1 *The goals of vocabulary learning*

How much vocabulary do learners need to know?

Whether designing a language course or planning our own course of study, it is useful to be able to set learning goals that will allow us to use the language in the ways we want to. When we plan the vocabulary goals of a long-term course of study, we can look at three kinds of information to help decide how much vocabulary needs to be learned: the number of words in the language, the number of words known by native speakers and the number of words needed to use the language.

How many words are there in the language?

The most ambitious goal is to know all of the language. However, even native speakers do not know all the vocabulary of the language. There are numerous specialist vocabularies, such as those of nuclear physics or computational linguistics, which are known only by the small groups who specialise in those areas. Still, it is interesting to have some idea of how many words there are in the language. This is not an easy question to resolve because there are numerous other questions which affect the way we answer it, including the following.

What do we count as a word? Do we count *book* and *books* as the same word? Do we count *green* (the colour) and *green* (a large grassed area) as the same word? Do we count people's names? Do we count the names of products like *Fab*, *Pepsi*, *Vegemite*, *Chevrolet*? The few brave or foolish attempts to answer these questions and the major question 'How many words are there in English?' have counted the number of words in very large dictionaries. *Webster's Third New International Dictionary* is the largest non-historical dictionary of English. It contains around 114,000 word families excluding proper names (Goulden, Nation and Read, 1990). This is a very large number and is well beyond the goals of most first and second language learners.

There are several ways of deciding what words will be counted.

Tokens

One way is simply to count every word form in a spoken or written text and if the same word form occurs more than once, then each occurrence of it is counted. So the sentence 'It is not easy to say it correctly' would contain eight words, even though two of them are the same word form, *it*. Words which are counted in this way are called 'tokens', and sometimes 'running words'. If we try to answer questions like 'How many words are there on a page or in a line?' 'How long is this book?' 'How fast can you read?' 'How many words does the average person speak per minute?' then our unit of counting will be the token.

Types

We can count the words in the sentence 'It is not easy to say it correctly' another way. If we see the same word again, we do not count it again. So the sentence of eight tokens consists of seven different words or 'types'. We count words in this way if we want to answer questions like 'How large was Shakespeare's vocabulary?' 'How many words do you need to know to read this book?' 'How many words does this dictionary contain?'

Lemmas

A lemma consists of a headword and some of its inflected and reduced (*n't*) forms. Usually, all the items included under a lemma are the same part of speech (Francis and Kučera, 1982: 461). The English inflections consist of plural, third person singular present tense, past tense, past participle, *-ing*, comparative, superlative and possessive (Bauer and Nation, 1993). The Thorndike and Lorge (1944) frequency count used lemmas as the basis for counting, and the more recent computerised count on the *Brown Corpus* (Francis and Kučera, 1982) has produced a lemmatised list. In the Brown count the comparative and superlative forms are not included in the lemma, and the same form used as a different part of speech (*walk* as a noun, *walk* as a verb) are not in the same lemma. Variant spellings (*favor*, *favour*) are usually included as part of the same lemma when they are the same part of speech.

Lying behind the use of lemmas as the unit of counting is the idea of learning burden (Swenson and West, 1934). The learning burden of an item is the amount of effort required to learn it. Once learners can use the inflectional system, the learning burden of for example *mends*, if

the learner already knows *mend*, is negligible. One problem in forming lemmas is to decide what will be done with irregular forms such as *mice*, *is*, *brought*, *beaten* and *best*. The learning burden of these is clearly heavier than the learning burden of regular forms like *books*, *runs*, *talked*, *washed* and *fastest*. Should the irregular forms be counted as a part of the same lemma as their base word or should they be put into separate lemmas? Lemmas also separate closely related items such as the adjective and noun uses of words like *original*, and the noun and verb uses of words like *display*. An additional problem with lemmas is what is the headword – the base form or the most frequent form? (Sinclair, 1991: 41-42).

Using the lemma as the unit of counting greatly reduces the number of units in a corpus. Bauer and Nation (1993) calculate that the 61,805 tagged types (or 45,957 untagged types) in the *Brown Corpus* become 37,617 lemmas which is a reduction of almost 40% (or 18% for untagged types). Nagy and Anderson (1984) estimated that 19,105 of the 86,741 types in the Carroll, Davies and Richman (1971) corpus were regular inflections.

Word families

Lemmas are a step in the right direction when trying to represent learning burden in the counting of words. However, there are clearly other affixes which are used systematically and which greatly reduce the learning burden of derived words containing known base forms. These include affixes like *-ly*, *-ness* and *un-*. A word family consists of a headword, its inflected forms, and its closely related derived forms.

