The Science and Technology of Corpus, and Corpus for Science and Technology

Ramesh Krishnamurthy (COBUILD, University of Birmingham and HarperCollins Publishers)

1. Introduction

Corpus Linguistics is a young discipline. The earliest work was done in the 1960s, but corpora only began to be widely used by lexicographers and linguists in the late 1980s, by language teachers in the late 1990s, and by language students only very recently.

This course in corpus linguistics was held at the Departamento de Linguistica Aplicada, E.T.S.I. de Minas, Universidad Politecnica de Madrid from June 15-19 1998. About 45 teachers registered for the course. 30% had PhDs in linguistics, 20% in literature, and the rest were *doctorandi* or qualified English teachers. The course was designed to introduce the use of corpora and other computational resources in teaching and research, with special reference to scientific and technological discourse in English.

Each participant had a computer networked with the lecturer's machine, whose display could be projected onto a large screen. Application programs were loaded onto the central server, and *telnet* and a web browser were available. COBUILD gave us permission to access the 323 million word Bank of English corpus, Mike Scott allowed us to use his Wordsmith Tools software, and Tim Johns gave us a copy of his MicroConcord program.

2. Traditional Resources

2.1 Dictionaries

Dictionaries have long been a standard resource for language teachers and students. Traditionally, dictionaries and other language reference books were written on the basis of intuition (or copied from existing sources!) by so-called experts. Evidence was sometimes used, and the sources were quoted. But collecting evidence is not as easy as it sounds. James Murray, editor of the Oxford English Dictionary, asked volunteer readers to read the books given to them and copy out any good examples of words in use. However, he later complained:

> *The editor or his assistants have to search for precious hours for examples of common words, which readers passed by… Thus, of abusion, we found in the slips about 50 instances: of abuse not five… There was not a single quotation for imaginable - a word used by Chaucer, Sir Thomas More, and Milton.* (James Murray, Address to the Philological Society, 1879)

This shows our natural human tendency to notice unusual things, while overlooking ordinary, everyday things, whether in language or in other fields. The methodology also raises other problems: who chooses the readers, and who chooses the texts to be read. Another problem with using citations as evidence is that we can select only examples that prove our point. Dr Johnson, in his Dictionary (1757), even altered quotes from famous authors (such as Milton and Dryden) and well-known texts (such as the Bible) to suit his purpose.

2.2 General Dictionaries and Specialist Dictionaries

What are the differences between general dictionaries and specialist dictionaries? Bo Svensen makes the following points in *Practical Lexicography* (OUP 1993):

> p. 21 ff "It is easier to determine which words are NOT specialist terms within a given specialist area than to determine which words ARE specialist terms and should be omitted from a general dictionary... The difference is not just in the selection of entries; they also have different methods of description and explanation."
>
> p. 49 ff "Words can move from *general* to *special* or vice versa. They often move from *special* to *general* a) via education and the media: politics, history, natural sciences, technology, economics, computing. b) via consuming goods/services: food, clothing, building, transport, commerce, law,

> medicine. c) via leisure activities: sports, art, literature, music, drama, hobbies. d) via sudden or brief innovations and discoveries... Not only the individual words are different, but also the constructions, collocations, and idioms... Technical language resources often emphasize terms, rather than other aspects."

Most traditional resources focus on terminology, and much less attention is paid to the "constructions, collocations, and idioms". For example, mountaineering texts often use the verb *take* in expressions like *take the north face of the mountain* (where *take* = climb, attempt to climb), yet this is not recorded in most mountaineering dictionaries (i.e. the movement of words from general language to special language is less likely to be noticed).

2.3 Specialist Dictionaries

So what do specialist dictionaries look like, and what information do they contain? Examples from computing and medical dictionaries highlighted various features: different typefaces (bold for headwords, roman for definitions, italics for `trademark', etc); no grammar or pronunciation; no examples; diagrams (why do certain entries have them and others not?); technical terms in the definitions in capital letters (i.e. acting as cross-references). The term *architecture* is now used very commonly in the domain of computing and therefore, if you do a web search for example, most documents found will be about computing, not about buildings! The definition for *Applications Program Interface* in the computing dictionary was much more detailed than the entry in the medical dictionary.

A major problem with printed reference books, especially in technical domains, is that technology changes very quickly, so the books soon become out-of-date. Fields like computing really require a new edition every year or so, whereas editions of general language dictionaries usually appear at much longer intervals.

2.4 Academic and scientific writing

Examining a typical academic text on architecture, we noticed that the style of writing involved long, complex sentences, containing many technical terms (e.g. *motif, relief, figure, plane, contour, composition*). In fact, many of these terms are actually general words with special meanings (see 2.2) in the field of architecture. Not only the words, but the phraseology and expressions (e.g. *represents a borrowing from...*) are typical. This kind of nominalization (cf. *borrows from*) is a common feature (other examples included *provides a deeply shadowed reinforcement* instead of *reinforces*, and *The treatment of these flat, decorative features* instead of *These flat decorative features are treated*).

Halliday and Martin, in *Writing Science* (Falmer Press 1993), discuss the main features of academic writing:

1. Not just terminology, but `wording'... technical grammar deploys nominal groups and clauses in rhetorical structures to form arguments.
2. Verbs and adjectives are transformed into nouns, which allows "new" information from previous discourse to be reused as "given".
3. Extending the nominal group, using prepositional phrases, embedded clauses, and recursion. The noun group generates "objectivity" which allows reasoned argument.

According to Halliday and Martin, the characteristics of scientific English are:

1. interlocking definitions (e.g. *This distance is called the radius. The diameter of a circle is twice the radius.*)
2. technical taxonomies (superordination, composition; e.g. suffixes such as -berry, -fish, etc)
3. special expressions (technical grammar; e.g. *The process of finding the truth set is called "solving the open sentence over D".*)
4. lexical density (i.e. the number of lexical items or content words per clause)
5. syntactic ambiguity (e.g. *Lung cancer death rates are associated with/reflected in smoking*. Does this mean they are the cause or just the evidence? Does rate mean "number" or "speed"?)
6. grammatical metaphor (i.e. one class or structure is substituted for another, abandoning the traditional association of verbs with processes, nouns with participants, adjectives with qualities, adverbials with

circumstances, conjunctions with process relations, and modals with assessment; e.g. *he departed > his departure.... unstable > instability*)
7. semantic discontinuity (i.e. scientific writing makes semantic leaps which readers are expected to follow)

2.5 Traditional Linguistics and Corpus Linguistics
In traditional linguistics, the use of empirical methods was suppressed for several decades by the dominance of Chomsky's ideas. Chomsky said that language data consists of utterances, which are part of "performance", and are therefore susceptible to error, idiosyncrasy, and local circumstances, whereas linguists should be studying "competence", the thought processes which generate utterances. But Chomsky's ideas are now being strongly challenged:

> "The Chomskyan position on induction is closely related to the langue-parole and competence-performance distinctions. But what … frequency data make very clear is the ultimate inseparability of system and use." (Halliday 1993)

> "Chomsky's (1957, 1965) rejection of induction, by machines or humans, is still widely assumed to be valid. In his attack on American structuralism, he rejects the concept of "discovery procedures". But he provides no real arguments against such methods, merely stating that linguistic theory is not "a manual of procedures", and asserting that there are simply no practical and mechanical ways of extracting a grammar from a corpus of utterances (1957: 50ff; 1965: 18)…" (Michael Stubbs, *Functions of Language*, 2, 1: 1-33; 1995)

Stubbs later stresses that "no procedures can ever be entirely automatic. We always start with intuitions about what is interesting to study, and intuition re-enters in designing procedures and interpreting findings." So corpus linguistics does not reject intuition, but its predominance is reduced.

Traditional linguistics (including Chomsky) focussed on grammar, and grammar was regarded primarily as a sentence level feature. Grammaticality was regarded as the ultimate test of validity for any invented example.
The other linguistic levels above and below the sentence (especially the word level or *lexis*) were largely ignored or subordinated. One modern British linguistics tradition, founded by J.R. Firth and continued by Michael Halliday and John Sinclair, recognizes a continuum called *lexico-grammar*. Lexis and grammar are ways of looking at language "through different ends of the same telescope", and so lexis is "the most delicate grammar". At Birmingham University, John Sinclair pioneered studies at both the discourse level and the lexical level. In the Cobuild Grammar (1990), he linked the levels of language from morpheme to discourse with structural items (e.g. morpheme: noun, adjective, verb and adverb inflections) and linguistic functions (concept identification, message building, and so on).

2.6 Why bother with a corpus?
This is a brief review of arguments that have taken place over the past 20 years or so. Most linguists now concede that there are some interesting and useful findings coming out of corpus/computer-based research.
1. Even "expert speakers" have only a partial knowledge of a language. A corpus can be more comprehensive and balanced.
2. Even expert speakers tend to notice the unusual and think of what is possible. A corpus can show us what is common and typical.
3. Even expert speakers cannot quantify their knowledge of language. A corpus can give us accurate statistics.
4. Even expert speakers cannot remember everything they know. A corpus can store and recall all the information that has been input.
5. Even expert speakers cannot make up natural examples. A corpus can provide us with a vast number of real examples.

Note the shift from "grammaticality" to "naturalness". Many grammatically correct sentences are very unlikely to be uttered, not only because of semantics (e.g. Chomsky's example *Colourless green ideas dream furiously*) but also because of naturalness. Corpus evidence often shows that invented examples tend

to be “unnatural” in some way: e.g. *He argued well* was found in a dictionary, but the Cobuild Bank of English corpus of 323 million words (see below) had no evidence for *argue well* at all, although there were many examples for *argue strongly, passionately, forcefully, successfully, convincingly*, etc.

Corpora give us reliable information about the language: which words are common, and which are rare; new words coming into the language, and old ones dropping out of use; which meanings of a word are common, and which are rare; words developing new meanings, or losing old ones; typical contexts and grammatical patterns for each word; which words are more commonly used in speech than in writing, in British rather than American English, in popular magazines rather than in academic books, etc.

