USE AND USEFULNESS : CORPUS ANALYSIS AND THE TEACHING OF ENGLISH

Graeme Kennedy
Professor of Applied Linguistics
Victoria University of Wellington

Language teaching theories and good professional practice usually reflect facts or opinions about three related matters:

1. The conditions necessary for learning languages
2. The teaching approaches, procedures and techniques which seem to help people learn languages most efficiently
3. What we learn when we learn a language

At various times and places one or more of these three cornerstones of language teaching receive greater emphasis. Over the last twenty years or so, for example, among teachers of English in many parts of the world it has become part of the conventional wisdom to say that a language is best learned through spoken communication, and particularly through interactive, communicative tasks involving the negotiation of meaning. Others have focused less on the process of learning and have set out to teach learners to do certain things with language. Such teaching of the functions of language has included helping learners to show approval or to apologize appropriately or to get by in a restaurant or a post office. Sometimes topics which reflect current concerns such as the hole in the ozone layer or the consequences of drift net fishing make up the content of instruction, in a sense illustrating various topics which a language serves to communicate.

The emphasis on language as communication has sometimes been associated with an apparent lack of interest in what is being learned. Instead of seeing their role as teachers of vocabulary and structure as well as the use of language, some teachers appear to have seen themselves as organizers of opportunities for learning, leaving nature (or chance) to determine what words or grammar are learned. If there is unlimited time for learning, nature can indeed take its course, but in normal circumstances the reason we teach something is so learning can be speeded up. In this sense teaching tries to be more efficient than nature. By analysing large spoken or written samples or corpora of the language with the aid of computers it is now possible to show which language items are most likely to be needed by language learners and which therefore deserve the investment of teaching time.

CORPUS ANALYSIS

Between 1920 and about 1970, it was fairly commonly recognized by English language teachers that knowledge of 3,000 words would enable a reader to recognize about 85% of the words in any text, or that about 100 very frequent words accounted for half of the words in any text. Table 1 summarizes an analysis of a 5,000,000-word corpus undertaken by Carroll and others (1971) of the reading material to which American schoolchildren were exposed. The finding that a few high frequency words accounted for most of the words which occur in texts was generally consistent with research by Thorndike (1921), Palmer (1933), West (1953), Fries (1952) and many others who had earlier

undertaken analyses involving the laborious counting of words in collections of texts totalling millions of words. West's General Service List of English Words (1953) became one of the most well-known reference works used by English language teachers, not only to find out the most frequently used words but also the most frequently used meanings of words with multiple uses. Table 2, for example, provides information from this famous work on the use of the word right in written texts. Such objective analysis, of course, sometimes runs counter to the intuitions of English language teaching textbook writers and teachers themselves.

Table 1 : A word count of children's reading material in a 5,000,000-word corpus (Carroll, Davies & Richman, 1971)

No. of Different Words	% of Total Words in the Corpus
86,741	100.0
43,831	99.0
5,000	89.4
3,000	85.2
2,000	81.3
100	49.0
10	23.7

Table 2 : Frequency of use of major meanings of RIGHT (West, 1953)

right	% of use
(adjective)	
= correct	9
= suitable	8
= morally right	2
(noun)	
= legal right(s)	44
= moral right(s)	3
(adjective)	
= correctly	4
= entirely	4
(adjective, adverb, noun)	
= direction (opposite of left)	14
other meanings	12

Early corpus analysis, needless to say, had to be done without the aid of computers and was directed mainly at English vocabulary. There was some counting of how often various grammatical items were used, however. Tables 3, 4 and 5 summarize some of the findings of early corpus studies of verb form use which were done without the aid of computers and which still have important implications for

English language teaching. Verbs typically make up about 20% of the words used in spoken or written texts in English and learning to use such morphological changes as those associated with tense and voice is often hard for learners of English.

