N. Schmitt (ed)
An Introduction to Applied Linguistics
Arnold, London (2002)

3 Vocabulary

Paul Nation
Victoria University of Wellington

Paul Meara
University of Wales, Swansea

What is Vocabulary?

One of the most difficult questions to answer in vocabulary studies is 'What is a word?' and there are a variety of only partly satisfactory answers depending on the reasons for asking the question. If we want to count how long a book is, or how fast someone can speak or read in words per minute, then we need to count 'tokens'. The sentence 'To be or not to be, that is the question' contains ten tokens. Even though the same word form *be* occurs twice, it is counted each time it occurs. When counting tokens, it is necessary to decide if we count items such as *I'm* or *we'll* as two tokens or one. If we are counting tokens in spoken language, do we count *um* and *er* as tokens, and do we count repetitions such as *I . . . I . . . I said* as tokens? We can only answer these questions by examining our reasons for counting.

Often, we are interested in how many different words someone knows or uses. For example, if we are interested in how much 'sight vocabulary' a learner has (words that are known well enough to be recognized quickly and accurately) then we would count word 'types'. The sentence 'To be or not to be, that is the question' contains eight word types. Both *be* and *to* occur twice, and so they are not counted after their first occurrence. Some of the problems with counting types include deciding what to do about capital letters (are *High* and *high* two types or one?). And, what do we do with identical types that have different meanings (*generation* (of electricity) and (the younger) *generation*).

If our reason for counting is related to vocabulary learning then we need to choose a unit of counting that reflects the kind of knowledge that language users draw on. There is evidence (Nagy *et al.*, 1989) that language users see closely related word forms (*mend*, *mends*, *mended*, *mending*) as belonging to the same word family and it is the total frequency of a word family that determines the familiarity of any particular member of that family. In other words, the regular word-building devices create items that are seen as being very closely related to each other. A major problem with counting word families is in deciding what should be counted as a member of a family. The most conservative way is to count 'lemmas'. A lemma is a set of related words that consists of the stem form and inflected forms that are all the same part of speech. So, *approach*, *approaches*, *approached*, *approaching* would all be members of the same lemma because they all have the same stem, include only the stem and inflected forms, and are all verbs. A less conservative definition of a word family would also include items made with derivational affixes, such as *un-* and *non-*, and suffixes, such as *-ness* and *-ly*. Bauer and Nation (1993) suggest that as learners become more proficient, the number of items included in their word families will also tend to increase.

There are some groups of words, such as *good morning* and *at the end of the day*, which seem to be used like single words. Some of the groups may be items that have not been analysed into parts but are just learned, stored and used as complete units. Others may be constructed from known parts but are used so often that it is more efficient to treat them as a single unit. Pawley and Syder (1983) suggest that native speakers speak appropriately and fluently because they have stored very large numbers of these 'multi-word units' (MWUs) which they can draw on when using the language. These MWUs go by several names:

- 'Preformulated language' (emphasizing how MWUs can be stored as single units which are 'ready to go').
- 'Formulas' (emphasizing how MWUs can be repeatedly used instead of having to generate new ways of saying things).
- 'Lexical phrases' (emphasizing how certain phrases are typically used to achieve particular functions in everyday life, for example *Have you heard the one about* ____ is commonly used to introduce a joke).

A key feature of these MWUs is that the words in the unit cannot be freely substituted with other words; rather they have strong partnership connections, a property called 'collocation'. Although we have always been aware of some MWUs, for example idioms, research into MWUs has only blossomed with the advent of corpus analysis, which has brought extended lexical patterning into the light (*see* Chapter 6, *Corpus Linguistics*). Because this is a new area, counting these MWUs is still difficult, as no defining criteria have yet gained general acceptance. So, what is considered to be a MWU will depend on the purpose of the counting. If the goal is to count items that would require learning for comprehension then the MWUs would need to be to some degree non-compositional (the meaning of the unit could not be inferred from the meaning of its parts). This criterion would result in a rather short list of high-frequency items, if the frequency cut-off point was the same as that for single words. That is, MWUs are much less frequent than single high-frequency words. If the goal of counting is to come up with a list of items that could contribute to fluency and a native-like turn of phrase then the MWUs need at least to be frequent and grammatically coherent. Only a small list of such items would get within the most frequent 2000 words and phrases of English (Nation, 2001).

What Vocabulary Should be Learned?

What vocabulary to focus on should be determined by two major considerations – the needs of the learners and the usefulness of the vocabulary items. The traditional way of measuring the usefulness of items is to discover their frequency and range in a relevant corpus. The most striking features of the results of a frequency-based study are:

- The very wide spread of frequencies, with some items occurring many many times and some occurring only once.
- The relatively small number of words needed to cover a very large proportion of the tokens in a text.
- The very large number of low frequency items that account for a very small proportion of the tokens in a text.