The major problem in counting using word families as the unit is to decide what should be included in a word family and what should not. Learners' knowledge of the prefixes and suffixes develops as they gain more experience of the language. What might be a sensible word family for one learner may be beyond another learner's present level of proficiency. This means that it is usually necessary to set up a scale of word families, starting with the most elementary and transparent members and moving on to less obvious possibilities.

How many words do native speakers know?

A less ambitious way of setting vocabulary learning goals is to look at what native speakers of the language know. Unfortunately, research on measuring vocabulary size has generally been poorly done (Nation, 1993c), and the results of the studies stretching back to the late nine-

teenth century are often wildly incorrect. We will look at the reasons for this later in this book.

Recent reliable studies (Goulden, Nation and Read, 1990; Zechmeister, Chronis, Cull, D'Anna and Healy, 1995) suggest that educated native speakers of English know around 20,000 word families. These estimates are rather low because the counting unit is word families which have several derived family members and proper nouns are not included in the count. A very rough rule of thumb would be that for each year of their early life, native speakers add on average 1,000 word families a year to their vocabulary. These goals are manageable for non-native speakers of English, especially those learning English as a second rather than foreign language, but they are way beyond what most learners of English as another language can realistically hope to achieve.

How much vocabulary do you need to use another language?

Studies of native speakers' vocabulary seem to suggest that second language learners need to know very large numbers of words. While this may be useful in the long term, it is not an essential short-term goal. This is because studies of native speakers' vocabulary growth see all words as being of equal value to the learner. Frequency based studies show very strikingly that this is not so, and that some words are much more useful than others.

Table 1.1 shows part of the results of a frequency count of just under 500 running words of the Ladybird version of *The Three Little Pigs*. It contains 124 different word types.

Note the large proportion of words occurring only once, and the very high frequency of the few most frequent words. When we look at texts our learners may have to read and conversations that are like ones that they may be involved in, we find that a relatively small amount of well-chosen vocabulary can allow learners to do a lot. To see this, let us look at an academic reading text and examine the different kinds of vocabulary it contains. The text is from Neville Peat's (1987) *Forever the Forest. A West Coast Story* (Hodder and Stoughton, Auckland).

Sustained-*yield* management ought to be long-term government **policy** in *indigenous* forests *zoned* for production. The adoption of such a **policy** would represent a *breakthrough* – the boundary between a *pioneering*, **extractive phase** and an *era* in which the *timber* industry **adjusted** to living with the forests in *perpetuity*. A forest **sustained** is a forest in which harvesting and *mortality* combined do not **exceed** *regeneration*. Naturally enough, faster-growing forests produce more *timber*, which is why attention

Table 1.1. *An example of the results of a frequency count*

the	41	met	3	come	1
little	25	myself	3	door	1
pig	22	not	3	down	1
house	17	on	3	fell	1
a	16	pigs	3	go	1
and	16	please	3	grew	1
said	14	pleased	3	had	1
he	12	shall	3	hair	1
I	10	soon	3	here	1
me	10	stronger	3	him	1
some	9	that	3	houses	1
wolf	9	they	3	huff	1
build	8	three	3	knocked	1
't	8	want	3	live	1
third	8	who	3	long	1
was	8	with	3	mother	1
of	7	won	3	must	1
straw	7	yes	3	my	1
to	7	yours	3	next	1
you	7	big	2	off	1
man	6	by	2	once	1
second	6	care	2	one	1
catch	5	chin	2	puff	1
first	5	day	2	road	1
for	5	does	2	set	1
will	5	huffed	2	so	1
bricks	4	let	2	their	1
built	4	'm	2	them	1
himself	4	no	2	there	1
now	4	puffed	2	took	1
sticks	4	strong	2	up	1
than	4	take	2	upon	1
very	4	then	2	us	1
asked	3	time	2	walked	1
carrying	3	too	2	we	1
eat	3	along	1	went	1
gave	3	are	1	were	1
give	3	ate	1	which	1
his	3	blow	1	your	1
in	3	but	1	yourselves	1
it	3	came	1		
'll	3	chinny	1		

would tend to swing from *podocarps* to *beech* forests regardless of the state of the *podocarp* **resource**. The colonists cannot be blamed for *plunging* in without thought to whether the **resource** had limits. They brought from *Britain* little experience or understanding of how to **maintain** forest **structure** and a *timber* supply for all time. Under *German* management it might have been different here. The *Germans* have practised the **sustained approach** since the seventeenth century when they faced a *timber* shortage as a result of a **series** of wars. In *New Zealand* in the latter part of the twentieth century, an **anticipated** shortage of the most valuable native *timber*, *rimu*, prompts a **similar response** – no more **contraction** of the *indigenous* forest and a balancing of yield with *increment* in **selected areas**.