Table 3 : Most Frequent Verbs (adapted from Ota, 1963)

Verbs	% of the 17,166 verbs in Ota's corpus	Verb form use (%)				
		Simple present	Present prog.	Simple past	Past prog.	All other tenses
be	30.7	79.0	0.05	17.0	0	3.95
think	5.0	87.8	1.3	9.7	0.3	1.00
have	4.0	66.0	1.3	23.7	0.1	8.9
know	3.6	88.7	0	8.2	0	3.1
say	2.6	37.4	2.5	50.0	0.7	9.4
want	2.4	81.1	0	17.7	0.2	1.0
go	1.7	32.1	27.5	26.5	2.4	11.5
get	1.5	47.7	9.2	36.6	1.1	5.4
do	1.5	29.3	22.4	23.6	1.9	22.8
come	1.4	33.6	11.2	39.0	3.3	12.9
have to	1.3	74.8	0.4	24.3	0	0.5
see	1.3	65.0	0.9	23.0	0	11.1
make	1.3	36.6	8.8	29.2	0.9	24.5
mean	1.2	82.6	0	15.4	0	2.0
feel	0.9	68.7	3.3	22.7	0.7	4.6
take	0.9	27.5	9.4	38.9	2.0	22.2
	61.3					

Table 3 summarizes part of a study of American English by Ota (1963). This corpus of about 100,000 words was mainly made up of broadcast radio interviews and TV dramas but included some academic writing. Seven verbs (be, think, have, know, say, want, do) accounted for 50% of all the 17,166 verb occurrences in the corpus, with be accounting for over 30%. The 16 verbs in Table 3 accounted for over 60% of all the verbs used in Ota's corpus.

Ota's study also showed just how varied the verb form use can be with different verbs. Only a few of the verbs (go, get, do, come, make, take) are used much in the progressive, and even then rarely as often as the simple present or past. For some verbs (e.g. say, come), the past tense was more frequently used than the present. But perhaps most important of all, for most verbs, the simple present or past was used much more frequently than all the other tenses combined.

The fact that most English tenses have low frequency of use is very important information for teachers of English. Joos (1964) undertook a study of all the verbs used in a book which was in part a transcript of a legal trial. Joos found that of 224 possible verb forms or so-called "tenses" in English, only 79 actually occurred in his corpus, and, of these, only 15 forms accounted for over 90% of all the verb form uses. (See Table 4.) The other 64 verb form types in Joos's corpus, including ones like might (say), could have (asked), can be (done), accounted for the remaining 10%.

In Joos's study, verb form types 1, 2, 4, 5, 13, 14, the simple present and past make up only 70.9% of the verb forms used. Perhaps because it recorded a courtroom trial, Joos's study had more modal auxiliary verbs (e.g. can, may) than are found generally in spoken or written English. Normally, over 80% of all verb forms used in spoken or written English are in the simple present or past. This is illustrated in Table 5, which summarizes the findings of two corpus studies by George (1963) and Ota (1963). George constructed a corpus of about half a million words of British English. His corpus was made up from written English texts including some scripts of plays which might be assumed to be more like spoken English than the other texts. Both George and Ota found that in spoken English (whether transcribed from recordings or from play scripts) the simple present was more frequently used than the simple past. In written English, the simple past was typically more frequent than the present.

Table 4 The Most Frequent Simple and Complex Finite Verb Forms (Joos, 1964)

Types	Examples	Number of occurrences	% of finite forms used
1	I always say no good comes of these cases	2,853	35.5
2	When the doctor went away did he leave	2,143	26.7
3	the defence have decided not to call the doctor	319	4.0
4	Morphia and heroin were commonly used	292	3.6
5	both morphia and heroin are administered to people	249	3.1
6	If there were, I would take them and destroy them	219	2.7
7	the answers sound as colourless as one can make them	208	2.6
8	the period when he was prescribing for her	177	2.2
9	are you standing there and saying as a trained nurse	175	2.2
10	had you made any inquiries before giving evidence	164	2.0
11	I will certainly help you	115	1.4
12	asks if he may put a further question to the witness	90	1.1
13	And you still say so? I do	83	1.0
14	did the doctor ask you for anything? - He did	80	1.0
15	would you have expected the doses to have a fatal result	77	1.0
			90.1

Overall, because George's corpus was mainly from written sources, the relative frequency of present and past is different from that in Ota's corpus. However, as Table 5 shows, in both studies all other

tenses, including perfect and progressive, are shown as being comparatively infrequent in spoken and written English.