These three points are illustrated in Table 3.1 and Table 3.2. Table 3.1 is the result of a frequency count of a 500-token section of this chapter. The 500-word section contained 204 different word types which made up 169 word families. Table 3.2 lists the frequency, the number of words with that frequency, and the cumulative coverage of the tokens. In Table 3.1 not all the words occurring once or twice are listed because there were too many of them to show here.

Table 3.1 A frequency list of a 500-word text

The	22	Family	6
Of	18	What	6
To	17	If	5
And	16	Same	5
Is	16	Types	5
A	14	Vocabulary	5
That	12	All	4
We	12	Do	4
Word	11	I	4
Or	10	It	4
Are	8	Occurs	4
Be	8	Related	4
In	8	With	4
As	7	Words	4
Count	7	Counted	3
Counting	7	For	3
Tokens	7	Forms	3

Table 3.1 – *continued*

How	3	Each	2
Include	3	Form	2
Items	3	. . .	
Language	3	About	1
Like	3	Affixes	1
Not	3	After	1
One	3	Also	1
Only	3	Anderson	1
Question	3	Answers	1
Stem	3	Any	1
Then	3	Approach	1
Twice	3	Approached	1
Would	3	Approaches	1
Answer	2	Approaching	1
Can	2	Asking	1
Closely	2	Bauer	1
Conservative	2	Because	1
Contains	2	Being	1
Deciding	2	. . .	
Different	2		

Table 3.2 Number of words and coverage for each frequency

Frequency	Number of types	Cumulative coverage of text (%)
10 and above	10 word types	29.6
8 occurrences	3	34.4
7	4	40.0
6	2	42.4
5	4	46.4
4	8	52.8
3	16	62.4
2	32	75.2
1	125	100

By making frequency counts of large relevant corpora, it is possible to come up with lists of words that will be very useful for people in the early stages of learning a language. Several such lists exist and they provide a very useful basis for course design. The classic list of the most useful words of English is Michael West's (1953) *A General Service List of English Words* (GSL) which contains 2000 high-frequency words. There is plenty of evidence that 2000 words is an appropriate size for such a list, but the list needs to be based on a corpus where spoken language is well represented. The GSL is based on written language, and so needs to be updated by a new list based on both spoken and written discourse.

The information from frequency studies suggests a cost–benefit approach to dealing with vocabulary. If we use frequency counts to distinguish high-frequency from low-frequency words then it seems clear that the high-frequency words need to be the first and main vocabulary goal of learners. These words are so frequent, so widespread and make up such a manageable group that both teachers and learners can usefully spend considerable time ensuring that they are well learned. The low-frequency words are so infrequent, have such a narrow range of occurrence and make up such a large group that they do not deserve teaching time. Of course, learners need to keep on learning low-frequency words after they have learned the high-frequency words, but they should do this incidentally or deliberately in their own time. Teachers should focus on strategies that help learners do this 'incidental' or 'deliberate' learning. These strategies include guessing from context, learning from word cards, using word parts and dictionary use. We will look at these in more detail later in this chapter.

It is possible to increase the number of high-frequency words that teachers and learners should give attention to by looking at the needs of the learners and making special purposes vocabulary lists. The most useful of these lists is the Academic Word List (Coxhead, 2000) which is designed for learners who intend to do academic study through the medium of English. The list consists of 570 word families which account for 8.5–10% of the tokens in a wide range of academic texts. The list includes words such as *evaluate*, *invest*, *technology* and *valid*. These words are a very important learning goal for learners with academic purposes who have learned the high-frequency words of English. On average, there are 30 of these words on every page of an academic text. Some of these words have more than one largely unrelated meaning, for example *issue* ('problem'), *issue* ('produce, send out'), but almost invariably one of these meanings is much more frequent than the other.

How Should Vocabulary be Learned?

Many teachers would assume that vocabulary learning stems mainly from the direct teaching of words in the classroom. However, vocabulary learning needs to be more broadly based than this. Let us look at four strands of vocabulary learning in turn.

Learning Vocabulary from Meaning-focused Input (Listening and Reading)

Learning from meaning-focused input, that is, learning incidentally through listening and reading, accounts for most first language vocabulary learning. Although this kind of learning is less sure than deliberate study, for native speakers there are enormous opportunities for such learning (Nagy, Herman and Anderson, 1985). For such learning to occur with non-native speakers, three major conditions need to be met. First, the unknown vocabulary should make up only a very small proportion of the tokens, preferably around two per cent, which would mean one unknown word in 50 (Hu and Nation, 2000; *see* Chapter 13, *Reading*). Second, there needs to be a very large quantity of input, preferably one million tokens or more per year. Third, learning will be increased if there is more deliberate attention to the unknown vocabulary through the occurrence of the same vocabulary in the deliberate learning strand of the course and through consciousness-raising of unknown words as they occur through glossing (Watanabe, 1997), dictionary use and highlighting in the text. It is important to remember that incidental learning is cumulative and therefore vocabulary needs to be met a number of times to allow the learning of each word to become stronger and to enrich the knowledge of each word.