This is not to say the idea is being *aired* here for the first time. Over a century ago the first *Conservator* of Forests proposed **sustained** harvesting. He was cried down. There were far too many trees left to bother about it. And yet in the *pastoral* **context** the dangers of *overgrazing* were **appreciated** early in the piece. *New Zealand geography* students are taught to this day how *overgrazing* causes the *degradation* of the soil and hillsides to slide away, and that with them can go the *viability* of hill-country sheep and cattle farming. That a forest could be *overgrazed* as easily was not widely accepted until much later – so late, in fact, that the *counter* to it, **sustained**-*yield* management, would be forced upon the industry and come as a shock to it. It is a simple enough **concept** on paper: balance harvest with growth and you have a natural *renewable* **resource**; forest products forever. **Plus** the social and **economic benefits** of regular work and **income**, a regular *timber* supply and relatively **stable** markets. **Plus** the **environmental benefits** that *accrue* from **minimising** the **impact** on soil and water qualities and wildlife.

In practice, however, **sustainability** depends on how well the **dynamics** of the forest are understood. And these **vary** from **area** to **area** according to forest make-up, soil *profile*, *altitude*, *climate* and **factors** which forest science may yet discover. *Ecology* is deep-felt.

We can distinguish four kinds of vocabulary in the text: high-frequency words (unmarked in the text), academic words (in bold), and technical and low-frequency words (in italics).

High-frequency words

In the example text, these words are not marked at all and include function words: *in*, *for*, *the*, *of*, *a*, etc. Appendix 6 contains a complete list of function words. The high-frequency words also include many content words: *government*, *forests*, *production*, *adoption*, *represent*, *boundary*. The classic list of high-frequency words is Michael West's (1953a) *A General Service List of English Words* which contains around 2,000 word families. Almost 80% of the running words in the text are high-frequency words.

Academic words

The text is from an academic textbook and contains many words that are common in different kinds of academic texts: *policy*, *phase*, *adjusted*, *sustained*. Typically these words make up about 9% of the running words in the text. The best list of these is the *Academic Word List* (Coxhead, 1998). Appendix 1 contains the 570 headwords of this list. This small list of words is very important for anyone using English for academic purposes (see chapter 6).

Technical words

The text contains some words that are very closely related to the topic and subject area of the text. These words include *indigenous*, *regeneration*, *podocarp*, *beech*, *rimu* (a New Zealand tree) and *timber*. These words are reasonably common in this topic area but not so common elsewhere. As soon as we see them we know what topic is being dealt with. Technical words like these typically cover about 5% of the running words in a text. They differ from subject area to subject area. If we look at technical dictionaries, such as dictionaries of economics, geography or electronics, we usually find about 1,000 entries in each dictionary.

Low-frequency words

The fourth group is the low-frequency words. Here, this group includes words like *zoned*, *pioneering*, *perpetuity*, *aired* and *pastoral*. They make up over 5% of the words in an academic text. There are thousands of them in the language, by far the biggest group of words. They include all the words that are not high-frequency words, not academic words and not technical words for a particular subject. They consist of technical words for other subject areas, proper nouns, words that almost got into the high-frequency list, and words that we rarely meet in our use of the language.

Let us now look at a longer text and a large collection of texts.

Sutarsyah, Nation and Kennedy (1994) looked at a single economics textbook to see what vocabulary would be needed to read the text. The textbook was 295,294 words long. Table 1.2 shows the results. The academic word list used in the study was the *University Word List* (Xue and Nation, 1984).

What should be clear from this example and from the text looked at earlier is that a reasonably small number of words covers a lot of text.

Table 1.2. *Text coverage by the different kinds of vocabulary in an economics textbook*

Type of vocabulary	Number of words	Text coverage
1st 2000 word families	1,577	82.5%
Academic vocabulary	636	8.7%
Other vocabulary	3,225	8.8%
Total	5,438	100.0%

Table 1.3. *The coverage by the different kinds of vocabulary in an academic corpus*

Type of vocabulary	% coverage
1st 1000 words	71.4%
2nd 1000 words	4.7%
Academic Word List (570 words)	10.0%
Others	13.9%
Total	100.0%

Coxhead (1998) used an academic corpus made up of a balance of science, arts, commerce and law texts totalling 3,500,000 running words. Table 1.3 gives the coverage figures for this corpus.

Figure 1.1 presents the proportions in a diagrammatic form. The size of each of the sections of the right-hand box indicates the proportion of the text taken up by each type of vocabulary.

Table 1.4 gives the typical figures for a collection of texts consisting of five million running words.

Some very important generalisations can be drawn from Table 1.4 and the other information that we have looked at. We will look at these generalisations and at questions that they raise. Brief answers to the questions will be given here but will be examined much more closely in later chapters.