Corpora also help to make dictionaries more accurate, reliable, and authoritative: words included in the dictionaries are in regular use, words that are omitted are very rare or out-of-date; normal, everyday meanings are given first, and historical or technical meanings later; new words and meanings are noticed immediately and are carefully monitored, so that only genuine additions to the language are included, but one-off inventions or temporary fads are filtered out; usage notes are kept up-to-date, reflecting changes in social attitudes; authentic examples are taken directly from the corpus.

2.7 Corpus Research Theory
Two quotes effectively summarize the methodology used in corpus research:

> “Rather than imposing such a model on the data, Sinclair looks for ways of deriving the theory from the data: the concept of data-driven description becomes central.” (Stubbs 1996: 28)

> “Linguistics usually operates with … abstract categories … But … it is good policy to defer the use of them for as long as possible, to refrain from imposing analytical categories from the outside.” (Sinclair 1991: 29)

2.8 The shift from “right or wrong” to “common or rare”
Corpus evidence for the use of *hopefully* shows that its use as a sentence adverb is very common. However, some traditional linguists would still say that this use is “wrong”, because they don't want to accept the evidence, even substantial corpus evidence. Corpus linguists prefer to use descriptive and quantitative terms, and talk about *frequent*, *common*, or *unmarked* uses as opposed to *rare* or *marked* uses, rather than using evaluative and prescriptive terms such as *right* and *wrong*.

Prescriptivism denies the right of the individual language user to be creative in his or her use of the language, and ignores the fact that languages change over time, according to the needs and preferences of the language community, and not in obedience to any rules that individual speakers may wish to preserve. Some countries now try to control language change through their national language academies. They generally fail, because language is a means of communication and will change to meet the changing communicative needs of its users. This problem is not new. The Roman poet Horace mentioned it in his Ars Poetica (c. 65-68 BC):

> *“Many things are resurrected which once had passed away, and expressions which are now respected in turn will pass, if usage so decrees - the usage over which the authority and norm of daily speech have final jurisdiction.”*

His sentiments are remarkably in tune with corpus linguistics.

3: Corpora: History, Design, Construction, Availability

3.1 Introduction
Language has been transmitted in various forms in the past. Oral traditions preceded written ones, writing systems utilised a variety of surfaces: tree-bark, palm leaves, stone, animal skins, paper. Printing technology, sound recording systems, visual recording techniques, and transmission systems have all now entered the electronic age through computers, the internet and email. Computer technology has progressed at breathtaking speed. Computers have become smaller, faster, and cheaper, and are able to do more

processes simultaneously, with greater memory and data storage facilities. We now see converging interactive technologies combining television, telephone, and computer.

Computer use was concentrated initially among mathematicians, physicists and engineers. Then it became available to large administrative/bureaucratic systems, then to commercial organizations and finally to domestic applications. Computers were first used for mathematics. The first computerized language work was done in the 1950s and 1960s. Professor John Sinclair of Birmingham University was one of the pioneers, putting together one of the earliest corpora of spoken language - 120,000 words, all input on punched cards.

3.2 A Brief History of Language Corpora

Much of the early corpus work was done on the English language, but now most languages have corpus developments in progress. Parallel and translation corpora have been developed and multilingual corpora are emerging. Specialist corpora have been built for a variety of purposes: for Natural Language Processing, for diachronic language studies, and for research into language variety, first language acquisition, and second language acquisition.

The average size of corpora has increased vastly in the past 30 years, from 1 million words in the 1960s (e.g. the Brown University corpus of written American English) to several hundreds of millions of words (e.g. Cobuild's Bank of English was 323 million words in 1998, 418 million words in October 2000). Many other languages now have corpora of over 100 million words.

The speed of development in corpus linguistics has been immense. This is reflected in the numerous conferences, publications and websites related to the field. The notions of "balance" and "representativeness" continue to undergo considerable attention: if one wishes to make general statements about a language, one must be sure that one has sampled an adequate population of data.

3.3 Corpus Design

There are various models for corpus construction. The model in Figure 1 was used for the Brown corpus, the LOB (London-Oslo-Bergen Universities) corpus, and the ICE (International Corpus of English) corpus at London University. This is an ***a priori*** model, based on the idea that "we know what language is like" and therefore we know what the main constituents of a corpus should be.

Figure 1: The Brown corpus model, also used for the LOB and ICE corpora (Gerald Nelson 1991)

SPOKEN	300	**DIALOGUE**	180	**private**	100	direct	90
						distance	10
				public	80	class lesson	20
						broadcast discussion	20
						broadcast interview	10
						parliamentary debate	10
						legal cross-examination	10
						business transaction	10
		MONOLOGUE	120	**unscripted**	70	spontaneous commentary	20
						unscripted speech	30
						demonstrations	10
						legal presentation	10
				scripted	50	broadcast news	10
						broadcast stories	10
						broadcast talks	20
						speeches (not broadcast)	10
WRITTEN	200	**NON-PRINTED**	50	**non-professional**	20	student untimed essay	10
						student exam essay	10

				correspondence	30	social letters	15
						business letters	15
		PRINTED	150	**informational (learned)**	40	humanities	10
						social sciences	10
						natural sciences	10
						technology	10
				informational (popular)	40	humanities	10
						social sciences	10
						natural sciences	10
						technology	10
				informational (reportage)	20	press news reports	20
				instructional	20	administrative/regulatory	10
						skills/hobbies	10
				persuasive	10	press editorials	10
				creative	20	novels/stories	20

For convenience, the model envisages 1 million words as 500 texts of 2000 words. But many texts are much longer than 2000 words (a novel is about 50-80,000 words, a broadsheet newspaper is over 100,000 words, many university textbooks are even longer), so will 2000 words be sufficient to typify a speaker's or an author's style? Which part of a longer text should be selected? The beginning, middle or end? Many texts are shorter than 2000 words (e.g. telephone calls, advertisements, student exam essays, social letters, business letters, or press editorials), so how do we ensure a representative selection?

This model is **idealistic** (it ignores the fact that some data types may be difficult to obtain, such as business correspondence), **static** (it does not allow for new types of text, such as email or mobile phone text messages), and **synchronic** (the texts are taken from one brief time period - one year, in fact). So it will not be suitable for many types of language study. But the main weakness of the model lies in the proportions allocated to the different text types: why 300 spoken texts and 200 written texts, why 180 dialogues and 120 monologues, and so on? These proportions are rather arbitrary.

The Longman-Lancaster corpus model (Della Summers, 1992) has a 50% "selective" section and a 50% "microcosmic" section. The "selective" section of the written corpus is a variation on the Brown model, being divided into *imaginative* (i.e. fiction) and *informative* (i.e. non-fiction) components. Individual texts are chosen according to "real-world" criteria such as i*nfluentialness*, *popularity*, and *educational status*. The "microcosmic" section uses a computerized random number generator to pick texts from a large catalogue. The selective section of the spoken corpus takes predetermined proportions of certain text types (lectures, sales talks, interviews, debates, radio and TV broadcasts, etc) and the microcosmic section takes randomly selected texts from a representative sample of the population.

Cobuild's first corpus (Looking Up, ed. J.M. Sinclair, Collins ELT, 1987) was built in order to write an EFL dictionary, so did not include technical texts, drama, poetry, or children's literature. Of the total of 214 books in the corpus, 7 were published before 1950, 9 in 1950-59, 32 in 1960-69, 33 in 1970-74, 72 in 1975-79, and 61 in 1980-81; 164 books were written by male authors, 40 by female authors; 157 books were written in British English, 45 in American English. However, the acquisition of data has been continuous, and the corpus has grown from 7.3 million words (1981-83) to 18 million words (1985), 120 million words (1993), 167 million words (1994), 211 million words (1995), 323 million words (1996) and 418 million words (2000).

3.4 Corpus Creation Processes

The main processes involved in creating a corpus are: selection, permission, accession, data input (data conversion, scanning, keyboarding, transcription) and data processing (spellchecking, encoding, and indexing). Cobuild believes in minimal coding, but other corpora are heavily encoded and annotated.

Typical costs in the mid-1990s in the UK were £60 per million words for data conversion, £1,500 for optical scanning, £3,300 for keyboarding, and £25,000 for transcription. The automatic conversion of electronic data is by far the cheapest process and transcription of speech is the most expensive. Optical scanning works for good quality printed texts, but keyboarding is necessary for poorer quality printed texts (e.g. local newspapers, ephemera), or where the text is complex (including tables, graphs, etc), disjointed (e.g. magazine articles interrupted by advertisements), or printed in odd formats (e.g. overlay, slanted). Transcription of audio data is so expensive that the Brown model (3/5 spoken data) is simply not economically viable for a 400 million-word corpus!

Figure 2 shows a summary of the current Bank of English corpus contents. Note the variety of data sources (column 1), variations in subcorpus size (column 2) and number of texts (column 3), average text sizes for each subcorpus (column 4), and dates (column 5). What is "a text"? For example, is a newspaper a text (probably edited into a common "house style"), or is each article in it a text (individual journalists, and especially external contributors may have their own styles)?

Figure 2: Composition of the Cobuild Bank of English 418 million word Corpus, 2000

American Books **32m words** **327 texts** **mostly 1990 >**
266 non-fiction, 61 fiction; 169 male authors, 63 female, 12 joint male-female, 83 other.

American Radio (NPR) **22m words** **726 texts** **1990-93**
national; National Public Radio, Washington, USA; 1990 Sep-Dec (88 broadcast programs),
1991 Jan-Dec (244 programs), 1992 Apr-Dec (259 programs), 1993 Jan-May (135 programs)

BBC World Service **18.5m words** **143 texts** **1990-91**
international; News, Current Affairs, Sport, Medicine Now, Science Now, Regional broadcasts;
1990 Apr-Sep (130 broadcast programmes), 1991 Jul-Aug (13 programmes)

British Books **43m words** **578 texts** **mostly 1990 >**
384 non-fiction, 188 fiction; 300 male authors, 189 female, 27 joint male-female, 56 other.