The core of the meaning-focused input strand of a course is a well-organized, well-monitored, substantial extensive reading programme based largely, but not exclusively, on graded readers (for substantial reviews, *see* Waring, 1997a; Day and Bamford, 1998). Graded readers are particularly helpful for learners in the beginning and intermediate stages, as they best realize the three conditions for learning outlined above. Typically, a graded reader series begins with books about 5000 words long written within a 300–500-word family vocabulary. These go up in four to six stages to books about 25,000–35,000 words long written within a 2000–2500-word family vocabulary. Nation and Wang (1999) estimate that second language learners need to be reading at least one graded reader every two weeks in order for noticeable learning to occur. In the past, graded readers have been accused of being inauthentic reduced versions of texts which do not expose learners to the full richness of the English language, and are poorly written. These criticisms all have a grain of truth in them, but they are now essentially misinformed. There are currently some very well-written graded readers which have key advantages: even beginning and intermediate learners with limited vocabulary sizes can read simplified readers for pleasure, which is an authentic usage, even if the text itself is not purely 'authentic'. Learners find it impossible to respond authentically to texts that overburden them with unknown vocabulary.

Listening is also a source of meaning-focused input and the same conditions of low unknown vocabulary load, quantity of input and some deliberate attention to vocabulary are necessary for effective vocabulary learning. Quantity of input, which directly affects repetition, may be partly achieved through repeated listening, where learners listen to the same story several times over several days. Deliberate attention to vocabulary can be encouraged by the teacher quickly defining unknown items (Elley, 1989), noting them on the board or allowing learners the opportunity to negotiate their meaning by asking for clarification (Ellis, 1994, 1995; Ellis and Heimbach, 1997; Ellis and He, 1999). Newton (1995) found that although negotiation is a reasonably sure way of vocabulary learning, the bulk of vocabulary learning was through the less sure way of non-negotiated learning from context, simply because there are many more opportunities for this kind of learning to occur.

Learning Vocabulary from Meaning-focused Output (Speaking and Writing)

Learning from meaning-focused output, that is, learning through speaking and writing, is necessary to move receptive knowledge into productive knowledge. This enhancement of vocabulary through the productive skills can occur in several ways. First, activities can be designed, such as those involving the use of annotated pictures or definitions, which encourage the use of new vocabulary. Second, speaking activities involving group work can provide opportunities for learners to negotiate the meanings of unknown words with each other. Such negotiation is often successful and positive (Newton, 1995). Third, because the learning of a particular word is a cumulative process, the use of a partly known word in speaking or writing can help strengthen and enrich knowledge of the word.

Joe, Nation and Newton (1996) describe guidelines for the design of speaking activities that try to optimize vocabulary learning by careful design of the written input to such activities. These guidelines include predicting what parts of the written input are most likely to be used in the task, using retelling, role-play or problem-solving discussion which draws heavily on the written input, and encouraging creative use of the vocabulary through having to reshape the written input to a particular purpose.

There are no studies of the learning of particular vocabulary through writing, but written input to the writing task could play a role similar to that which it can play in speaking tasks.

Deliberate Vocabulary Learning

Studies comparing incidental vocabulary learning with direct vocabulary learning characteristically show that direct learning is more effective. This is not surprising as noticing and giving attention to language learning generally makes that learning more effective (Schmidt, 1995). Also, deliberate learning is more focused and goal-directed than incidental learning. There is a long history of research on deliberate vocabulary learning, which has resulted in a very useful set of learning guidelines (Nation, 2001). These guidelines are illustrated below through the use of word cards.

1. *Retrieve rather than recognize.* Write the word to be learned on one side of a small card and its translation on the other side. This forces retrieval of the item after the first meeting. Each retrieval strengthens the connection between the form of the word and its meaning (Baddeley, 1990). Seeing them both together does not do this.
2. *Use appropriately sized groups of cards.* At first start with small packs of cards – about 15 or 20 words. Difficult items should be learned in small

groups to allow more repetition and more thoughtful processing. As the learning gets easier increase the size of the pack – more than 50 seems to be unmanageable simply for keeping the cards together and getting through them all in one go.

3. *Space the repetitions.* The best spacing is to go through the cards a few minutes after first looking at them, and then an hour or so later, and then the next day, and then a week later, and then a couple of weeks later. This spacing is much more effective than massing the repetitions together into an hour of study. The total time taken may be the same but the result is different. Spaced repetition results in longer lasting learning.
4. *Repeat the words aloud or to yourself.* This ensures that the words have a good chance of going into long-term memory.
5. *Process the words thoughtfully.* For words which are difficult to learn, use depth of processing techniques like the keyword technique (*see below*). Think of the word in language contexts and situational contexts. Break the word into word parts if possible. The more associations you can make with an item, the better it will be remembered.
6. *Avoid interference.* Make sure that words of similar spelling or of related meaning are not together in the same pack of cards. This means days of the week should not be all learned at the same time. The same applies to months of the year, numbers, opposites, words with similar meanings, and words belonging to the same category, such as items of clothing, names of fruit, parts of the body and things in the kitchen. These items interfere with each other and make learning much more difficult (Higa, 1963; Tinkham, 1997, Waring, 1997b; Nation, 2000).
7. *Avoid a serial learning effect.* Keep changing the order of the words in the pack. This will avoid serial learning where the meaning of one word reminds you of the meaning of the next word in the pack.
8. *Use context where this helps.* Write collocates of the words on the card too where this is helpful. This particularly applies to verbs. Some words are most usefully learned in a phrase.