High-frequency words

There is a small group of high-frequency words which are very important because these words cover a very large proportion of the running words in spoken and written texts and occur in all kinds of uses of the language.

Sustained-*yield* management ought to be long-term government **policy** in *indigenous* forests *zoned* for production. The adoption of such a **policy** would represent a *breakthrough* – the boundary between a *pioneering*, **extractive phase** and an *era* in which the *timber* industry **adjusted** to living with the forests in *perpetuity*. A forest **sustained** is a forest in which harvesting and *mortality* combined do not **exceed** *regeneration*. Naturally enough, faster-growing forests produce more *timber*, which is why attention would tend to swing from *podocarps* to *beech* forests regardless of the state of the *podocarp* **resource**. The colonists cannot be blamed for *plunging* in without thought to whether the **resource** had limits. They brought from *Britain* little experience or understanding of how to **maintain** forest **structure** and a *timber* supply for all time. Under *German* management it might have been different here. The *Germans* have practised the **sustained approach** since the seventeenth century when they faced a *timber* shortage as a result of a **series** of wars. In *New Zealand* in the latter part of the twentieth century, an **anticipated** shortage of the most valuable native *timber*, *rimu*, prompts a **similar response** – no more **contraction** of the *indigenous* forest and a balancing of yield with *increment* in **selected areas**.

This is not to say the idea is being *aired* here for the first time. Over a century ago the first *Conservator* of Forests proposed **sustained** harvesting. He was cried down. There were far too many trees left to bother about it. And yet in the *pastoral* **context** the dangers of *overgrazing* were **appreciated** early in the piece. *New Zealand geography* students are taught to this day how *overgrazing* causes the *degradation* of the soil and hillsides to slide away, and that with them can go the *viability* of hill-country sheep and cattle farming. That a forest could be *overgrazed* as easily was not widely accepted until much later – so late, in fact, that the *counter* to it, **sustained**-*yield* management, would be forced upon the industry and come as a shock to it.

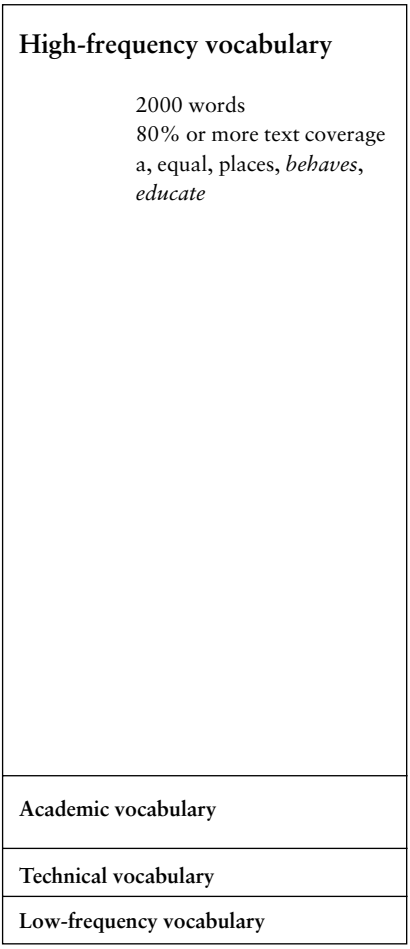

Figure 1.1 Vocabulary type and coverage in an academic text

How large is this group of words? The usual way of deciding how many words should be considered as high-frequency words is to look at the text coverage provided by successive frequency-ranked groups of words. The teacher or course designer then has to decide where the coverage gained by spending teaching time on these words is no longer worthwhile. Table 1.5 shows coverage figures for each successive 1,000 lemmas from the *Brown Corpus* – a collection of various 2,000-word texts of American English totalling just over one million tokens.

Usually the 2,000-word level has been set as the most suitable limit for high-frequency words. Nation and Hwang (1995) present

Table 1.4. *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

Table 1.5. *The percentage text coverage of each successive 1000 lemmas in the Brown Corpus*

1000 word (lemma) level	% coverage of text (tokens)
1000	72
2000	79.7
3000	84
4000	86.7
5000	88.6
6000	89.9

evidence that counting the 2,000 most frequent words of English as the high-frequency words is still the best decision for learners going on to academic study.

What are the words in this group? As has been noted, the classic list of high-frequency words is Michael West's *General Service List* which contains 2,000 word families. About 165 word families in this list are function words such as *a*, *some*, *two*, *because* and *to* (see appendix 6). The rest are content words, that is nouns, verbs, adjectives and adverbs. Older series of graded readers are based on this list.

How stable are the high-frequency words? In other words, does one properly researched list of high-frequency words differ greatly from another? Frequency lists may disagree with each other about the frequency rank order of particular words but if the research is based on a well-designed corpus there is generally about 80% agreement about