Table 5 Uses of Finite Verb Forms in Two Corpora (percentages)

	George (1963)		Ota (1963)			
Verb forms	Plays	Overall (mainly written)	Radio interviews	TV plays	Academic writing	Overall (mainly spoken)
Simple present	67.6	38.4	65.1	63.6	26.4	59.0
Simple past	14.4	48.2	13.5	23.1	58.5	23.6
Present perfect	5.3	3.1	7.2	2.5	2.7	4.8
Past perfect	0.9	4.1	0.6	0.2	3.4	0.9
Present prog.	4.4	1.4	4.4	6.4	0.9	4.7
Past prog.	0.4	1.4	0.8	1.1	1.1	0.9
Pres.perf.prog.	0.6	0.1	0.5	0.5	0.1	0.4
Past perf.prog.	-	0.1	0.02	-	0.2	0.03
All others	4.4	3.2	7.9	2.6	6.6	5.7

Note: George includes active and passive together. Ota has active only.

COMPUTER CORPUS ANALYSIS

As we have seen, corpus linguistics is concerned both with the description of a language (i.e. what words and structures are possible) as well as with how frequently these words and structures are used in particular contexts (i.e. what words or structures are likely to occur).

The availability of computers has given a new impetus to corpus analysis because it has revolutionized the speed with which items can be counted in texts. Computer corpus linguistics has become one of the most vigorous branches of language study in the last few years and promises to contribute significantly to curriculum development and classroom practice.

Several quite large corpora are available for study and are being used by English teachers in different parts of the world on personal computers. The most well known of these are the one-million-word Brown corpus of written US English, the parallel one-million-word Lancaster-Oslo-Bergen (LOB) corpus of written British English, and the half-million-word London-Lund (LLC) corpus of spoken British English. There are also several major projects to make new corpora which can be taken as representative of different varieties of contemporary English. The most important of these is probably the International Corpus of English (ICE) which will contain one million word samples of spoken and written English from each of 15 different parts of the world, including New Zealand. The British National Corpus (BNC) which will have a total of 100 million words of text from spoken and written sources is being compiled by several British publishers and universities. The Cobuild corpus which contains over 20 million words is part of an ongoing project between Birmingham University and

Collins to produce corpus-based dictionaries, grammars and teaching materials. The study of New Zealand English will also be made easier by the availability of the recently completed one-million-word Wellington corpus of written New Zealand English.

As corpora become increasingly available on CD-Rom disks, teachers are finding that personal computers can be used to access huge amounts of text with astonishing speed to retrieve examples of words or phrases in authentic contexts. Such material can, if desired, be organized relatively quickly into teaching materials for learners of English.

However, it is the possibilities opened up by computer corpora in providing statistical information on the use of language which may be of greater use at present. Until the 1960s, as we have seen, counting the occurrence of words or grammatical items in texts was an article of faith for many language teachers and was a foundation for vocabulary-based approaches to language teaching. Analysis with the aid of a computer of large representative samples of spoken or written English can provide almost unlimited opportunities for teachers of English to direct their teaching towards likely areas of significant use. This can be seen in a widely-respected study by Coates (1983) who studied the use of modal verbs such as must and should in the LOB corpus of written British English and the London-Lund corpus of spoken British English. Must and should are typically taught to speakers of other languages as ways of expressing obligation, e.g.

You must stop when you see a red light.
You should get your back X-rayed.