British Ephemera **4.5m words** **2359 texts** **mostly 1991-6**
junk mail, leaflets, newsletters, brochures.

British Magazines **44m words** **1113 texts** **1992-00**
c. 75 different monthly/weekly titles; women's, men's, sports, hobbies, ethnic, lifestyle, music, DIY.

British Spoken **20m words** **2670 texts** **mostly 1991-96**
radio phone-ins, lectures, meetings, interviews, private phone calls, social events; 2200 items from 1991-6;
64 from 1978-89; 314 not yet dated.

Economist **15.5m words** **229 texts** **1991-99**
international weekly journal; 1991 Jan-Dec (51 issues), 1992 Jul-Dec (26 issues), 1993 Jan-Jun (25 issues),
1994 Jan-Jun (25 issues), 1995 Jan-Dec (51 issues), 1998 Jul-Dec (25 issues), 1999 Jan-Jun (26 issues)

Independent **30m words** **260 texts** **1990-99**
national daily broadsheet newspaper; Independent (weekdays) and Independent on Sunday;
1990 Jan, Nov (20 issues), 1995 Jul-Dec (71 issues), 1998 Jul-Dec (85 issues), 1999 Jan-Jun (84 issues)

Times **30m words** **208 texts** **1995-00**
national daily broadsheet newspaper; Times (weekdays) and Sunday Times; 1995 Nov, Dec (29 issues),
1996 Jan-Apr (41 issues), 1999 Mar, May, Jul, Sep, Nov (78 issues), 2000 Jan, Feb (60 issues)

Today **26m words** **794 texts** **1992-95**
national daily tabloid newspaper, ceased publication in c. 1996; 1992 Jan-Dec (285 issues),
1993 Dec (3 issues), 1994 Jan-Dec (255 issues), 1995 Feb-Nov (230 issues)

Guardian **32m words** **332 texts** **1995-99**
national daily broadsheet newspaper; 1995 Jan-Nov (180 issues), 1999 Jan-Jun (152 issues)

New Scientist **7.9m words** **188 texts** **1992-99**
international weekly journal; 1992 Jan-Dec (51 issues), 1993 Jan-Apr (16 issues), 1994 Jan-Jun (25 issues),
1995 Jan-Dec (46 issues), 1998 Jul-Dec (25 issues), 1999 Jan-Jun (25 issues)

Australian newspapers **34m words** **406 texts** **1995-99**
regional daily newspaper, Queensland; Courier Mail (weekdays) and Sunday Mail;
1995 Jan, Sep, Oct (80 daily issues), 1998 Oct-Dec (91 daily issues), 1999 Jan-Aug (235 daily issues)

American Ephemera **3.5m words** **2786 texts** **mostly 1995-96**

junk mail, leaflets, newsletters, brochures; 1053 items from 1995, 347 from 1996, 22 from 1989-94; 1200 not yet dated (but mostly 1995/96); 111 date unknown.

American newspapers **10m words** **281 texts** **1994-96**
regional and local website newspapers; Palo Alto News 1994 Jan-Nov (84 daily issues), Seattle Times 1996 Feb-May (122 daily issues), Houston Chronicle 1996.

Sun and News of the World **31m words** **597 texts** **1997-00**
national daily tabloid newspaper; Sun (weekdays), News of the World (Sundays); 1997 Nov (1 daily issue), 1998 Apr-Dec (203 daily issues), 1999 Jan-Dec (333 daily issues), 2000 Jan, Feb (60 daily issues)

American Academic textbooks **6m words** **31 texts** **1990-96**
6 Politics & Government, 5 Psychology, 5 Sociology/Anthropology, 5 History, 4 Gender studies 3 Economics, 1 Health, 1 Biology, 1 Oceanography)

American Spoken **2m words** **16 texts** **1994-97**
professional, mainly political/journalistic and academic; University committee meetings 1995-97; White House press briefings 1994-97.

TOTAL **418,449,873 words** **14,605 texts**

The Cobuild system of annual, retrospective or ***post hoc*** balancing is a model for a ***dynamic*** corpus. The general aim each year is to increase the overall size of the corpus, to replace older data with newer data where possible, to increase the range of data types and sources, and to reduce the imbalances between subcorpora (e.g. British Books data was 25% of the whole corpus in 1993, 13% in 1996; British Spoken was 3% in 1993 and 5% in 1996). A dynamic corpus is not ideal for some academic research, where a stable dataset is needed over a period of several years, but the main purpose of the Bank of English is to provide the basis for up-to-date new dictionaries, reflecting the immediately current language.

3.5 The Design of Specialist Corpora

Figure 3 is a design document showing the different text types which may need to be collected in order to construct a corpus of Business Language. The upper half of the diagram represents spoken texts, and the lower half consists of written texts. The left side of the diagram represents texts produced for "real world" activities (e.g. buying and selling, ordering, supplying), while the right side shows texts produced mainly or solely for academic purposes (e.g. seminars, exam essays, and text books). Some types of language activity that may take place in both: presentations, meetings, case studies. Some texts (e.g. business journals and the financial press) serve as resources for both the real world and the academic community, and the authors or contributors may belong to either camp.

Figure 3: A Corpus of Business English (English Department, University of Birmingham)

"REAL WORLD"		**ACADEMIC PURPOSES**	
Social	Travel		
Committees	Presentations		
Dealing with customers	Buying and selling		Seminars and lectures
	Meetings		
Telephone			
Telex			
Letters	Order, Supply, Remind, Complain, Apologize …		Exam essays
			Assignments
Memos	Report writing		
Email	Case studies		
	Summary writing		
Documentation	Import/Export		
Promotional literature			Text books
Job specifications and advertisements			
	Business journals and magazines		

	Financial press	

3.6 Corpus Size

How big should a corpus be? There is no obvious way of answering this question, and any answer will obviously have to reflect the purpose for which the corpus was built. Before proceeding any further, we must be more exact in our terms. What is a "word"? The sentence *The cat sat on the mat* contains 6 "words" in total, but it has two occurrences of the same "word": *the*. In corpus linguistics, we prefer to use the terms ***types*** and ***tokens***. The sentence has 6 ***tokens***, but only 5 ***types***, because there are two ***tokens*** of the ***type*** *the*.

Figure 4 shows that a corpus of 18 million tokens yields information for about 19,808 potential dictionary entries (this is an over-estimate: many of the types in the corpus will be proper names and other encyclopaedic items), assuming we need at least 10 examples of a word before we can define it and describe its behaviour (and this is an under-estimate: we often need more than 10 examples, especially if a word has several meanings or uses). The 323 million-word corpus is 18 times bigger, but the potential dictionary entries have only increased by a factor of less than 4.

Figure 4: Types and tokens in a corpus, and headwords in a dictionary

	Collins COBUILD Bank of English corpus			
	1987	**1993**	**1995**	**1996**
	18 million words	**120 million words**	**211 million words**	**323 million words**
tokens	18,000,000	120,000,000	211,505,963	323,302,789
types	247,069	475,633	638,901	812,452
hapax types	131,299	213,684	296,436	383,356
non-hapax types	115,770	261,949	342,465	429,096
types with frequency 10>	43,579	104,201	134,942	164,963
There are approx. 2.2 types per dictionary headword (or lemma).				
potential dictionary headwords	19,808	47,364	50,458	74,833

Compare these figures with Figure 5, showing the number of headwords in various dictionaries.

Figure 5: How many words are there actually in dictionaries?

ENGLISH LEARNER'S DICTIONARIES:	
Oxford (OALD 1995)	*63,000 references*
Collins COBUILD (CCED 1995)	*75,000 references*
Longman (LDOCE 1995)	*80,000 words/phrases*
Cambridge (CIDE 1995)	*100,000 words/phrases*
NATIVE-SPEAKER ENGLISH DICTIONARIES:	
Collins English Dictionary (CED 1992)	*180,000 references (190,000 numbered definitions, 3.5 million words of text)*
American Heritage Dictionary (AHD 1992)	*350,000+ entries/meanings*
BILINGUAL DICTIONARIES:	
Collins Spanish-English, English-Spanish (1992)	*230,000+ references, 440,000+ translations*
Collins Sansoni Italian-English, English-Italian (1988)	*240,000+ references, 570,000+ translations*
Oxford Hachette French-English, English-French (1994)	*350,000+ words/phrases, 530,000+ translations*

Of course, increasing the size of the corpus does help. Words with few examples in a smaller corpus have many more examples in larger corpora, so we can do more accurate and detailed analyses.

Figure 6: A larger corpus gives more detailed information

	Collins COBUILD Bank of English corpus		
	1986	**1995**	**2000**
	18 million words	**211 million words**	**418 million words**
impending	77	922	2251
clemency	7	188	464
lilting	10	93	209
facilitator	1	96	195
matriarchal	9	68	157
regretful	17	70	155

Although it is important to know the bases of corpus design, most people will not need to construct their own general corpora, because many of these already exist. However, they may want to create small corpora for specific purposes, and fortunately these are quite easy to create nowadays, with the wealth of material available on the Web.

4: Corpus: resources and facilities

4.1 Frequency

We assume for the moment that there must be some relationship between the frequency with which a word occurs in a corpus, and the importance of that word in the linguistic system. The most frequent words in the language remain in roughly the same order in a frequency table, whatever the size of the corpus (and whatever the date; the most frequent words take much longer to change in use): in English, *the, of, and*, etc.

Figure 7: The most frequent words of English

20m corpus (1987)		**211m corpus (1995)**		**323m corpus (1996)**		**418m corpus (2000)**	
the	1,081,654	***the***	11,611,078	***the***	17,845,265	***the***	22,849,031
of	535,391	***of***	5,359,185	***of***	8,321,813	***of***	10,551,630
and	511,333	***to***	5,180,130	***to***	8,034,738	***to***	10,429,009
to	479,191	***and***	4,941,561	***and***	7,852,198	***and***	9,787,093
a	419,798	***a***	4,537,660	***a***	7,061,412	***a***	9,279,905
in	334,183	***in***	3,796,752	***in***	5,854,481	***in***	7,518,069
that	215,322	***that***	2,226,871	***that***	3,323,182	***that***	4,175,495
it	198,578	***it***	1,954,556	***is***	2,970,503	***s***	4,072,762
i	197,055	***is***	1,940,162	***it***	2,297,913	***is***	3,900,784
was	194,286	***for***	1,794,630	***for***	2,794,483	***it***	3,771,509

Note: For computational reasons, in the first list ***i*** includes all the occurrences of ***I***, the first person subject pronoun (because all words are lowercased); and in the fourth list, ***s*** includes all occurrences of ***'s*** (both the genitive marker and the contraction for *is*).