Deliberate vocabulary learning is a very important part of a vocabulary learning programme. It can result in a very quick (and long-lasting) expansion of vocabulary size which then needs to be consolidated and enriched through meaning-focused input and output, and fluency development. The meaning-focused and context-based exposure also complements deliberate learning in that deliberate learning by itself usually does not provide the knowledge of grammar, collocation, associations, reference and constraints on use that may be best learned through meeting items in context.

Deliberate vocabulary teaching is one way of encouraging deliberate vocabulary learning. Such teaching can have three major goals. First, it can aim to result in well-established vocabulary learning. This requires what has been called 'rich instruction' (Beck, McKeown and Omanson, 1987: 149). This involves spending a reasonable amount of time on each word and focusing on several aspects of what is involved in knowing a word such as its spelling, pronunciation, word parts, meaning, collocations, grammatical patterns and contexts of use. Such rich instruction is necessary if pre-teaching of vocabulary is intended to have the effect of improving comprehension of a following text (Stahl and Fairbanks, 1986). Because of the time involved in rich instruction, it should be directed towards high-frequency words. Second, deliberate vocabulary teaching can have the aim of simply raising learners' consciousness of particular words so that they are noticed when they are met again. Here, vocabulary teaching has the modest aim of beginning the process of cumulative learning. Third, deliberate vocabulary teaching can have the aim of helping learners gain knowledge of strategies and of systematic features of the language that will be of use in learning a large number of words. These features include sound-spelling correspondences (Wijk, 1966; Venezky, 1970; Brown and Ellis, 1994), word parts, (prefixes, stems and suffixes), underlying concepts and meaning extensions, collocational patterns and types of associations (Miller and Fellbaum, 1991).

Deliberate vocabulary teaching can take a variety of forms including:

- Pre-teaching of vocabulary before a language use activity.
- Exercises that follow a listening or reading text, such as matching words and definitions, and creating word families using word parts or semantic mapping.
- Self-contained vocabulary activities like the second-hand cloze (Laufer and Osimo, 1991).
- Word detectives where learners report on words they have found.
- Collocation activities.
- Quickly dealing with words as they occur in a lesson.

Developing Fluency with Vocabulary across the Four Skills

Knowing vocabulary is important, but to use vocabulary well it needs to be available for fluent use. Developing fluency involves learning to make the best use of what is already known. Thus, fluency development activities should not involve unknown vocabulary. The conditions needed for fluency development involve a large quantity of familiar material, focus on the message and some pressure to perform at a higher-than-normal level. Because of these conditions, fluency development activities do not usually focus specifically on vocabulary or grammar but aim at fluency in listening, speaking, reading or writing.

There are two general approaches to fluency development. The first relies primarily on repetition and could be called 'the well-beaten path approach' to fluency. This involves gaining repeated practice on the same material so that it can be performed fluently. This includes activities such as repeated reading, the 4/3/2 technique (where learners speak for four minutes, then three minutes, then two minutes on the same topic to different learners), the best recording (where the learner makes repeated attempts to record their best-spoken version of a text) and rehearsed talks. The second approach to fluency relies on making many connections and associations with a known item. Rather than following one well-beaten path, the learner can choose from many paths. This could be called 'the richness approach' to fluency. This involves using the known item in a wide variety of contexts and situations. This includes speed-reading practice, easy extensive reading, continuous writing and retelling activities. The aim and result of these approaches is to develop a well-ordered system of vocabulary. Fluency can then occur because the learner is in control of the system of the language and can use a variety of efficient, well-connected and well-practised paths to the wanted item. This is one of the major goals of language learning and is not easily achieved.

This discussion has focused on the learning of individual words, but learning MWUs can occur across the four learning strands as well. Most learning of such units should occur through extensive meaning-focused language use rather than deliberate study. Fluency development activities provide useful conditions for establishing knowledge of these units.

Strategy Development

There are four major strategies that help with finding the meaning of unknown words and making the words stay in memory (*see* Chapter 10, *Focus on the Language Learner*, for more on strategies). These strategies are guessing from context clues, deliberately studying words on word cards, using word parts and dictionary use. These are all powerful strategies and are widely applicable. Because they provide access to large numbers of words, they deserve substantial amounts of classroom time. Learners need to reach such a level of skill in the use of these strategies that it seems easier to use them than not use them. These strategies are useful for the high-frequency words of the language and they are essential for the low-frequency words. Because there are thousands of low-frequency words, and each word occurs so infrequently, teachers should not spend classroom time teaching them. Instead, teachers should provide training in the strategies so that learners can deal with these words independently.