They have other uses, however, including the expressing of the speaker's degree of certainty (the epistemic use), e.g.

You must be our new neighbours.

It's five o'clock. - That means they should be in Sydney by now.

Should can also, of course, be used to express conditions or hypothetical situations, e.g.

If you should hear of anyone who has a Lada for sale, please let me know.

As Table 6 shows, Coates found that must and should are used to express obligation only about half the time. Other uses which are not often taught are frequent enough to be part of English courses.

Table 6 : Percentage of use of the meanings of MUST and SHOULD (based on Coates, 1983)

	Meanings							
	Obligation		Degree of certainty		Hypothesis		Other	
	Spoken	Written	Spoken	Written	Spoken	Written	Spoken	Written
must	53	65	46	31	-	-	1	4
should	42	51	18	12	21	9	19	28

Another well-known problem for learners of English is conditional sentences using if.... Hill (1960) described 324 possible verb form combinations for conditional sentences, ranging from the simple present tense in both clauses (If you heat air, it rises) to combinations such as If you could have been there you mightn't have enjoyed it anyway.

In a computer corpus analysis which compared American and British written English in the Brown (US) and LOB (British) corpora, Wang (1991) found that two-thirds of the possible combinations never occurred at all and 10 constructions accounted for over 75% of all the conditional sentences in the two corpora. There were no significant differences between US and British usage, as Table 7 shows.

Computer corpus analysis can also suggest possible reasons why learning to use English prepositions is normally difficult. Although the differences in meaning between on and in, between and through or to and from may seem easy enough to establish in theory, learners still seem to find them difficult to use. Studying how prepositions are used in a huge corpus by means of a concordance programme can be done quickly and with quite surprising findings. As Table 8 shows, between and through, for example, differ not so much in their meaning but in the company they keep. Between is most commonly preceded by nouns (with difference being the most common in the one million word LOB corpus). Through, on the other hand, is most commonly preceded by verbs.

Table 7 : Verb Form Use in Conditional Sentences in the Brown and LOB Corpora (Wang, 1991)

Verb form in if-clause	Verb form in main clause	Percentage of conditional occurrences	
		Brown (US)	LOB (UK)
present simple	present simple	22.0	22.0
present simple	will/shall/be going to + stem	13.2	12.5
past simple	would/could/might + stem	11.3	11.1
past simple	past simple	6.7	6.8
present simple	should/must/can/may/ought to + stem	10.0	6.4
past perfect	would/could/might have + past participle	3.9	4.1
were/were to	would/could/might + stem	4.0	4.0
can + stem	present simple	1.1	3.2
present simple	would/could/might + stem	1.9	2.4
present simple	imperative	1.7	2.0
		75.8	74.5

Computer corpus analysis can also throw light on wider semantic matters. In communicative approaches to language teaching, for example, most teachers would agree that learning to talk about temporal frequency (how often something happens) is quite important and therefore a goal for learning and teaching. But how does English express temporal frequency? Most course books include the five words always, usually, often, sometimes, never. In academic English, however, at high school and tertiary level, learners might expect to meet a number of other words or phrases which are also used as often as some of the five words commonly taught. Table 9 summarizes some of the ways in which temporal frequency is expressed, and shows how often each of these items occurs in a 320,000-word corpus of written academic English, suggesting which items might be frequent enough to deserve teaching time.

Table 8 : Words which occur most frequently immediately before BETWEEN and THROUGH in the LOB Corpus (from Kennedy, 1991)

No. of occurrences			No. of occurrences		
* difference	between	59	go	through	36
* relationship		25	pass		33
distinction		19	come		20
* relation		16	be		15
* gap		12	and		13
* agreement		11	get		12
* contrast		11	break		10
* distance		11	*him		10
* place		11	run		10
be		10	*way		9
* comparison		9	*it		8
exist		9	fall		7
* meeting		9	lead		7
* contact		8	look		7
* link		8	out		7
and		7	in		6
in		7	live		6
as		6	only		6
* conflict		6	*them		6
* correlation		6	all		5
* gulf		6	carry		5
lie		6	cut		4
that		6	down		4
* time		6	flash		4
agree		5	*line		4
* connection		5	one		4
distinguish		5	or		4
* interval		5	right		4
pass		5	see		4
* border		4	shoot		4
* exchange		4			

Nouns are recorded in their singular form, verbs in their stem form.