The relation between rank and frequency of items in a corpus frequency list has been investigated and the principal formula is known as *Zipf's Law*: "Frequency is inversely proportional to rank" (Zipf, Human Behaviour and the Principle of Least Effort, 1949), which can be expressed mathematically as: $a_n = k*^1/_n$. This formula seems to work for most of the items in corpus frequency list except for the most frequent words.

The computer can arrange the corpus frequency lists in many different ways. The most widely and most commonly used format is the one shown above, the ***frequency* ordered** list (strictly speaking, *reverse* frequency ordered list, because the list is in descending order, starting with the highest frequency words). The items at the bottom of the list have only a single occurrence in the corpus, and many of these will be eccentric or rare words, or typographical errors. Figure 8 shows the top 60 most frequent words in the 418 million word Bank of English corpus:

Figure 8: Corpus frequency list: in frequency order

the	22849031	you	2132436	there	1097305
of	10551630	by	2024320	who	1059554
to	10429009	but	2021398	which	1052720
and	9787093	have	1928894	all	1051499
a	9279905	his	1855728	were	1043408
in	7518069	are	1843433	she	1026560
that	4175495	from	1784772	been	1019904
s	4072762	they	1743087	up	964141
is	3900784	this	1601898	her	951899
	3898632	not	1540940	when	942884
it	3771509	has	1428566		933232
for	3690466	we	1375655	if	916009
i	3216005	had	1372891	would	913635
was	3092967	an	1340348	more	905792
on	2936269	t	1216283	so	885702
he	2729706	will	1179145	out	879763
with	2717188	or	1171981	about	873031
as	2382198	their	1145185	can	868740
be	2207680	one	1142348	what	827793
at	2183544	said	1119996	no	764972

Note: See notes to Figure 9 for ***i*** and ***s***; similarly, here, ***t*** includes all occurrences of ***'t*** (as in *can't*, *won't*, etc); and are the codes used in the corpus to indicate the beginning and end of a paragraph.

Linguistically, this list is very interesting: all the items are manifestly *grammatical* words or *function* words, i.e. articles, prepositions, pronouns, modal verbs, auxiliary verbs, the question words (which, when, what, etc) and so on.

The frequency list can also be displayed in alphabetical order. This is particularly helpful to lexicographers: before writing dictionary entries, they can see which words occur frequently enough to be included in the dictionary. These lists can also be helpful to teachers in deciding which items to teach.

Figure 9: Corpus frequency list: in alphabetical order

technocracy	32
technocrat	172
technocrates	4
technocratic	175
technocratica	1
technocratically	4
technocrats	438

Using the cut-off point of a minimum of 10 examples (mentioned earlier), the candidates for our dictionary would be: *technocracy* 32, *technocrat* 172, *technocratic* 175, and *technocrats* 438.

Another useful display is words ending in particular suffixes. Such lists can provide extremely useful classroom teaching materials.

Figure 10: Corpus frequency list: words ending in *-ness*, in frequency order

business	167046	darkness	7424	wilderness	3819
illness	15273	happiness	7031	goodness	3776
fitness	11821	weakness	6941	guinness	3614
witness	11578	madness	5011	sickness	3386
awareness	9423	willingness	4957	sadness	3256
consciousness	8797	effectiveness	4197	fairness	2764

Let a group of students look at the list of most frequent items ending in *-ness*, and ask them what all the words have in common. Once they have identified *-ness*, the teacher can discuss morphology in more general terms. What part of speech do the words in *-ness* belong to? What do nouns ending in *-ness* mean? If we remove *-ness* from words, what part of speech do the root words belong to? One interesting feature of using frequency lists in teaching is that they tend to include important exceptions. In the above list, *business* not only requires the substitution of *i* for *y*, but its main meaning is no longer related to the main meaning of *busy*; *wit* is not an adjective, and *witness* is not connected to it in meaning; *harness* is not a suffixed item at all, etc. (Note that *guinness*, lowercased for computational purposes, and other proper nouns may need to be explained or omitted).

A similar exercise can be created from the list of words ending in *-ical*. Again, discussions of morphology, wordclass (which words are both adjectives and nouns?) and semantics can be stimulated by lists such as these.

Not only suffixes, but other productive elements in word formation can be investigated. Surprising new elements like *-buster* can lead to amusing but instructive research by the students themselves. Which words ending in *-buster* refer to human beings, and which do not? What do the non-human words refer to? What part of speech do *-buster* words belong to? Do they belong to more than one wordclass? In what type of texts do *-buster* words occur in? Who invents them or uses them?

Figure 11: Corpus frequency list: words ending in *-buster*

blockbuster	2038	tankbuster	13	virusbuster	5
buster	1366	trustbuster	12	gumbuster	5
filibuster	302	chartbuster	10	barkbuster	5
bonkbuster	33	stressbuster	7	ballbuster	5
ghostbuster	27	snorebuster	7	heatbuster	4
sleazebuster	19	leafbuster	7	drugbuster	4
crimebuster	18	sodbuster	6	nosebuster	3
fraudbuster	15	gangbuster	6	fuzzbuster	3
dustbuster	15	gamebuster	6	fatbuster	3
dambuster	14	cheatbuster	6	witchbuster	2

Looking at corpus data in other languages is a useful way of pointing out to students the fact that some linguistic features are language-dependent: for example German has some different characters (ß, *ü, ä, ö*), grammatical gender, and is much more inflected; capital letters can be grammatically significant (e.g. *Leben* is a noun, *leben* is a verb), so should be preserved in German corpus displays. This frequency list is from the Munster newspaper corpus (a 36 million word corpus containing c. 300,000 types) of *Die Zeit* and *Frankfurter Allgemeine*.

Figure 12: A German corpus frequency list

der	1219472	des	273429	als	177761
die	1039479	sich	259388	an	160760
und	787438	für	247160	auch	158335
in	621484	auf	242308	es	156934
den	418040	im	237872	Der	144669
von	368004	nicht	236532	aus	141765
zu	297413	dem	236007	daß	139291
mit	292663	ist	224073	werden	133465
das	279250	ein	194417	hat	127035
Die	274900	eine	184293	nach	124868

4.2 Lemmatization

Earlier, we grouped together the ***tokens*** (individual instances, examples, or occurrences of a word) into the ***types*** of which the tokens are realizations. But what term can we use to refer to all the forms associated with the same `root word'? Some people use the term ***lemma***. It is similar in meaning to the linguistics term ***lexeme***, but without the specification of semantic properties. It is also similar to the dictionary concept of ***headword*** or ***entry word***. Some people limit ***lemma*** to the inflected forms of a word, others include derivational and compound forms as well. Patrick Hanks (How common is common, 1988:6) discusses the problem:

> "We still do not have a term to account for the different forms that a word can take. The term used by linguists for this is `lemma'. Thus, *umbrella* and *umbrellas* are the two types of the lemma UMBRELLA. *Go, goes, going, went,* and *gone* are the five types of the lemma GO, and so on... Dictionaries traditionally regard *strong, stronger, strongest* as one lemma, *strength* as another lemma, and *strengthen, strengthens, strengthening* as a third. In these dictionaries the status of a derived form such as *strongly* is indeterminate."

In corpus linguistics, different analyses may require different sets of forms, so automatic *lemmatization* usually remains an optional facility in corpus software systems.

4.3 Part-of-Speech Tagging

Most corpus systems have the facility to tag each word (token) in the corpus with its part-of-speech in each context. Tagging programs have achieved a very high rate of accuracy in recent years. From a tagged corpus, we can generate frequency lists with grammar information. Here are the top items in the Bank of English tagged frequency list:

Figure 13: Corpus frequency list with part-of-speech tags

the	DT	22842362	<p>	NONE	3898584
of	IN	10548182	is	BEZ	3895398
and	CC	9783949	s	BEZ	3769495
a	DT	9269619	to	IN	3706518
in	IN	7471445	for	IN	3690461
to	TO	6717200	i	PPS	3146978

There are various taggers available, and each tagger may have its own set of tags, or may be able to use more than one tagset. Here are some examples from the current Cobuild tagset:

Figure 14: Part-of-speech tags

BE	verb 'to be', base form: *be*
BED	verb 'to be', past tense: *were*
BEDZ	verb 'to be', past tense: *was*

BEG	verb: 'to be', -ING form: *being*
BEM	verb 'to be', 1st person, present tense, singular: *am*
BEN	verb 'to be', past participle: *been*
BER	verb 'to be', 3rd person, present tense, plural: *are*
BEZ	verb 'to be', 3rd person, present tense, singular: *is*
CC	co-ordinating conjunction (*and*, *or*, etc)
CD	(cardinal) number
CS	subordinating conjunction (*unless*, *although*, etc)
DEM	demonstrative pronoun (*this*, *that*, etc)

Once the corpus has been tagged, it is also possible to produce a ***lemmatized*** word-and-POS-tag corpus frequency list:

Figure 15: Lemmatized corpus frequency list

```
the   DT 17845179
be    V 11989968  VB be 1702992     VBD were 828676   VBD weren 14366
      VBDZ was 2423790  VBG being 274115  VBM am 68169      VBM m 207711
      VBN been 801065   VBR are 1407720   VBR aren 20257    VBR art 80
      VBR re 246235     VBZ am 7823 VBZ is 2970439    VBZ isn 50691
      VBZ s 910625      VBZ was 13  VBZ wasn 55201
of    IN 8321789
and   CC 7582146
a     DT 7060847
in    IN 5826066
to    TO 5132283
have  V 4087004   VB hast 146 VB have 1465990   VB haven 29534
      VB ve 231081      VBD had 1095919   VBD hadn 18131    VBD haved 2
      VBG haveing 4     VBG having 101243 VBN had 16726     VBN haved 6
      VBZ has 1075870   VBZ hasn 14970    VBZ hath 550      VBZ s 36832
to    IN 2902437
for   IN 2794483
```

4.4 Language Change

The most frequent words may not change much over time, but corpus frequency lists can be very helpful for detecting changes lower down the list. In particular, comparative frequency lists over a period of time can identify new words. Changes in technology have a particularly enormous impact on society. In the last century, the pace of technological change has increased dramatically. Another recent phenomenon has been the global nature of such changes. The effects of technological changes on society, on the economy in particular, are usually well documented in the media.

