What's New About Word Frequency?

Philip Scholfield

FREQUENCY PAST AND PRESENT

Following the recent popularisation of the term 'lexical syllabus', one might be forgiven for thinking that the use of word frequency in ESL/EFL is totally new. Far from it. Frequency has long been regarded as one key criterion in deciding what are the more or less important words (or phrases etc.) to include in a language course of a particular level, or what are the basic words of English, appropriate to replace rarer ones in simplified readers. Frequency also guides good dictionary makers in various ways, such as the order in which to present information about words, and is used by some stylisticians to characterise different varieties of English.

What then **is** new about frequency? It is not the use of frequency in itself: rather it is the **quality** of the frequency information that is now becoming available, and exciting **new kinds** of frequency information that can now be obtained.

FREQUENCY OF WORDS

The simplest frequency information is just about how often individual word forms occur. In the past this was based on counts from collections of just a few million words of English, e.g. most recently in book form from the *Brown* and LOB corpora (Hofland and Johansson 1982). But we can now find this out by computer from corpora containing as many as 100 million words (e.g.*the British National Corpus*). This is very important as many words in a language are very infrequent and occur only once or not at all in small corpora.

If one looked just at the figures for writing, one would think that these words occurred with about the same overall frequencies.

Frequencies per million of 'certain' and 'sure' in written and spoken English (Longman Lancaster Corpus).

	Written	**Spoken**
certain	259	292.5
sure	234	426.9

However, we can now see that 'sure' is much more common in spoken English. Of course the overall figures disguise the fact that the high spoken frequency of 'sure' is due partly to its use in the sense of 'yes' rather than 'certain', e.g. Do you have the time? Sure.

However, if you call up figures for *BE SURE* you find that in spoken English it outnumbers *BE CERTAIN* by about 8 to 1; and *MAKE SURE* outnumbers *MAKE CERTAIN* by about 100 to 1 !

This sort of information is already informing dictionaries such as the recent *Longman Language Activator* (Summers 1993), an innovative production aid for intermediate/ advanced learners of English, which has an ear logo for items found to be particularly frequent in speech.

The misleading nature of overall frequencies is further confirmed by the following:

Frequencies per million of certain and sure in social science and fiction texts.

	Social sc.	**Fiction**
certain	358.7	178.5
sure	73.8	353.1

We can see now that the superficially similar frequency of 'certain' and 'sure' in written language is itself really a conflation of quite different frequencies in different written varieties. If examined further this turns out to be due in part to the higher frequency of 'certain' meaning' particular but unspecified' as in 'certain kinds of...' or 'in certain cases' in social science text (see further Biber et al. 1994).

FREQUENCY OF COLLOCATIONS

Teachers and learners' dictionary makers are today very interested in finding out the commonest combinations in which words occur. It is widely felt that native speakers - and learners - may operate in language much more than used to be thought by using 'ready-made' combinations of words, rather than making everything up from individual words via grammatical rules all the time when speaking or writing.

Today one may readily call up from a corpus information about what words occur most often next to a word of interest (or within three words before, etc.) - a simple kind of collocational information. This can be obtained in simple frequency form, or by some more sophisticated calculation such as 'mutual information' which gives extra weight to co-occurrence of words that are in themselves less frequent.

Top words occurring immediately before money in the BNC Spoken English Corpus, with raw frequencies.

Most frequently co-occurring		**Highest 'mutual information' (for words co-occurring min 9 times):**		
word	***freq***	***word***	***freq***	***m.i.***
the	113	pocket	51	244.76
of	779	spending	38	132.1
some	300	redundancy	10	123.08
more	245	earn	14	110.13
your	225	saving	13	99.7
any	210	losing	16	99.67
that	205	raise	23	98.05
much	161	extra	49	67.37
no	140	save	31	57.56
my	111	petrol	9	51.64
for	99	borrow	10	46.53
their	82	spend	27	44.17
enough	80	enough	80	37.05

As can be seen, the simple frequency information records co-occurrence with function words that are very frequent anyway. Much more interesting for the teacher and dictionary maker are usually the mutual information figures, as they highlight something much closer to the linguistic notion of 'collocation'.