* = noun or pronoun

Table 9 : How temporal frequency is expressed in academic sections of the Brown and LOB Corpora (number of occurrences in 320,000 words) (from Kennedy, 1987a)

Continual frequency		Usual occurence (usually)		High frequency		Low frequency (sometimes)		Zero frequency (never)	
when ...	189	(in) general	124	often	133	sometimes	69	never	66
during	186	normal	119	frequently	36	in/under (certain)		at no time	2
always	110	usually	105	repeated	36	circumstances	29	not ever	2
constant	76	common	104	frequent	12	on some occasions	27	not at all	1
after ...	59	generally	78	repeatedly	9	scatter	22	without ever	1
before ...	35	typical	38	repetition	8	rare	19	under no	
consistent	24	ordinary	30	a great deal	3	unusal	17	circumstances	1
ever	21	usual	27	many times	3	occasionally	14		
as ...	19	normally	26	again and again	2	rarely	14		
continuously	14	commonly	24	time and again	2	in some cases	12		
permanent	13	regular	18	time and again	2	seldom	10		
continuous	12	for the most part 1	0	rife	1	occasional	8		
whenever ...	11	custom	8	thick and fast	1	at times	7		
invariably	10	more often	8	innumerable times	1	periodic	7		
constantly	8	more often than not	8	often times	1	strange	7		
consistently	7	ordinarily	6	session after session	1	from time to time	6		
at all times	5	mostly	6	repetitions	1	odd	6		
wherever ...	4	regularly	6	as much as possible	1	sporadic	5		
permanently	4	customary	5			intermittent	5		
while ...	4	routine	5			irregular	5		
continual	4	in most cases	4			scarcely 5			

Table 10 similarly gives a corpus-based description of how we express the notion of approximation in English and which words or phrases we are most likely to use in educational contexts.

Both Tables 9 and 10 show that counting by computer on its own is not enough. Indeed, useful corpus analysis is rarely done entirely by the computer. It is often necessary to use a combination of the teacher's analytic skills aided by the capacity of the computer to quickly find and count occurrences of items in huge collections of texts.

Table 10 : The most frequent words used for expressing approximation in the academic sections of the Brown and LOB Corpora (number of occurrences in 320.000 words) (from Kennedy, 1987b)

(in) general	189
about	181
average	107
almost	100
estimate	82
relatively	81
generally	78
(50) or (60)	60
approximately	55
a maximum of	54
somewhat	51
nearly	46
at least	42
(50) to (60)	41
near	39
a minimum of	31
within	29
some	27
from (50) to (60)	27
up to	26
between (50) and (60)	24
virtually	24
approximation	21
or more	10
more or less	10
say	10
approximate	10
roughly	10
much the (same)	21
neighbourhood	18
all but	15
generalization	15
approach	14
broad	13
not quite	13
quasi-	11
of the order of	11

Using the capacity of the most recent CD-Rom versions of corpora, it takes only a few seconds on a personal computer to find all the instances of a particular word or phrase in context in a million words of text. For example, if a teacher or learner wanted to know how the word circumstance is used, it would take only a few seconds to discover that circumstance seems to be used mainly in the plural in both the LOB and Brown corpora, as the following word counts show.

	Occurrences in Brown corpus of US English	Occurrences in LOB corpus of British English
circumstance	15	5
circumstances	83	107

This information is probably not very remarkable in itself. What may be of interest to language teachers, however, is information from these same corpora of words which collocate or occur with circumstances.

