Vocabulary in ESP

Last updated 6 November 2009

1. Counting words

a. **Tokens**, ie running words, as in word count. Used for words in a book, typing speed, reading speed etc ***don't*** counts as one word. Tip. If you want to see this for a Microsoft Word document, go to File, Properties, Statistics and then you can see the word count.

b. **Types** ie different words, each word form eg a singular and plural being counted separately.

c. **Lemmas** ie a headword and some of its inflected and reduced forms (n't). Usually, this term means that only the same parts of speech are included, but variant spellings are included. There is a problem with irregular forms such as nice, is, brought, beaten, best.

d. **Word families**. These consist of a headword, its inflected forms, and its closely related derived forms eg adding various prefixes and suffixes.

2. How much vocabulary do you need, to use another language?

The following division of vocabulary is widely used.

a. **High frequency words**. These belong to about 2000 word families.

b. **Academic words**. These are those used which are common in the academic world, irrespective of the subject. eg policy, phase, adjusted, sustained.

c. **Technical words**. These are closely related to the topic and subject area of the text. They differ between the subjects. Each subject typically has 1000 word families.

d. **Low frequency words**. eg zoned, pioneering, perpetuity, pastoral. They make up about 5% words in an academic text. They include all non-academic, non-technical and not high-frequency words.

3. **Example: An economics textbook of 295,294 words.**
   82.5% of the words were from the first 2000 word familes, and 1577 of these word families were used.
   8.7% of the words were from an Academic list,
   8.8% of the words came from a huge variety of 3,225 rarer word families.

4. **Example: The Coxhead corpus of 3.5m words, balancing science, arts, commerce and law text.**
   10 word forms/families accounted for 23.7% of the corpus
   100 word forms/families accounted for 49% of the corpus
   1000 word families accounted for 74.1% of the corpus
   2000 word families accounted for 81.3% of the corpus
   3000 word families accounted for 85.2% of the corpus
   4000 word families accounted for 87.6% of the corpus
   5000 word families accounted for 89.4% of the corpus
   12448 word families accounted for 95.0% of the corpus
   86741 word families accounted for 100% of the corpus

5. **High frequency words**

   a. **How large is this group?** ie where do you draw the line? Usually it is drawn at 2000 word families.

   b. **How stable are high frequency words?** ie how much do frequency lists differ depending on the corpus (database)? This is a question of validity - how generalisable are the results from one large corpus to the expected results in another large corpus? We must allow for minor disagreements in the actual frequencies. The agreement is that 2000 words families account for around 80% of all words in a text, irrespective of the genre. And this rough figure has been widely substantiated.

6. **Less frequent words**

   a. Notice how in the above economics corpus, an extra 3000 words ie 2000 to 5000 words, accounted for only (89.4-81.3 =) 9.1% of the extra words. Also, rare words (the 81741 word families beyond the 5000 word families level) were 10.6% (364,000) of the word tokens.

   b. There is a problem of drawing the line between the different groups. For instance, any one of several thousand moderate/low frequency words could be considered to be candidates for inclusion in the high frequency list. A different corpus (database) would give different suggestions. eg curious, wing, to arm, gate, approximately.

   c. Many low frequency words are proper names. About 4% in the ‘Brown Corpus’.

d. “One person's technical vocabulary is another person's low frequency word”. The vocabulary of an individual grows partly as a result of our personal interests.

e. Some low frequency words are genuinely so. They are words which almost every language user rarely uses. eg eponymous, bifurcate, plummet, ploy.

f. Webster's third New International Dictionary has 267000 entries, 113,161 word families. To read with minimal disturbance from unknown vocabulary, Nation thinks 15,000 word families at least need to be known.

7. **Specialised vocabulary**

It is possible to make a specialised vocabulary list of most frequent technical words in that subject. Either do a frequency count, or use a group of experts, or both. Notice how the academic list is very important in the economics textbook but the extra words in 2000-3000 word level do not add much.

Nation (1990:18) says that little research has been done on the size of technical vocabularies per subject. He estimates 1000 words (families). This immediately raises some questions, including

a. defining the bounds of a subject. If it is small enough Nation’s estimate may be true!
b. --some subjects are more vocabulary intense than others. eg anatomy, pharmacy, physiology, biochemistry, therefore are likely to be considerably more
c. Nation says nothing about semi-technical vocabulary, eg force, cell. These are the common words which have a specialised sense.
d. in all this no account is taken of polysemy (two or more distinct meanings).

8. Missing concepts: semi-technical words and faux amis

The classification scheme above is very useful, but two significant classes of words are overlooked.

a. The semi-technical words are overlooked because the lists take no account of polysemy. It should be obvious why: it is not yet practical to automate the differentiation of words according to several senses of a word, and this means it is not easy to count them in a large database. So this is a real, and understandable, limitation of the use of these lists of words.

Semi-technical words, are, by definition, polysemous. Since researchers have known for some time that the semi-technical words give notable problems to first and second language learners, it would make a lot of sense to identify them and systematically teach them. To my mind this should be a high priority for research.

