: IKIP Kristen Satya Watjana.

Richards, I. A. 1943. *Basic English and its Uses.* London: Kegan Paul.

Schonell, F. J., I. G. Meddleton. & B. A. Shaw. 1956. *A Study of the Oral Vocabulary of Adults.* Brisbane: University of Queensland Press.

Sinclair, J. M. (ed. in chief). 1995. *Collins COBUILD Dictionary* (Second edition). London: Harper Collins.

Stubbs, M. 1986. "Language Development, Lexical Competence and Nuclear Vocabulary". In K. Durkin (ed.) *Language Development in the School Years.* London: Croom Helm.

Summers, D. (ed.). 1993. *Longman Language Activator.* Harlow: Longman.

Sutarsyah, C., P. Nation. & G. Kennedy. 1994. "How Useful is EAP Vocabulary for ESP? A Corpus Based Study". *RELC Journal* 25: 34-50.

Waring, R. 1997. "A Comparison of the Receptive and Productive Vocabulary Sizes of some Second Language Learners". *Immaculata (Notre Dame Seishin University, Okayama)* 1: 53-68.

West, M. 1935. *Definition Vocabulary* (Bulletin of the Department of Educational Research, University of Toronto, 4). Toronto: University of Toronto.

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

West, M. 1965. *An International Reader's Dictionary.* London: Longman.