In the Brown corpus of written US English, circumstances is preceded by under or in on 46 out of the 83 times it occurs (55%), as in under these circumstances, in normal circumstances.

In the LOB corpus, under or in precede circumstances in 53% of the 107 examples. The big difference, however, between US and UK usage seems to be that in US usage 65% of the examples are like under (these) circumstances. In UK usage, on the other hand, only 28% of the occurrences of circumstances are preceded by under. This tendency for circumstances to collocate with in in British English and under in US English is illustrated in Table 11 which summarizes the most frequent combinations of words occurring before circumstances in the two corpora.

Table 11 : Collocations with CIRCUMSTANCES in the Brown and LOB corpora

	Brown (US)	LOB (UK)		Brown (US)	LOB (UK)
in the circumstances	0	6	under the circumstances	7	1
in no circumstances	0	1	under no circumstances	5	1
in these circumstances	2	6	under these circumstances	2	3
in such circumstances	1	4	under such circumstances	0	0
in certain circumstances	0	3	under certain circumstances	2	2
in some circumstances	0	3	under some circumstances	0	0
in different circumstances	0	1	under different circumstances	2	0
in similar circumstances	1	1	under similar circumstances	2	0
in any circumstances	1	1	under any circumstances	0	0
in normal circumstances	1	0	under normal circumstances	1	0

Other words which occur on one or more occasion between in or under and circumstances in one or both of the corpora include adverse, all, appropriate, changing, exceptional, given, identical, modest, other, particular, present, rare, special, specified, unusual, various.

Thus the availability of representative samples of English in computer corpora can help curriculum designers and teachers make better informed judgements about what items to teach and when to teach them, the assumption being that frequent use in corpora may indicate usefulness for learning. Of

course, it is not always the case that the most frequently used words or structures in the language should receive the greatest teaching time or emphasis. Sometimes less frequently used items might deserve more teaching time simply because they are hard to learn or have a wide range of uses. Nevertheless, because teachers can now search huge samples of English in use very quickly to get a picture of what words or phrases occur most frequently and what other items they are most likely to occur with, they are able to pay more attention again to what should be learned as well as to the processes of learning and teaching.

REFERENCES

Carroll, John B., P. Davies and B. Richman. (1971). The American Heritage Word Frequency Book. New York: Houghton Mifflin.

Coates, Jennifer. (1983). The Semantics of the Modal Auxiliaries. Beckenham: Croom Helm.

Fries, C. C. (1952). The Structure of English: An Introduction to the Construction of English Sentences. New York: Harcourt Brace & Co.

George, H. V. (1963). Report on a Verb-Form Frequency Count. Hyderabad: Central Institute of English. Monograph 1.

Hill, L. A. (1960). 'The Sequence of Tenses with "if"-clauses'. Language Learning10,3.165-178.

Joos, Martin. (1964). The English Verb: Form and Meanings. Madison: University of Wisconsin Press.

Kennedy, G. D. (1987a). 'Quantification and the Use of English: A Case Study of One Aspect of the Learner's Task'. Applied Linguistics 8,3.264-286.
—. (1987b). 'Expressing Temporal Frequency in Academic English'. TESOL Quarterly 21,1.69-86.
—. (1991). 'Between and Through: The Company They Keep and the Functions They Serve', in: K. Aijmer and B. Altenberg (eds.) English Corpus Linguistics: Studies in Honour of Jan Svartvik, 95-110. London: Longman.

Ota, Akira. (1963). Tense and Aspect of Present Day American English. Tokyo, Kenkyusha.

Palmer, H. E. (1933). Second Interim Report on English Collocations. Tokyo, Institute for Research in English Teaching.

Thorndike, E. L. (1921). Teacher's Word Book. New York: Columbia Teachers College.

Wang Sheng. (1991). A Corpus Study of English Conditionals. Unpublished MA thesis. Victoria University of Wellington.

West, M. (1953). A General Service List of English Words. London: Longman.