What is less excusable is that polysemy has been little researched. I do not know, for each word family in the frequent list, how many major senses there are. Maybe it exists, but I could not find it after over an hour of searching. Of course, the situation is made even more complicated in that derivations may themselves be polysemous.

While linguists do distinguish between polysemy and homonymy, the question is rather academic and artificial to the user. Polysemy refers to several distinct senses to one word. Homonymy refers to two distinct words which just happen to have the same spelling. Either way, to the user, the effect is the same.

b. Faux amis

Faux amis, the false friends, the words which have similar spelling but different meanings, or worse, have senses which are similar and senses which are different, have been well researched, and lists and books of them - at least for major language pairs such as French and English - are widely available.

What I would like to know as a teacher of English in a Francophobe country, is which words are faux amis in the High Frequency Word List, and in the Academic Word List. It is these faux amis which should be given priority for teaching.

9. Inflections

Reminder: Inflexional affixes, in English are:

a) the noun plural s
b) the indicator of the past tens-ed,
c) -ing (present participle)
d) the verbal s (third person singular)
e) the ‘s of possession (genitive case, Indirect Object)
f) the comparative suffix -er
g) the superlative suffix-est

As an estimate, 21.9% of different types in a written text are inflected. Therefore it should be obvious that these inflections need to be mastered by the learner. It is hard to imagine any course of English which does not cover these basic affixes.

9. Derivational Affixes

They often change the part of speech of the word they are added to. www.southampton.liunet.edu/academic/pau/course/webpre.htm provides a good annotated list of the more common prefixes and www.southampton.liunet.edu/academic/pau/course/websuf.htm has a list of the more common suffixes. See also www.almacen-eoicartagena.com/miguelangel/08-09%20avanzado%202/Prefixes%20&%20Suffixes%20(CALD%203rd%20Ed).pdf which has a scanned list from the Cambridge Advanced Learners Dictionary, third edition, 2008.

It is quite common for over 10% of a text to be made up of words with a prefix or a suffix. In addition, just like lists of most frequent words, it is possible to make lists of the most common affixes, as in the links above.

Therefore it makes sense to teach the common affixes.

10. **Do learners see words as being made of parts?**

Eg do we store and retrieve ***government*** as a single form, or ***govern*** + ***ment*** ? There is evidence that, at least for low frequency, regular, semantically consistent (transparent) suffixed words, they are recomposed every time they are used. For those who have studied psycholinguistics, this is similar to the phonological route.

Nagy et al (1989) investigated whether the speed at which a word is recognised depends on a. the frequency of the word form alone, or does familiarity with the cognates also help. In other words, is it the combined frequencies of the members of the word family which helps word recognition rather than each word separately. To take an example, is the speed of recognising ***argue*** dependant on its frequency alone, or also on the frequency of ***argues, arguing, argument*** etc. The answer will significantly affect preferred language teaching methods.

In fact the speed of recognition depended on the frequency of the word family. This suggests that "morphological relations between words are represented in the [mental] lexicon. Evidence suggests that inflexional and derived relationships significantly affected speed of recognition, suggesting that inflected and derived forms are stored under the same entry or are linked to each other in the mental lexicon." Nation (2001:269).

This means that in teaching, making students aware of word families is very important.
Also, if a stem can take many affixes, it is worth developing in teaching,

eg port. -> export/-able/-er/-ation etc
re-/im-/trans-/de-/sup-port.

11. **Dictionaries, and translation. Nation 2001:289-90**

Nation tackles the question of how useful it is to work through translation, either of the whole text, or for bilingual word lists.

a. Bilingual dictionaries are often criticised

1) they encourage the use of translation, which is thought to be counter-productive in the language classroom
2) they encourage the idea that L2 words are equivalent in meaning to L1 words.
3) they provide little information on how words are used.
4) synonyms are not explained or differentiated

b. But:

1) they provide meanings in a very accessible way
2) they can be bidirectional
3) Nation claims that numerous studies have shown that vocabulary learning is more effective using L2-L1 pairs than in L2- L2 definitions. I will come back to this point later.

12. Learning vocabulary as phrases

Commonly, learning single words in a bilingual list is discouraged. One forceful reason is that if students, especially beginners, focus on single words, they will never develop the fluency they need. Unless vocabulary learning keeps in step with grammar, then students will end up with a huge vocabulary, but still not be able to communicate; still not able to put words together into a coherent sentence.

Therefore, the mantra is repeated: never learn single words, always learn a new word in context, always learn a phrase. Therefore, if students really must use bilingual vocabulary lists, at the very least they should write down the new word in the context of a typical phrase.

13. Commentary by Nation

The extract is from http://www.asian-efl-journal.com/interviews-11-2006-paul-nation.php

> The next most useful step in the study and teaching of multi-word units will be the development of clear and reliable definitions of the different types of multi-word units. ... Recent work by Grant and Bauer ((Grant & Bauer, 2004) found that there are just over 100 true or core English idioms where the meaning of the parts do not provide access to the meaning of the whole. The most frequent of these core idioms are as well (as), by and large, so and so, out of hands.
>
> There is a very large group of multi-word units we can call figuratives where with some effort a relationship can be seen between the parts and the whole. These include multi-word units like at the end of my tether, give the green light, and just what the doctor ordered. These multi-word units have both a literal meaning and a figurative

meaning and the connection between these two meanings can be found through teacher explanation, or by the learners applying an interpretation strategy.