Every single innovation not only has to has to have a word to describe it, but usually generates several words, and often a whole domain of vocabulary. In some cases, old words are given new meanings, like the computer *mouse*. *Quark* was coined by James Joyce in Finnegan's Wake and later adopted as a technical term by physicists. Old words or elements are often placed together to form new combinations or compounds, e.g. *roller skates* and *telephone*. The name of the inventor is sometimes used (e.g. *biro*), or a trademark (e.g. *hoover, lycra*). Words may be imported from another language (e.g. *karaoke* from Japanese). Nouns often come first, but may acquire other wordclasses, e.g. the noun *email* has acquired a verb use; sending text messages on mobile phones has generated a new verb: *She was texting her friends*. Here is a selection of new words from the technology domain in the period 1985-1995:

Figure 16: New Technology generates New Words

	1985	**1995**		**1985**	**1995**
	18 million words	**211 million words**		**18 million words**	**211 million words**

camcorder	0	1214	microsurgery	0	50
cyborg	2	31	mobile phone	0	455
email	0	39	palmcorder	0	86
gopher	2	35	palmtop	0	25
helipad	0	27	satellite dish	0	236
hypertext	0	13	smart card	0	68
imaging	7	463	teleworker	0	46
laptop	0	184	videophone	0	144

If we select from a larger variety of domains, we find the new words of the 1990s include such items as: *alcopops, bull bars, chaos thory, clone, gridlock, karaoke, listeriosis, pro-choice, shell suit, snail mail, snowboarding, TESSA,* and *white-knuckle rides.*

The corpus is not only a useful source of new words, it can also reveal new uses or meaning changes in common, established words. Look at these examples of *take a bath*: *...those Japanese who took a bath in Bombay... Shareholders have already taken a bath... Investors announced that they were taking a bath* Why were we being told about the personal hygiene of Japanese people in Bombay, shareholders, and investors? A few more examples soon clarify the new meaning: *The entire insurance broking sector took a bath yesterday on Sedgwick's depressed interim results...The Bank of Ireland took a bath in New England, America's most depressed banking market...People who kept on buying in `87 and took a bath, piled further into UK property and took another bath...USair took a bath over American Airlines' price war but not a bad bath.* In the business world, *taking a bath* obviously means “losing a lot of money”.

4.5 Concordances

This is a standard facility of most corpus systems, the ability to display every *token* for a particular *type* (in this case every *token* for the *type*/word *alcopops*), with some context for each *token*. This is called a KWIC concordance (Key Word In Context).

Figure 17: KWIC Concordance for *alcopops*

```
tim N2000960127    brews up a storm over new `alcopops" </h> <b> Olga Craig and Nicholas
tim N2000960127      have followed, with new `alcopops" being launched at the rate of one
tim N2000960127 these products," he said. <p> Alcopops have been criticised by doctors an
tim N2000960127  of Physicians, has described alcopops as `insidious", while 4,000 member
tim N2000960127 ge drinkers are trying to buy alcopops. There is a serious risk of increa
tim N2000960127 he promotion and packaging of alcopops by March. <p> However, Rae, a form
tim N2000960127 of alcoholic lemonade dubbed `alcopops" such as the bluntly named Hooch.
tim N2000960127 living it up. Another case of alcopops. <p> Reporting Scotland's Cameron
tim N2000960131 may be fuming at the rise of `alcopops" such as the alcoholic lemonades
tim N2000960131  Liquid Gold. <p> Already, 30 alcopops are on the market. These include
tim N2000960131  e disappointed. I would give alcopops one more summer." <p> But Paul Mil
tim N2000960131  ager when it was introduced. Alcopops offer people the same `quaffabilit
tim N2000960131 ker's Liquid Gold agrees that alcopops have passed the consumer test. So
tim N2000960131  Research, traces the rise of alcopops to the `beer boredom" which he cla
tim N2000960203 ured alcoholic pops so-called alcopops. <p> He saw the success Bass had w
tim N2000960203 `summer fad" by most experts, alcopops turned out to be the drinks sensat
tim N2000960203 ve half the calories of other alcopops. At the same time, Drnec wants to
tim N2000960203 es and reckons the market for alcopops could top 250m cans and bottles, o
```

Note: This particular display from the Bank of English has some additional information at the beginning of each context or “concordance line”: the first two columns contain corpus codes indicating the source and date of the examples: all of them are from the Times newspaper of 27th January, 31st January, and 3rd February 1996. Other corpus codes are evident: <p> is the paragraph marker, </h> signifies the end of a headline. In general, the Bank of English uses very few such codes compared to some other corpora.

In four of the lines, we notice the use of inverted commas around the word *alcopops* itself; in the 7th and 15th line, we have *dubbed* and *so-called*. These are clear signals that *alcopops* is regarded as a new word in these contexts. In the first two lines, the word is qualified by the adjective *new*, and in other lines we have

words like *rise*, *introduced*, and so on, confirming that *alcopops* are a new product. If we didn't know what *alcopops* were, we have a ready-made brief definition in the two occurrences of *alcoholic lemonade*. In most texts, whether journalistic, scientific, or technological, authors introducing new terms generally give some such explanation to assist their readers. Finally, there is also evidence in the lines of the social attitudes and impact of *alcopops* in such phrases as *brew up a storm*, *criticised by doctors*, *described as `insidious"*, *serious risk*, *fuming*, and so on. These 18 examples were the total evidence for *alcopops* in the corpus (apart from one example of the use of the singular form *alcopop*), and we can see that there is just about enough evidence to create a basic dictionary entry.

When people use corpus examples for academic purposes, they need accurate information about their sources. When looking at the example `usb B9000001383 FBI suggestions to `report all information relating to,` a single keystroke tells us its origin:

```
TITLE: The End of Victory Culture - Cold War America and the Disillusioning of a
Generation
AUTHOR: Engelhardt, Tom
PUBLISHER: BasicBooks (HarperCollins)
YEAR (OF PUBLICATION): 1995
COUNTRY (OF PUBLICATION):  USA
```

4.6 Sorting Concordances

Most corpus software initially displays concordance lines in random order. This may be acceptable when there are only a few examples, as in the case of *alcopops* above, or when only a cursory examination is intended. However, for words with many more examples, or for more detailed analyses, we can get more help. Here are 10 of the 50 lines in the corpus for the word *arrant*, first in random order:

Figure 18: Unsorted Concordance for *arrant*

```
  claims that they had bungled as `arrant nonsense". Criticism centred on
 but given special dispensation is arrant nonsense. Most worrying for
      are on view, but despite the arrant Scottishness of such painters as
   I decided that it was a work of arrant racism, and forgot about it. Last
nd the Army Catering Centre. <p> W arrant Officer Class One Mark STEPHENS,
rganisation". <p> And that is just arrant nonsense. <p> Mrs Mathison is at
    of law; <p> Illustrious heroes arrant vagrants seem'd, <p> And gentlest
         This may have seemed like arrant chauvinism but in fact it is a
  of position. This, of course, is arrant nonsense. Some illnesses are the
silence # <p> So if anyone has the arrant stupidity to ever tell you that
```

Note: The corpus code # in line 10 is used to replace complex punctuation sequences. Line 5 is a corpus processing error: the space between `W` and `arrant` may be an error in the original text, or may have been introduced during the optical scanning, manual editing, or computer processing of the text. The frequency of such errors in the corpus is generally quite low (c. 1-2%), so Cobuild's policy is not to bother to correct them, as they are difficult to find, and expensive to correct, and further errors may be introduced during the correction process. Users are expected to notice and disregard such errors, which do not affect the corpus statistics to any great extent.

We can see that *arrant* is used to qualify the word *nonsense* in 4 of the 10 lines. We can also see some of the other nouns that it qualifies: *Scottishness*, *racism*, *vagrants*, *chauvinism*, *stupidity*. However, we get a much clearer picture if we ask the software to present the lines with the nouns qualified by *arrant* in alphabetical order. For example, we can see that many of the nouns belong to one or two semantic sets: negatively
evaluated people (*bigot*, *coward*, *drunkards*, *knave*) and negative abstract qualities (*bullshit*, *chauvinism*, *effrontery*, *lunacy*, *melodrama*, and of course *nonsense*). Note that, by being used after *arrant*, even normally neutral or positively evaluated words can acquire a negative flavour: *beginner*, *Communists*, *democracy*.

Figure 19: Concordance for *arrant*: sorted by the word to the right

```
   his rivals that even against an arrant beginner, he would sell his
  as an authentic believer, not an arrant bigot. He saw a place for women in
```

```
 Monarchy # which by the way is an arrant Bull, a contradiction in adjecto,
her. She loathed them, called them arrant bullshit, said they'd lick
         This may have seemed like arrant chauvinism but in fact it is a
hich comes down to the same thing. Arrant Communists, all four of them,
xalted by it -- and now I can say, arrant coward that I am, that at least I
  was denouncing some proposal as `arrant democracy"; and no woman under 30
 Seithenin, one of the three great arrant drunkards of the isle of Britain',
endeavour is dogged by a degree of arrant effrontery and there is no reason
endeavour is dogged by a degree of arrant effrontery and there is no reason
welling in all Denmark But he's an arrant knave. <f> Horatio: <f> There
      earth and <f> heaven? We are arrant knaves all; believe none of us.
hastity oath of hers was an act of arrant lunacy, Joanna. She'll never hold
    quickly turns into the kind of arrant melodrama which would be laughable
```

Note: The corpus code `<f>` represents a change of font or typeface, usually used in texts to highlight foreign words, quotations, etc.