Guessing from Context

Guessing a meaning for a word from context clues is the most useful of all the strategies. To learn the strategy and to use it effectively, learners need to know 95–98% of the tokens in a text. That is, the unknown word to be guessed has to have plenty of comprehensible supporting context. The results of using the guessing strategy have to be seen from the perspective that learning any particular word is a cumulative process. Some contexts do not provide a lot of information about a word, but most contexts provide some information that can take knowledge of the word forward. Nagy, Herman and Anderson (1985) estimated that native speakers gain measurable information for up to ten per cent of the unknown words in a text after reading it. Although this figure may seem low, if it is looked at over a year of substantial amounts of reading, the gains from such guessing could be 1000 or more words per year. For second language learners, learning from guessing is part of the meaning-focused input strand and this should be complemented by direct learning of the same words, and for the higher frequency words, opportunity to use them in meaning-focused output.

Training in the skill of guessing results in improved guessing (Fukkink and de Glopper, 1998; Kuhn and Stahl, 1998). Such training should focus on linguistic clues in the immediate context of the unknown word, clues from the wider context, including conjunction relationships, and common-sense and background knowledge. Word part analysis is not a reliable means of guessing, but it is a very useful way of checking on the accuracy of a guess based on context clues.

Successful guessing from context is also dependent on good listening and reading skills. Training learners in guessing from context needs to be a part of the general development of these skills. Training in guessing needs to be worked on over several weeks until learners can make largely successful guesses with little interruption to the reading process.

Learning from Word Cards and Using Word Parts

The strategy of learning vocabulary from small cards made by the learners has already been described in the section on the deliberate study of words. Although such rote learning is usually frowned on by teachers, the research evidence supporting its use is substantial (Nation, 2001). There are also very useful mnemonic strategies that can increase the effectiveness of such learning. The most well-researched of these is the 'keyword technique' which typically gives results about 25 per cent higher than ordinary rote learning. The keyword technique is used to help link the form of a word to its meaning and so can be brought into play once the learner has access to the meaning of the word. To explain the technique let us take the example of a Thai learner of English wanting to learn the English word *fun*. In the first step, the learner thinks of a first language word that sounds like the foreign word to be learned. This is the keyword. Thai has a word *fun* which means 'teeth'. In the second step, the meaning of the keyword is combined in an image with the meaning of the foreign word. So, for example, the learner has to think of the meaning of the English word *fun* (happiness, enjoyment) combining with the Thai keyword *fun* (teeth). The image might be a big smile showing teeth, or a tooth experiencing a lot of enjoyment.

Using word parts to help remember the meaning of a word is somewhat similar. If the learner meets the word *apposition* meaning 'occurring alongside each other', the learner needs to find familiar parts in the word, *ap-* (which is a form of *ad-* meaning 'to' or 'next to'), *pos* (meaning 'to put or to place'), and *-ition* (signalling a noun). The word parts are like keywords, and the analysis of the word into parts is like the first step of the keyword technique. The second step is to relate the meaning of the parts to the meaning of the whole word which is a simple procedure for apposition. This is done by restating the meaning of the word including the meaning of the parts in the definition –

'placed next to each other'. To make use of word parts in this way the learner needs to know the most useful word parts of English (20 or so high-frequency prefixes and suffixes are enough initially), needs to be able to recognize them in their various forms when they occur in words and needs to be able to relate the meanings of the parts to the meaning of the definition. Like all the strategies, this requires learning and practice. Because 60 per cent of the low-frequency words of English are from French, Latin or Greek and thus are likely to have word parts, this is a widely applicable strategy.

Dictionary Use

Dictionaries may be monolingual (all in the foreign language), bilingual (foreign language words–first language definitions and vice versa) or bilingualized (monolingual with first language definitions also provided). Learners show strong preferences for bilingual dictionaries and research indicates that bilingualized dictionaries are effective in that they cater for the range of preferences and styles (Laufer and Hadar, 1997; Laufer and Kimmel, 1997).

Dictionaries may be used 'receptively', to support reading and listening, or 'productively', to support writing and speaking. Studies of dictionary use indicate that many learners do not use dictionaries as effectively as they could, and so training in the strategies of dictionary use could have benefits. Dictionary use involves numerous subskills such as reading a phonemic transcription, interpreting grammatical information, generalizing from example sentences and guessing from context to help choose from alternative meanings.

Training learners in vocabulary use strategies requires assessment to see what skill and knowledge of the strategies the learners already have, planning a programme of work to develop fluent use of the strategy, helping learners value the strategy and be aware of its range of applications, and monitoring and assessing to measure progress in controlling the strategy. Each of the strategies described above are powerful strategies that can be used with thousands of words. They each deserve sustained attention from both teachers and learners.