The third type of multi-word unit can be called literals. Examples include weak tea, late arrival, naughty boy and the following ideas. Some of these may have word-for-word translations in the learners' first language while others may be unpredictable from the learners' first language.

Each of these three types of multi-word units need different learning approaches
– core idioms need to be learned unanalysed by finding out their meaning from other sources
--figuratives need to be approached by a learning strategy where the literal meaning is related to the figurative meaning
--literals should not need any special learning for receptive purposes, but for productive purposes learners will need to give some special attention to the unit especially where there is not a first language equivalent.

14. Commentary

Word lists, and work spent memorising them, perhaps using a system of flash cards where the word is written on one side and the translation is written on the other side, are not popular with teachers, for reasons given above. Yet Nation and probably others would argue that they can be a highly effective way of learning vocabulary. And even words which are only half learned, and are unavailable for reproduction may well be available in the recognition modes of reading and listening. Word lists should be extremely useful for intermediate/advanced learners who need to make a big push to learn as quickly as possible several hundred or more words. My experience is that the motivated learner, practicing 30-60 minutes per day, can easily learn over a hundred new words and phrases per week.

As knowledge of L2 becomes advanced, the more the learning methods and learning abilities approximate to learning a new field in L1. Many new fields require a vast increase in vocabulary. as reported in www.scientificlanguage.com/adult1la/adultl1acquisition.html when I studied anatomy I was learning over 50 new terms per hour of lecturing, which meant at least 100 per week (reinforced by practical work in the dissection lab). And that was just one subject among many! For homework I used to draw and label, and practice doing this from memory, just as if I were learning a bilingual list of words. Therefore, I agree with Nation that vocabulary lists can be a very effective language learning method.

I also dare to believe that even the beginner can benefit from bilingual word lists. A beginner needs to speedily recognise and use the basic words - the type of skill greatly aided by memorising word lists. A beginner does not yet have the language skills to analyse words. Word analysis skills are not needed for the 2000-3000 frequent word families, since these words need learning to the point of instant recognition. A beginner also needs to be up and working as soon as possible, in order to survive in the country of the new language. Therefore I think it is quite realistic for full time language students to learn at the rate of 100 word families per week.

More detailed word analysis skills belong to the advanced classes.

15. **Importance of phrases**

a. A large amount of language knowledge is collocational knowledge. Ellis 2001 argues that language knowledge and language use can be accounted for by the storage of chunks of language in long-term memory and by experience of how likely particular chunks are to occur with other particular chunks, without the need to refer to underlying rules. Language knowledge and use is based on associations between sequentially observed language items. This viewpoint sees collocational knowledge as the essence of language knowledge. A good learner recognises more phrases, and can use phrases quicker.

b. All fluent and appropriate language requires collocational knowledge. Arguably, the best way to explain how language users produce native-like sentences and use the language fluently is that, in addition to knowing the language rules, they store hundreds of preconstructed clauses in their memory and draw on them in language use. Thus each word is likely to be stored many times, as a word, and as part of a chunk. It seems that memorised clauses are an important marker of fluency.

c. Many words are used in a limited set of collocations and knowing these is part of what is involved in knowing the words. The most frequent ones, ie comparable in frequency to the first 2000 words, deserve attention in class.

Frequent collocations of frequent words also deserve attention.

give up
get off
heavy rain

Collocations are obviously easier to learn if the total meaning is the sum of the parts.

eg ***take medicine***

We need dictionaries of collocations, especially because many of them are unpredictable. We need to know also their frequency so that we can give priority to the most common ones. And this is where Corpus linguistics comes into its own.

High frequency words are memorised and recognised as whole words. Low frequency words such as ***unambiguousness*** are re-created by rules every time we need them. This means that beginners need to be aiming to speedily recognise words and phrases. Beginners habitually demand listening comprehensions that are slow, and reading comprehensions which are small and every word is important. Teachers within cultures where speed reading seems to be rare have then an uphill task of encouraging fast reading even for beginners.

Paradoxically, the advanced student should be able to function with a high level of instant recognition, and to use analysis skills - bottom up strategies - for new, rare words. My experience is that beginners want to analyse, and firmly resist recognition practice at speed - the type of skill greatly aided by memorising word lists. Perhaps a little more explanation of what the teacher is doing and why would help the adult learners.

References:

Ellis R 2001. *The study of second language acquisition*. First edition. Oxford University Press, UK

Nation ISP 1990. *Teaching and learning vocabulary* Heinle & Heinle, Boston USA.

Nation ISP 2001. *Learning vocabulary in another language*. CUP UK.

http://www.oup.com/elt/catalogue/teachersites/oald7/oxford_3000/business_and_finance?cc=gb
Gives a list, from Oxford University Press, of the 250 most common business and finance words.