Later in the same ***sorted concordance***, we find all the examples of *arrant nonsense* together, and we see that there are 24 of them (out of 50 lines in total for *arrant*), so *nonsense* is the most frequently qualified noun.
If we wanted to, we could then sort these lines again, by the word to the right of *nonsense* (i.e. 2 words to the right of *arrant*), and so on.

Sorting concordances by the word to the right of the "Key Word" (also called the ***node***) allows us to see the common adjective-noun combinations in this case, but it can also reveal other adjacent combinations such as verb-adverb, adverb-adjective, and verb-preposition.

More complex sorting operations can identify multi-word units such as separable phrasal verbs. Here we see all the inflected forms (another useful facility in the corpus software) of the verb *freak* followed two words later by *out*. Note that not all the lines are examples of the phrasal verb (e.g. *these freaks going out in a canoe*).

Figure 20: Concordance for forms of the verb *freak* followed 2 words later by *out*

```
  map will have changed. We're gonna freak 'em out.'' </p> <p> He leans
   rent was put up by ten quid which freaked everybody out and especially the
    The Exchange. <p> Mike and Sarah freaked everyone out when they appeared on
And she said of the huge cross: `It freaks everyone out." In an amazing
   the couple's bed, Patsy says: It freaks everyone out as the bed is where we
      On the other, there are these freaks going out in a canoe to harpoon the
    Ginola would merely succeed in `freaking Graham out," should doff their
  cleared. A friend said: `The bugs freaked her out. I think the house is
  secretary Elaine Wood, 47. <p> It freaked her out eventually. Going back
```

Sorting by the word to the left of the Key Word can show us other interesting combinations, such as the adjectives associated with a noun, the verbs that take a particular preposition, and so on. Sorting by the word to the left of the adverb *feelingly*, we discover that this adverb is most commonly used to modify verbs of speech.

Figure 21: Concordance for *feelingly*: sorted by the word to the left

```
  Standards Institution'', he added feelingly, ``improvements in riding
    <p> D Yes &hellip # Max agreed feelingly `It's high time you came out from
before she conked out, she appealed feelingly to the surgeon, saying, as she
    like Aldous Huxley, complained feelingly of the rich person's inability to
  looked around the lobby and ended feelingly - `woman?" <p> Brand acknowledged
```

4.7 Subcorpus distribution
In many corpora, data is held in various ***subcorpora***. At Cobuild, the subcorpora roughly approximate to linguistic notions of *genre* and *variety*. Frequency information is displayed for each subcorpus, so the word *defense* is seen to be more frequent in American subcorpora (npr, usnews, usbooks, usephem) than in British or Australian. Because the subcorpora are of varying sizes, the number of occurrences per million words is given as well as the actual frequencies (these figures are from the 418 million word corpus):

Figure 22: Subcorpus distribution for *defense*

```
Query is "defense"
16419 matching lines
Corpus         Total Number of       Average Number per
               Occurrences           Million Words

npr                7527                 84.6/million
usnews             2489                 61.5/million
usacad              963                 38.0/million
usspok              244                 30.1/million
usbooks            3772                 29.0/million
usephem             185                 13.2/million
newsci              180                  5.7/million
brbooks             342                  2.0/million
brmags              328                  1.9/million
indy                132                  1.1/million
bbc                  73                  1.0/million
today                49                  0.5/million
guard                50                  0.4/million
times                45                  0.4/million
brephem               5                  0.3/million
sunnow               19                  0.1/million
oznews               11                  0.1/million
econ                  5                  0.1/million
```

The figures for *defence* of course show exactly the opposite, with British (and Australian) subcorpora at the top and American subcorpora at the bottom of the list. It comes as a surprise to some people that *defense* is used at all in British texts, or that *defence* is used in American texts, but of course when British writers refer to American terms or proper nouns (e.g. the *defense* in American football, or the Secretary of State for *Defense*), or quote American sources directly, they have to use the American forms, and vice versa.

Similarly, subcorpus distribution figures show that *actually* is used more frequently in spoken data (both British and American) than in written data. For other words, we may notice other genre-specific qualities, for example that they occur most frequently in broadsheet newspapers rather than tabloids, in books rather than in newspapers, etc. A separate frequency list can also be generated for each subcorpus, so we can therefore look at which words are most important in a particular domain, genre, or text-type.

5: Collocation and units of meaning

5.1 What is collocation?

The term ***collocation*** was coined by Firth to describe one of his `Modes of Meaning'. Michael Halliday refers to it as one of the devices used to produce cohesion in a text. John Sinclair pioneered the investigation of collocation using quantitative techniques. Jeremy Clear at Cobuild helped to implement some of these techniques. Here are some useful quotes from the academic literature.

J.R.Firth:
(1957) You shall know a word by the company it keeps... I propose to bring forward as a technical term, meaning by `collocation', and to apply the test of `collocability'.

M.A.K.Halliday:
(1966) ... lexis seems to require the recognition merely of linear co-occurrence together with some measure of significant proximity, either a scale or at least a cut-off point. It is this syntagmatic relation which is referred to as `collocation'.
(1976) (Collocation, or collocational cohesion) ... is simply a cover term for the cohesion that results from the co-occurrence of lexical items that are in some way or other typically associated with one another, because they tend to occur in similar environments ... In a lexical analysis it is the lexical restriction which is under focus: the extent to which an item is specified by its collocational environment. ... It is the similarity of their collocational restriction which enables us to consider grouping lexical items into lexical

sets... The occurrence of an item in a collocational environment can only be discussed in terms of probability.

John Sinclair:
(1970) Collocations of very frequent words are positionally restricted ... Collocation which is positionally free ... will commonly be an indication of lexical patterning.
(1987) Collocation, as has been mentioned, illustrates the idiom principle. On some occasions, words appear to be chosen in pairs or groups and these are not necessarily adjacent... When A is node and B is collocate, I shall call this downward collocation - collocation of A with a less frequent word (B). When B is node and A is collocate, I shall call this upward collocation. ... Upward collocation of course is the weaker pattern in statistical terms, and the words tend to be elements of grammatical frames, or superordinates. Downward collocation by contrast gives us a semantic analysis of a word.
(1991) The basis of lexical patterning is the tendency of words to occur in the vicinity of each other to an extent that is not predicted by chance (Firth 1957). This tendency contributes substantially to the redundancy
of language, and so we can make an assumption that it is meaningful. Current work begs a lot of questions - what unit is to be counted, and what constitutes the optimal environment... Collocation is the occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening.

Jeremy Clear:
(1993) The study of collocation as a linguistic phenomenon has not found a central place in theoretical linguistics, perhaps because its proper province is the rather ill-defined area of linguistic patterning that is neither clearly syntactic nor clearly semantic. Moreover, the Firthian framework for studying collocation is dependent upon a model of linguistics which does not concern itself with the Chomskian competence/performance dichotomy and the consequent exclusive attention to competence. In the applied area of lexicography, however, collocation is widely acknowledged to be of central concern... Collocation, being a part of the study of lexis, badly needs computational assistance since the number of lexical items is so large... The algorithm is very simple: every instance of the keyword is located, and a frequency list is made of the words occurring within a defined span ... of the keyword. This raw list ... is then annotated with a "significance" statistic that is calculated from the frequency of the co-occurrence, the frequencies of the collocating word and the keyword, and the overall size of the corpus.

5.2 Intuition, Dictionaries, and Corpus evidence
Neither intuition nor traditional dictionaries (which are based on intuition) can give a good account of collocation. As an exercise to test this for yourself, make a list of the main collocates of the word *crisp* from your own intuition. Then look at the entry for *crisp* in a dictionary. Which collocates are shown in the dictionary? You may have to look at the entry quite carefully, as most dictionaries do not signal collocates explicitly. Now here are the main collocates of *crisp* from the 418 million word Bank of English corpus:

Figure 23: Main collocates of *crisp* (ordered by frequency)

```
Query is "crisp"
4490 matching lines
Main collocates (by frequency):
```

and	2095	white	295	by	142	air	91
the	1600	on	263	at	129	finish	91
a	1357	as	246	his	127	not	89
with	733	for	221	golden	119	salad	88
of	703	until	199	clean	113	i	82
in	512	are	177	quentin	111	clear	80
to	489	or	173	be	108	up	79
is	415	that	165	this	108	its	79
	377	but	160	they	104	brown	78
it	311	from	153	you	103	cotton	75
was	297	fresh	143	he	99	packets	75
s	296		143	dry	99	an	74

Note: The discrepancy between and (which one expects to be equal in frequency) is due to anomalous coding strategies used in the newspaper subcorpora.

Which of these words were in your list of collocates, and which were given in the entry for *crisp* in your dictionary? I looked at several EFL dictionaries and found examples with *biscuit, toast, apple, curls, snow,* and *lettuce* for the literal, physical meaning of the adjective; and *note, order, speech, answer* for the figurative meaning describing a brisk, clear and authoritative way of speaking.

Note that in the corpus collocate list above, many of the words are grammatical words (*and, the, a, with*, etc), which few of us would think of (in Firth's original terms, collocation with grammatical words is called ***colligation***). But even the content or vocabulary words are probably not the ones we immediately thought of: *white, fresh, golden, clean, dry, air, finish, clear, brown, cotton.* (We can exclude *packets*, which belongs to the sense of potato *crisps* in British English - *chips* in American English; and also *quentin*, which refers to the author Quentin Crisp).

The development of large language corpora and sophisticated, statistically-based software tools have revolutionised the study of collocation. The collocate list for *crisp* given above is based on simple frequency of co-occurrence within four words of the node word *crisp* and contained, as we noted, a high proportion of words that occur very frequently in the corpus anyway (i.e. the grammatical words), and obscure the lexical collocations that we are probably more interested in.