Assessing Vocabulary Knowledge

Vocabulary tests can have a range of purposes:

- To measure vocabulary size (useful for placement purposes or as one element of a proficiency measure).
- To measure what has just been learned (a short-term achievement measure).
- To measure what has been learned in a course (a long-term achievement measure).
- To diagnose areas of strength and weakness (a diagnostic measure).

Although no standardized vocabulary test has been truly well-researched, the following four have some research evidence supporting their validity (*see* Chapter 15, *Assessment*). They include the Vocabulary Levels Test (Schmitt, 2000; Nation, 2001; Schmitt, Schmitt and Clapham, 2001), the Productive Levels Test (Laufer and Nation, 1999), the *Eurocentres Vocabulary Size Test 10KA* (EVST) (Meara and Jones, 1990) and the vocabulary dictation tests (Fountain and Nation, 2000). Each of these tests samples from a range of frequency levels and tests learners' knowledge of the words. The Vocabulary Levels Test uses a matching format where examinees write the number of their answer in the blanks.

1	business		
2	clock	______	part of a house
3	horse	______	animal with four legs
4	pencil	______	something used for writing
5	shoe		
6	wall		

The test has five sections, covering various frequency levels, and so the results can help teachers decide what vocabulary level learners should be working on. Because teachers should deal with high-frequency and low-frequency words in different ways, the results of this test can also help teachers decide what vocabulary work they should be doing with particular learners or groups of learners.

The Productive Levels Test requires learners to recall the form of words using a sentence cue.

They keep their valuables in a va______ at the bank.

The first few letters of each tested word are provided to help cue the word and to prevent the learners from writing other synonymous words. This test format is useful in showing whether a learner's knowledge of a word has begun to move towards productive mastery.

The EVST uses a yes/no format where learners see a word on a computer screen and then have to decide if they could provide a meaning for the word. The test includes some imitation words that look like real words ('ploat') and learners' scores are adjusted downwards by the number of times they say that they know these non-words. The test gives an estimate of vocabulary size which can help inform placement decisions.

The vocabulary dictation tests each consist of five paragraphs, with each successive paragraph containing less-frequent vocabulary. The test is administered like a dictation but only the 20 target words at each level are actually marked. There are four versions of the test. It can be used for determining the extent of learners' listening vocabulary quickly.

As can be seen in the above examples, there is a wide variety of vocabulary test formats. Different test formats testing the same vocabulary tend to correlate with each other around 0.7, indicating that test format plays a considerable role in determining the results of a vocabulary test. This also suggests that different test formats may be tapping different aspects of vocabulary knowledge. There are a number of issues that complicate vocabulary testing and these are well covered by Read (2000) in his book devoted to assessing vocabulary.

Limitations on Generalizing Vocabulary Size Estimates and Strategies to Other Languages

It is worth pointing out that most of the research on vocabulary has been done within the broad context of English Language Teaching (ELT). This is rather unfortunate, since English is a very peculiar language in some respects, and particularly so as far as its vocabulary is concerned. This means that the findings reported in the earlier part of this chapter may not always be generalizable to other languages in a straightforward way.

The chief characteristic of English vocabulary is that it is very large. Consider, for example, the set of objects and actions that in English are labelled as: *book*, *write*, *read*, *desk*, *letter*, *secretary* and *scribe*. These words are all related semantically in that they refer to written language, but it is impossible to tell this simply by looking at the words. They share no physical similarities at all, and this means that learners of English have to acquire seven separate words to cover all these meanings. In other languages, this is not always the case. In Arabic, for example, all seven meanings are represented by words which contain a shared set of three consonants – in this case k–t–b. The different meanings are signalled in a systematic way by different combinations of vowels. This means that in Arabic all seven English words are clearly marked as belonging to the same semantic set, and the learning load is correspondingly reduced.

There are also some historical reasons which contributed to the complexity of English vocabulary. A substantial proportion of English vocabulary is basically Anglo-Saxon in origin, but after the Norman invasion in 1066, huge numbers of Norman French words found their way into English, and these words often co-existed side-by-side with already existing native English words. English vocabulary was again very heavily influenced in the eighteenth century when scholars deliberately expanded the vocabulary by introducing words based on Latin and Greek. This means that English vocabulary is made up of layers of words, which are heavily marked from the stylistic point of view. Some examples of this are:

cow	beef	bovine
horse	–	equine
pig	pork	porcine
sheep	mutton	ovine

The first column (Anglo-Saxon words), describes animals in the field, the second column (Norman French derivatives) describe the animals as you might find them in a feast, whereas the third column (learned words) describes the animals as you might find them in an anatomy text book. It is very easy to find examples of the same process operating in other lexical fields as well, since it is very widespread in English. Almost all the basic Anglo-Saxon words have parallel forms based on Latin or Greek, which are used in particular, specialist discourse.