FREQUENCY OF WORDS WITH PARTICULAR COMPLEMENTS

One area where one certainly **is** interested in frequency of co-occurrence with function words is the study of complements of vocabulary items. For instance, we know that many verbs can be followed by *that* clauses, but which are the most common? How often is the *that* omitted ?

e.g. I suggested (that) Tracy should do it.

Information like this is harder to extract automatically, because *that* following a verb will not always signal a complement clause:

cf. 'I suggested that yesterday', but new programs are being written all the time to help obtain the frequencies with less sorting by hand.

Some preliminary results on the above questions from the *Longman Lancaster Corpus* are as follows. In English generally, *say* is one of the most frequent verbs to be followed by *that* clauses, with *that* frequently left out. However, the other top verbs are not verbs of 'speaking' but of 'cognition' like *think, know, see, find, believe* and *feel* . In newspaper English *say* is well ahead of the rest, while in conversation *think* is ahead of *say*. This all makes sense when one thinks of the typical newspaper report of what someone said, and the commonness of *I think*... in speech.

In scientific writing, however, none of the above are very common. Instead some generally rarer verbs like *show, ensure* and *assume* are more prominent with that clauses, and that is rarely omitted. For stylistics and ESP materials this is clearly useful information (see further Aijmer et al. 1991, chs 7 and 9).

FREQUENCY OF WORDS IN PARTICULAR MEANINGS AND PHRASES

One does not usually teach or learn a word all at once, but perhaps just one sense first, followed by others later. We also saw above the value of looking at the meanings of *sure* separately when comparing it with *certain*. Thus, for most purposes one is not just interested in the frequency of words as wholes, but of words in particular parts of speech, meanings and fixed phrases separately. Superior information on these matters too is derivable from modern corpora (though once again not totally automatically) and is set to replace that well-known source of such information used for decades by coursewriters, *West's General Service List* (1953, reflected in Hindmarsh 1980).

Frequencies of different uses of lookout in the Longman Lancaster Corpus.

	frequency
BE on the lookout	44
'place where... '	37
'person who... '	25
KEEP a lookout	14
It's a bad/poor lookout for...	4
It's your/their own lookout	3

In this example we see that the most frequent use of the word is actually in a set phrase, though traditionally teachers and dictionaries would tend to explain the meaning of the word on its own first.

FREQUENCY IN LEARNER ENGLISH

It would not do to finish without mentioning word frequency in the English of learners themselves. Relevant corpora such as the *Longman Corpus of Learners' English* are not yet very large (1.86 million words), but nevertheless useful.

It is possible to see where learners are overusing words. For example, *moreover* occurs at a rate of 211.7 times per million in the learners' corpus, but less than one per million in the native speaker spoken corpus.

It is also possible to identify (by hand) and count the percentages of errors made when using particular words. For instance the phrase *kind(s) of* exhibits 30% error in the learners' corpus. Of this about half involves a failure of agreement between *kind* and its preceding determiner (e.g. *these/all kind of...*), and half involves a singular kind phrase preceding a plural noun (e.g. *this kind of programs*).

Facts of this sort can be used to guide the construction of teaching materials and reference books of common errors. They are also used by some learners' dictionary makers to revise their entries by considering if a competent dictionary user consulting their work could still make the error (see further Meara and English 1988).

USEFUL REFERENCES:

Aijmer, Karin & B. Altenberg (eds).
English Corpus Linguistics.
London: Longman, 1991

Biber, Douglas, S. Conrad & R. Reppen.
"Corpus-based Approaches to Language Issues in Applied Linguistics. "
Applied Linguistics 15 (1994): 169-189.

Hindmarsh, R.
Cambridge English Lexicon.
Cambridge: Cambridge University press, 1980.

Hofland, Knud, & Stig Johansson.
Word Frequencies in British and American English.
Bergen: Norwegian Centre for the Humanities, 1982.

Meara, Paul & F. English.
"Lexical Errors and Learners' Dictionaries."
(ERIC DOCUMENT NO. ED 654 321) 1988.

Summers, Della (ed).
Longman Language Activator.
London: Longman, 1993.

West, Michael.
A General Service List of English Words.
London: Longman, 1953.