Now here is the list of the main collocates of *crisp* from the same corpus, but placed in order of statistical significance (by ***t-score***, which will be explained a little later):

Figure 24: Main collocates of *crisp* (ordered by t-score)

```
Query is "crisp"
4490 matching lines
Main collocates (by t-score):
```

and	2095	27.416162	clear	80	8.034738	packet	40	6.272029
with	733	18.458867	shirt	67	8.006265	wine	43	6.134600
white	295	16.493395	linen	64	7.961940	bread	39	6.048025
a	1357	15.212960	light	70	7.535075	walkers	36	5.973876
until	199	13.083092	bacon	57	7.485936	liam	36	5.930212
fresh	143	11.711397	fried	55	7.374645	potato	36	5.927794
golden	119	10.752056	leaves	56	7.161853	palate	35	5.900076
quentin	111	10.521029	serve	54	7.069166	minutes	51	5.858404
clean	113	10.401892	cool	53	7.017203	sheets	34	5.731743
dry	99	9.742974	fruit	52	6.956489	bright	36	5.673348
salad	88	9.316576	skin	50	6.662531	soft	36	5.643075
finish	91	9.305565	green	54	6.661075	shot	40	5.508462
air	91	8.689444	vegetables	46	6.651474	or	173	5.504197
packets	75	8.639786	blue	50	6.456253	shirts	31	5.439707
cotton	75	8.542123	pastry	42	6.445349	delicious	30	5.378553
brown	78	8.291928	lettuce	41	6.377184	flavour	29	5.269566

Note: The second column shows the frequency of collocation, the third column shows the t-score.

We see that many of the grammatical words have now been eliminated from the list, and many more lexical words are now included. (We must omit *walkers* and *liam*, as they are both proper names). We can also more easily classify the collocates into semantic sets: colour words, types of cloth and clothing, fried food, and so on. And we can even identify some genre-specific collocations (e.g. *crisp shot* is used in sports journalism).

Various statistical measures have been used in collocation software. ***T-score*** is a measure of the confidence with which we can assert that an association exists between two events. In the case of collocation, the association is between the occurrence of one word and the occurrence of another word in close proximity (4 words in the Cobuild software) to it. It measures the number of standard deviations of the observed result from the expected result. The t-score will be higher when the co-occurrence is observed many times,

because it increases our confidence that the co-occurrence is not a freak of chance. However, t-score also takes into account the fact that a word is very frequent in the corpus anyway, hence *the, of, in, to, is,* etc are reduced in significance. Other statistical measures that have been used are *mutual information, log likelihood, and chi-square.*

Collocations can vary according to the specific wordform and the particular wordclass. For example, most of the collocates of the verb *file* are different from the noun *file*. Among the collocates of the verb are the infinitive-marker *to*, the prepositions *for* and *against*, the typical objects of the verb (*suit, complaint, charges, returns, reports, application*), *divorce* and *bankruptcy*. Whereas the collocates of the noun include *single, tape, retrieve, lowbar, feature, members, computer, letter*, and *open*; and the prepositions *on* and *from*. Note, however, that some collocates are shared by both verb and noun: *sports, fact, rank*, etc.

Figure 25: Main collocates for *file* as a verb (ordered by t-score)

```
Query is "file/VB"
2597 matching lines
Main collocates by t-score:
```

to	1389	23.375830	bankruptcy	79	8.861320	a	626	6.604859
sports	174	13.056627	letter	76	8.302097	divorce	44	6.537913
your	180	11.346618	fact	79	8.178710	must	60	6.474620
for	391	10.507314	charges	64	7.778834	will	133	6.456115
suit	94	9.578478	returns	60	7.643071	lawsuit	41	6.381987
against	114	9.327017	reports	61	7.280646	rank	41	6.347419
complaint	86	9.230359	tax	55	6.868946	application	41	6.281510

Figure 26: Main collocates for *file* as a noun (ordered by t-score)

```
Query is "file/NOUN"
8157 matching lines
Main collocates by t-score:
```

rank	939	30.606547	file	166	12.753934	members	108	8.900363
fact	546	22.518989	sports	150	11.793198	computer	94	8.873270
on	1057	18.427234	tape	118	10.607328	letter	97	8.693117
a	2178	15.659773	from	512	10.326860	open	110	8.599849
and	2251	15.275303	retrieve	93	9.613476	name	97	8.211145
the	4532	14.390380	lowbar	90	9.476280	getting	92	8.042332
single	195	13.086087	feature	91	9.207207	case	98	7.827673

Note: *file* is itself a significant collocate of the noun *file*. Many words are self-collocating, because we often need to repeat words: e.g. H*e pulled out file after file... There's a shaping file, a fine file... Click on the file, go to the File menu...*

Lists of collocates can be useful in themselves, but positional information about collocates can also be very important.

Figure 27: Main collocates of *linen*, by position in relation to the node (ordered by t-score)

```
Query is "linen"
3547 matching lines
Collocation picture by t-score:

from       cotton     bed        NODE       and        and        pound
pound      with       white      NODE       suit       pound      and
washing    and        dirty      NODE       shirt      public     wide
white      white      table      NODE       sheets     cotton     jacket
shirt      of         fine       NODE       mix        with       cotton
wash       silk       cotton     NODE       trousers   silk       linen
cotton     a          monogramme NODE       jacket     towels     with
fabrics    in         and        NODE       cotton     or         from
covered    photo      irish      NODE       shop       linen      silk
linen      pound      crisp      NODE       in         dress      white
with       linen      cream      NODE       dress      jacket     wool
wearing    crisp      blue       NODE       napkins    mix        cm
pure       their      or         NODE       or         which      12
```

table	wool	beige	NODE	cupboard	breasted	are
range	your	french	NODE	cloth	blankets	top
wool	fine	natural	NODE	union	cm	120
de	jouy	household	NODE	suits	shirt	39
in	fabrics	starched	NODE	tablecloth	waistcoat	thomas
bed	starched	embroidere	NODE	tablecloth	lace	50
made	her	clean	NODE	closet	wool	fabric
or	dirty	british	NODE	company	trousers	mix
laid	lace	black	NODE	skirt	tables	suit
wear	washing	a	NODE	napkin	from	95

Note: NODE is *linen*. The amount of information contained in this display means that long words are often truncated (e.g. embroidere is obviously *embroidered*). This display, called ***picture*** in the Cobuild software,
should be read vertically, as it shows the most significant collocates (by t-score) for each position to left and right of *linen* in the corpus texts (only 3 words to left and right are shown here, but the software can show up to 6).

The software allows us to focus on the collocate *dirty* (the 3rd most significant collocate one word before *linen*), and investigate the collocates of the phrase *dirty linen*. This highlights the expression *to wash one's dirty linen in public*, and reveals the variation *to air one's dirty linen in public*.

5.3 Semantic Prosody

The study of collocation has been extended in recent years. John Sinclair first recognized the linguistic feature he termed ***semantic prosody*** in `Looking Up' in 1987, where he noticed that the phrasal verb *set in* was nearly always associated with `negatively evaluated' events (*winter, decline, the rot*, etc). Since then, other scholars (e.g. Louw, Stubbs, Hunston) have investigated this phenomenon further and have found more examples: *bent on X, symptomatic of X, utterly X, without feeling X*, and even common and `neutral-seeming' verbs like *happen* and *cause*, all tend to be followed by negative expressions. One or two words with positive prosodies (e.g. *provide*) have also come to light.

With *bent on*, there is a predominance of "negative" collocates to the right: *destroying, revenge, destruction, domination, undermining, breaking, killing, destabilising, suicide, sabotaging, wrecking*, etc. Even the superficially `positive-seeming' collocates (*creating, proving, securing, winning, establishing, expansion, achieving, protecting*) are usually found, on closer inspection of the concordances, either to form part of a larger negative phrase or to be used ironically. The left-hand collocates generally support the negative evaluation (*hell-bent, terrorists, evil*).

The collocational profile of *provide* clearly exhibits a strongly positive semantic prosody: *information, services, opportunity, support, evidence, care, protection, free, perfect, best, good, assistance, insight, money, training, advice, relief*, etc.

5.6 Near-synonyms

Collocation can help to distinguish between near-synonyms. One example used in the academic literature is *strong* and *powerful*: you say *a strong cup of tea*, but rarely *a powerful cup of tea*; you say *a powerful car*, but rarely *a strong car*. In the following corpus comparison, as both words are very frequent, only a small (but roughly equal) sample of data was selected for analysis.

Figure 28: Main collocates of *strong* (ordered by t-score)

Query is "strong"
77695 matching lines
How many required:1% (777 lines)
Main collocates by t-score:

a	316	10.021643	there	34	3.035478	demand	8	2.629020
very	48	6.028691	position	34	3.035478	had	36	2.600991
enough	27	4.797076	too	15	2.815440	growth	8	2.551039
is	103	4.439355	team	12	2.791803	feelings	7	2.517104
has	42	3.206254	winds	8	2.790760	wind	7	2.511602
and	189	3.172484	have	47	2.676129	views	7	2.505605

support	13	3.154091	with	61	2.642252	leadership	7	2.500142

Figure 29: Main collocates of *powerful* (ordered by t-score)

```
Query is "powerful"
34536 matching lines
How many required: 2% (691 lines)
Main collocates by t-score:
```

most	92	8.978441	are	46	3.191680	than	20	2.592300
a	253	8.198574	lobby	10	3.133461	such	14	2.583193
more	56	5.884276	force	10	2.831012	shot	8	2.547589
very	29	4.355999	political	11	2.766980	so	24	2.510582
and	185	4.095588	but	45	2.727406	as	49	2.504227
world	26	4.081618	rich	8	2.646696	position	8	2.499584
is	84	3.542556	nation	8	2.616617	american	10	2.446174

Several collocates are more closely linked with ***strong****: support, team, winds, demand, growth, feelings, wind, views, leadership*. Others are linked with ***powerful***: *world, lobby, force, political, rich, nation, shot, American*. Again, one can detect some genre-specific or domain-specific tendencies: ***strong*** goes with emotions (*support, feelings*), sporting allegiances (*team*), economics (*growth, demand*) and climate (*winds*), whereas ***powerful*** goes with global politics (*world, lobby, political, nation, American*), wealth (*rich*), and sporting actions (*shot*). Only *position* is common to both lists.