English also has a tendency to use rare and unusual words where other languages often use circumlocutions based on simpler items. Thus, English uses *plagiarism* to describe stealing quotations from other people's literary works, *rustling* to describe stealing other people's cows and *hijacking* to describe stealing other people's airplanes. These terms are completely opaque in English: the words themselves contain no clues as to their meaning. In other languages, these ideas would often be described by words or expressions that literally translate as *stealing writing* or *stealing cows* or *stealing aircraft*. In these languages, the meaning of these expressions is entirely transparent, and they could easily be understood by people who knew the easy words of which these expressions are composed.

The Lexical Bar

Unfortunately for EFL learners, the opaque terms are not just an optional extra. A large part of English education is about learning this difficult vocabulary, which Corson (1995) called the 'lexical bar' or barrier, and educated English speakers are expected to know these words and to be able to use them appropriately. Trainee doctors, for example, need to master a set of familiar words for body parts (*eye*, *ear*, *back*, etc.) as well as a set of formal learned words for the same body parts (*ocular*, *auricular*, *lumbar*, etc.) They may also need to acquire a set of familiar words which refer to body parts which are regarded as taboo (*stomach/belly*, *bum*, *arse*, *bottom*, etc.). Some of these words will only occur in speech with patients, some would only be appropriately used with children, others will only appear in written reports, others might be appropriately used in a conversation with a medical colleague. Using a word in the wrong context can cause offence, make you look like an idiot or cause you to be completely misunderstood. All this represents a significant learning burden for non-native speakers, and one which is not always found to quite the same extent in other languages.

The basic problem here seems to be that English vocabulary consists of a large number of different 'items', which are layered according to the contexts in which they appear. In other languages, the number of basic items is smaller, but there is more of a 'system' for inventing new words (Ringbom, 1983). In languages with a rich morphology, for example, it is often possible to make a verb out of any noun by adding the appropriate verbal ending, or to make an adjective by adding an appropriate adjectival ending. You cannot always do this easily in English. In some other languages – German is a good example – it is possible to create new words by combining simple words into novel, compound forms. Native speakers learn these systems, and develop the ability to create new words as they need them, and to easily decode new words created by other speakers when they hear them. In these languages, having a large vocabulary may be less important than having an understanding of the process of word formation and having the ability to use these processes effectively and efficiently as the need arises.

An important consequence of this is that some of the statistical claims put forward for English will not apply straightforwardly to other languages. In English, for example, we would normally consider a vocabulary of 4000–5000 word families to be a minimum for intermediate level performance. But this may not be the case for other languages. It is possible, for example that in a language which makes extensive use of compounding, and has a highly developed morphological system, a vocabulary of 2000–3000 words might give you access to a very much larger vocabulary which could be constructed and decoded online. It is difficult to assess this idea in the absence of formal statistical evaluations, but it clearly implies that we need to evaluate the claims we make about English in the light of the particular lexical properties of other target languages.

Vocabulary Size and Language Proficiency

This means that the relationship between vocabulary size and overall linguistic ability may differ from one language to another. In English, there is a relatively close relationship between how many words you know, as measured on the standard vocabulary tests, and how well you perform on reading tests, listening tests and other formal tests of your English ability. In other languages, it is much less clear that this relationship holds up in a straightforward way. Let us imagine, for example, a language which had a relatively small core vocabulary – let's call it 'Simplish' – and let us say that Simplish has a core vocabulary of about 2000 core words but makes up for this by making very extensive use of compounding. In Simplish, anyone who had acquired the basic vocabulary and understood the rules of compounding would automatically have access to all the other words in the vocabulary as well. 'Difficult words' – in the sense of words that are infrequent – would exist in Simplish, but they would not be a problem for learners. These infrequent words would probably be long because they were made up of many components, but the components would all be familiar at some level. It might be difficult to unwrap the words at first, but in principle, even the most difficult word would be amenable to analysis. For L2 learners of Simplish, the vocabulary learning load would be tiny, and once they had mastered the core items, they would face few of the problems that L2 English speakers face. They would be able to read almost everything they encountered; they would be able to construct new vocabulary as it was needed, rather than learning it by rote in advance. For teachers of Simplish, it would be important to know how much of the core vocabulary their students could handle with ease and familiarity, but beyond that, the notion of 'vocabulary size' would be completely irrelevant. It would be useful to know whether your class had a vocabulary of 500 words or 1500 words, but once the learners had mastered the 2000 core words, it just would not make sense to ask how big their vocabulary was. It would also not make much sense to ask what words we need to teach: the obvious strategy would be to get students familiar with all the core vocabulary as quickly as possible. After that, we would need to concentrate on teaching learners how to unpack unfamiliar vocabulary and how to construct compound words in a way that was pleasing, elegant and effective.