We can similarly investigate the main collocates of *big*, *great* and *large*, and see which collocates they share and which are exclusive to one of the words:
big: `a, too, enough, bang, companies, one, business, difference, banks, board, there, very`
great: `a, deal, of, britain, was, it, there, success, is, many, grandfather, one, fun, man`
large:`a, of, number, scale, numbers, very, amounts, part, enough, small, and, in`

The corpus evidence can be used to check language reference books. For example, the entry for *big - large - great* in the Cobuild Usage book says: " (1)*Big, large,* and *great* are used to talk about size. They can all be used in front of count nouns, but only *great* can be used in front of uncount nouns. (2) *Big, large,* and *great* can all be used to describe objects. *Big* is the word you usually use in conversation. *Large* is more formal. *Great* is used in stories to indicate that something is very impressive because of its size. (3) You use *large* or *great* to describe amounts. You do not use *big* ... (4) When you are describing feelings or reactions, you usually use *great*. When *surprise* is a count noun, you can use either *big* or *great* ... You do not use *large* to describe feelings or reactions. (5) When you are talking about qualities, you use *great*. You do not use *big* or *large* ... (6) When you are describing a problem or danger, you use *big* or *great*. You do not usually use *large* ... (7) *Great* is also used to say that a person or place is important or famous..."

6: Grammar

6.1 Part-Of-Speech tagging

We have already had a brief look at how grammar is obtainable from a corpus. Section 3 referred to lemmatization and part-of-speech tagging. Part-of-speech tagging is done by using a dictionary lookup system, by applying grammar rules, or by referring to probability statistics; or usually by a combination of these techniques.

We can display part-of-speech tags in combination with concordance lines. If we take 100 lines for the wordform *report*, and ask to see the tags, we find that 89 lines are for noun uses (NN in the leftmost column) and only 11 for verb uses (VB in the leftmost column). This is a finding which we could not have arrived at by intuition. The proportion of noun-to-verb use for *report* is fairly stable. If we double the number of lines examined, we find that out of 200 lines for *report*, 179 are for noun uses, 21 for verb uses.

Figure 30: Concordance lines for *report* with part-of-speech tags displayed

```
Query is "report"
115414 matching lines
```

```
NN ghts. <p> The company's actuarial report must be made available on demand.
NN e resurrection of ecu bonds - The report of their death was exaggerated, but
NN         Deepak Tripathi sent this report from Kabul: </p> <h> REEL UTLEY
NN l Kev sets it up; Football; Match report </hl> <dt> 24 October 1999 </dt>
NN r high jump and shot. The 60-page report will be distributed to 1,700 clubs

VB The Daily Mirror will continue to report the news plain and unvarnished, as
VB  he noted. The company expects to report fiscal fourth-quarter profit of
VB man is like him, does he? I shall report this matter to Paul, then he'll see
VB nd a tabloid newspaper. They also report new calls for counseling from women
VB mits, had made up her mind not to report it. Three days after the attack,
```

Most current tagging programs claim success rates very close to 100%. The occasional error presents no problem when the results are being viewed by trained users, such as lexicographers or linguists, but may need to be manually corrected or omitted before presentation to students or when used as input to other computational processes.

Some corpus systems allow us to actually see the part-of-speech tagged text. It is not very reader-friendly, but can provide useful insights to specialists such as grammarians, corpus designers, programmers, etc. The output can be displayed horizontally or vertically, depending on the purpose:

Figure 31: Displays of part-of-speech tagged text

```
N_people V_sit #_,  LNK_or N_walk P_in N2_rings #_,

PPSS    WE
ABN     ALL
RB      now
VB      live
ININ    in
WDT     what
BEZ     is
RB      loftily
VBN     called
AT      a
JJ      global
NN      village
```

6.2 Parsing

A step further than part-of-speech tagging is parsing. In this process, each lexical item is additionally given a
clause function. Cobuild had a 200 million word corpus tagged and parsed by the University of Helsinki. In the output, for the clause *The kidnapper had left a trail*, *the* is not only labelled as a determiner (DET), and classified in more detail (central determiner, article, singular or plural in number), but the direction of the head noun in the nominal group is shown (@DN>). Similarly, *kidnapper* is identified as the subject (@SUBJ) of the clause, *had* as a finite auxiliary verb (@+FAUXV), *left* as a finite main verb (@-FMAINV), *trail*as the object (@OBJ) of the clause, and so on.

6.3 Grammar Changes

We are less sensitive to changes in grammar than in lexis. We all notice new words, but few of us notice new grammar patterns. Also, grammar changes take longer to establish themselves in the language. One or two minor changes have been noticed over the past twenty years or so: the pattern *persuade someone to do something* has been joined by *persuade someone into doing something*, no doubt following the pattern of other verbs with a similar meaning, such as *talk, cajole, tease, coax*, and so on:

```
     a limited immunity in a bid to persuade her into returning Abbie. <p> Ken
        called for more spending to persuade brighter scientists into becoming
as a child, he had the character to persuade his mother into providing, was
opinions on whether I should try to persuade her into doing it? Although giving
```

Similarly, the pattern *X (a larger unit) comprises (several) Ys (smaller units)* has been overtaken by *X comprises of (several) Ys*, probably echoing the pattern of *consist*, a commoner verb with a similar meaning:

```
Perpetual's Peps. The discount will comprise of 3 per cent rebated commission
  to 24 years a Silver Award. These comprise of specially designed ties or
 have volume control. A system must comprise of at least one Bellpush and one
  there. He said it would initially comprise of fifteen teams, each of between
```

7: Exploiting the corpus for teaching

All of the corpus displays and analyses discussed above can be used in some form or other by teachers to prepare teaching materials, to generate classroom activities and exercises, and to check doubts and queries that can never be adequately answered by traditional language reference sources. Information about the Bank of English corpus can be found at http://www.cobuild.collins.co.uk/, as well as some ideas for ways of using the corpus (e.g. the "Wordwatch" articles).

Increasingly, students are accessing corpora directly, and this allows a different mode of learning to take place (called *data-driven learning* by Tim Johns: see the Bibliography for his website, where many further examples of this technique are given). Students learn better by seeing many examples and recognising patterns for themselves, rather than by memorizing a rule given by a teacher or a book. And whereas normally students may come across an example of a particular word, structure, or pattern every few months (if they are lucky), corpus access allows them to see many examples of the same linguistic item or feature on a computer screen at the touch of the keyboard.

Large corpora grant students concentrated exposure to authentic language which even native-speakers might take a lifetime to experience. Corpora enhance linguistic awareness by presenting all the features of language simultaneously: every concordance line reinforces spelling, inflections, collocations, phraseology, grammatical patterns and natural contexts.

Corpora can be used for simple tasks like checking the correct spelling of a word. In a large corpus, the most frequent spelling is very unlikely to be wrong. Frequency can be used to ensure that only the commonest and most typical items are studied, learned, and practised. Concordance lines can be used to highlight the typical behaviour of words, wordclasses, and phrases. They can also be used in simple cloze tests: delete the keyword, and ask the students to guess the word from its contexts. Online exercises can include sorting a set of concordance lines into groups according to meaning. Contrastive exercises can focus attention on differences between near-synonyms, but also to raise awareness of differences in register, domain, genre and text-type.

With the advent of corpora, the teacher's role is gradually shifting towards facilitating the student's own learning process. And as students increasingly work on computers, their own essays can be analysed with corpus tools and compared with native-speaker corpora. In rapidly changing areas of language such as scientific and technological discourse, traditional resources quickly grow out-of-date. Teachers can easily assemble small corpora in specialized fields from documents available on the Internet, which will be more reliable both for terminology and current writing styles.

BIBLIOGRAPHY

Clear, J. (1993) From Firth Principles: Computational Tools for the Study of Collocation, in Baker et al (eds.) *Text and Technology*, John Benjamins, Philadelphia/Amsterdam

COBUILD: http://www.cobuild.collins.co.uk

Firth, J.R. (1957) Modes of Meaning, in F.R. Palmer (ed) *Papers in Linguistics 1934-51*, Oxford University Press, London

Halliday, M.A.K. (1966) Lexis as a linguistic level, in C.E. Bazell, J.C. Catford, M.A.K. Halliday, R.H. Robins (eds) *In Memory of J.R. Firth*, Longman, London

Halliday, M.A.K. (1976) *System and Function in Language*, ed. G. Kress, Oxford University Press, London

Halliday, M.A.K. and Martin, J.R. (1993) *Writing Science: Literacy and Discursive Power*, Falmer Press, London

Hanks, P.W. (1988) *How common is common*, Collins, Glasgow

Johns, T. *Microconcord*: http://web.bham.ac.uk/johnstf/

Nelson, G. (1991) Cobuild seminar, Birmingham University

Scott, M. *Wordsmith Tools*: http://www.lexically.net/
Sinclair, J.M., Jones, S. and Daley R. (1970) *English Lexical Studies, Report to OSTI on Project C/LP/08*
Sinclair, J.M. (ed) (1987) *Looking Up*, Collins ELT, London
Sinclair, J.M. (ed) (1990) *Collins Cobuild English Grammar*, Collins, London
Sinclair, J.M. (1991) *Corpus, Concordance, Collocation*, OUP, Oxford
Stubbs, M. (1995) Collocations and semantic profiles: On the cause of the trouble with quantitative studies, *Functions of Language*, 2, 1: 1-33
Stubbs, M. (1996) *Text and Corpus Analysis*, Blackwell, Oxford
Summers, D. (1993) Longman/Lancaster English Language Corpus – Criteria and Design, *International Journal of Lexicography 6/3*: 181-208
Svensen, B. (1993) *Practical Lexicography*, OUP, Oxford
Zipf, G. K. (1949) Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology, Addison-Wesley, Cambridge