Unfortunately, not many languages are as elegant as Simplish. However, if we think of English as being especially difficult as far as vocabulary is concerned then it seems likely that many of the languages that we commonly teach are much more like Simplish than English. This means that we would not always expect to find that vocabulary plays the same role in learning these languages as it does in English. Vocabulary size in English strongly limits the sorts of texts that you can read with ease: this might not be case in other languages, and this would make it unnecessary for teachers to invest in simplified readers. Advanced learners of English tend to exhibit richer vocabulary in their writing than less advanced learners do: in a language that makes more extensive use of a core vocabulary this relationship might not be so obvious, and this might have implications for the ways examiners evaluate texts written by learners of these languages. English has very different vocabulary registers for special areas of discourse and this makes it important for learners to acquire academic vocabulary, legal vocabulary, the vocabulary of business English and so on: in other languages, these special vocabularies may not be so obvious or necessary.

The general point here is that the sheer size of English vocabulary has a very marked effect on the way we teach English, and severely constrains the level of achievement we expect of learners. Most people agree that fluent English speakers need very large vocabularies, that it makes sense to pace the learning of this vocabulary over a long time and that we should rely principally on the learners' own motivation to get them to these very high levels of vocabulary knowledge. However, this would not be the best set of strategies to adopt if you believed that the language you were teaching was more like Simplish. In these cases, it would be worth putting a lot of effort into getting students learn the core vocabulary very quickly indeed, simply because the pay-off for this effort would be very great.

Our guess is that very many languages are much simpler than English is as far as their vocabulary structure is concerned, and that it would be wrong to assume that research findings based on English will generalize automatically to these languages. This means that teaching methods that take English vocabulary structure for granted will not always be the best way for us to approach the teaching of vocabulary in other languages.

This comparison underlines the importance of having a well-thought-out plan for helping learners with English vocabulary. The basis for this plan is an awareness of the distinction between high-frequency and low-frequency words, and of the strands and strategies which are the means of dealing with these words.

Further Reading

- Nagy, W.E., Herman, P., Anderson, R.C. (1985) Learning words from context. *Reading Research Quarterly* 20: 233–253. The classic first language study of guessing from context.
- Nation, I.S.P. (2001) *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press. A substantial recent survey of vocabulary teaching and learning.
- Read, J. (2000) *Assessing Vocabulary*. Cambridge: Cambridge University Press. A clear, well-informed study of vocabulary testing.
- Schmitt, N. (2000) *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press. An accessible introduction to vocabulary teaching and learning.
- Schmitt, N., McCarthy, M. (eds) (1997) *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press. An authoritative and accessible collection of articles on vocabulary.
- West, M. (1953) *A General Service List of English Words*. London: Longman. The classic second language 2000 word list. A model for future lists.

Hands-on Activity

Take this test (Goulden, Nation and Read, 1990) to estimate how many word families you know. You will find below a list of 50 words which is part of a sample of all the words in the language. The words are arranged more or less in order of frequency, starting with common words and going down to some very unusual ones.

Procedure

1. Read through the whole list. Put a tick next to each word you know, that is, you have seen the word before and can express at least one meaning for it. Put a question mark next to each word that you think you know but are not sure about. Do not mark the words you do not know.
2. When you have been through the whole list, go back and check the words with question marks to see whether you can change the question mark to a tick.
3. Then find the last five words you ticked (that is, the ones that are furthest down the list). Show you know the meaning of each one by giving a synonym or definition or by using it in a sentence or drawing a diagram, if appropriate.
4. Check your explanations of the five words in a dictionary. If more than one of the explanations is not correct, you need to work back through the list, beginning with the sixth to last word you ticked. Write the meaning of this word and check it in the dictionary. Continue this process until you have a sequence of four words (which may include some of the original five you checked) that you have explained correctly.
5. Calculate your score by multiplying the total number of known words by 500. Do not include the words with a question mark in your scoring.

Test

1 bag
2 face
3 entire
4 approve
5 tap
6 jersey
7 cavalry
8 mortgage
9 homage
10 colleague
11 avalanche
12 firmament
13 shrew
14 atrophy
15 broach
16 con
17 halloo
18 marquise
19 stationery
20 woodsman
21 bastinado
22 countermarch
23 furbish
24 meerschaum
25 patroon
26 regatta
27 asphyxiate
28 curricle
29 weta
30 bioenvironmental
31 detente
32 draconic
33 glaucoma
34 morph
35 permutate
36 thingamabob
37 piss
38 brazenfaced
39 loquat
40 anthelmintic
41 gamp
42 paraprotein
43 heterophyllous
44 squirearch
45 resorb
46 goldenhair
47 axbreaker
48 masonite
49 hematoid
50 polybrid

An estimate of vocabulary size is most informative if we know how much vocabulary it takes to use English in various ways. How much vocabulary do you think it takes to accomplish the following things?

- Engage in everyday conversations with your friends ________ word families
- Begin to move from graded readers to authentic texts ________ word families

- Read common authentic texts (newspapers, magazines, novels) without unknown words being a problem _________ word families
- Engage in sophisticated language use, such as studying at an English-medium university _________ word families