The 2nd ILASH Half-Day Workshop on
"Computational Language Resources"

University of Sheffield, Feb 8th 2002

OHP00 - Title

The Bank of English past, present, and future: corpus size, composition, annotation, and software

Ramesh Krishnamurthy

Honorary Research Fellow: Department of English, University of Birmingham
Consultant: Cobuild and the Bank of English corpus, Collins Dictionaries

ramesh@cobuild.collins.co.uk

OHP00 - NOTES:

In 1980, Professor John Sinclair initiated the COBUILD project (jointly funded by Collins Dictionaries and the University of Birmingham). In 1983, the BCET corpus of 7.3m words started to be used for dictionary compilation. Since 1987, when the corpus had reached around 20m words, three editions of the Cobuild dictionary and numerous other publications have appeared. The Bank of English corpus was launched in 1991, currently contains 450m words, and is planned to pass 500m words during 2002. This talk will look at some of the issues concerning corpus size, composition, annotation, and software in terms of past experience, current facilities, and possible future developments.

From the outset, Cobuild's aims have been:
a) to collect a large corpus of English language and analyse it
b) to publish the analyses for students and teachers of English

Why bother with a corpus?
Linguistics based on intuition and introspection alone has proved to be inadequate. "Users of a language are not necessarily accurate reporters of usage, even their own, Most of our skill in using language is unconscious, and therefore difficult to recall, though easy to recognize" (John Sinclair, *Introduction to Collins Cobuild English Language Dictionary*, 1987, p.xviii).

Why a large corpus?
"When COBUILD published a dictionary based on a corpus of 20 million words it was clear to all who worked on the project that our description of English lexis and grammar could be even better if more corpus data were available. The view in COBUILD has been that, where a corpus is concerned, the bigger the better. Bigger can be measured both by raw word count and by the breadth of coverage. The balancing of corpora - an issue within corpus linguistics that has recently started to attract some attention - is really a matter of establishing appropriate priority between increasing the word count and increasing the breadth of coverage. Of course, these

two measurements are not independent and COBUILD has managed to keep the corpus size rising while also increasing the range of text types sampled. No fine and delicate tuning of the relative proportions of varying text types is attempted, because we see no evidence that our samples can be significantly improved in that way, given that even a corpus of hundreds of millions of words is of a pitifully small size in comparison to the amount of English language being generated daily. We do, however, see firm evidence that the addition of large tranches of new corpus data from new sources enhances every aspect of our analysis." (Clear, Fox et al, *COBUILD: the state of the art*, IJCL 1:2, 1996)

OHP01 – Corpus Linguistics at Birmingham University

- **1960s : Sinclair: 135,000 words of spoken English**

Sinclair, Daley, Jones: *English Lexical Studies (The OSTI Report)* 1970

- **1970s - 1980s: academic research**

35,000 words: *Classroom Discourse* : Sinclair & Coulthard 1975
1m words: *Applied Science* : Roe 1977; Phillips 1983; Yang 1986
750,000 words: *Economics* : Tadros 1981
***Computers in Language Learning* : John Higgins & Tim Johns 1984**
(http://web.bham.ac.uk/johnstf/)

- **1980 onwards: Cobuild**

ed. Sinclair : *Looking Up* (Collins ELT) 1987
Clear, Fox, et al : *COBUILD: The State of the Art* (IJCL 1:2) 1996

- **1990s:**

German, Italian
***Forensic Linguistics* (http://www.builder.bham.acv.uk/forensiclinguistics/)**

- **2001: Centre for Corpus Linguistics**

OHP01 - NOTES:

1. There is a continuous history of corpus research at Birmingham since the 1960s. Sinclair's work on spoken data was started in Edinburgh and completed in Birmingham, and represented the earliest work on a spoken corpus. The focus of the research was on collocation. Sinclair, Daley, Jones *English Lexical Studies (The OSTI Report)* 1970 is currently being edited with the intention of re-publishing it soon. John Sinclair retired in 2000, but continues his research and teaching work at the Tuscan Word Centre in Italy.
2. Tim Johns is also responsible for concordancing software (Microconcord) and for pioneering "Data-Driven Learning" (DDL). He retired in 2001, but continues his research. His website contains a wealth of practical information and examples of research exercises and classroom activities.
3. The German corpus activities are mainly conducted by Bill Dodd.
4. Research into Forensic Linguistics was initiated by Malcolm Coulthard, Sue Blackwell etc.
5. The Centre for Corpus Linguistics is headed by Professor Wolfgang Teubert.
6. In 2002, the Dictionary Research Centre (which has moved from Exeter University)

will commence its activities in Birmingham.

OHP02 – COBUILD 1980-87

- ***Corpus design: selection, permissions, acquisition***

Trawl of universities for existing data
British Council libraries borrowing lists
For EFL: no drama, poetry, children
Balance Aims:
70% British 20% American, 10% Other
75% Male, 25% Female
75% Written, 25% Spoken

- ***Corpus data: inputs***

Optical scanning (Kurzweil Data Entry Machine) of books
Keyboarding of newspapers, magazines, and ephemera
Transcription of spoken data

- ***Corpus data: outputs***

Corpus Frequency Lists (alphabetical/frequency order; hardcopy)

Concordances (right-sorted, fixed length contexts:
microfiche and hardcopy):
a) 6 x 1m written batches; 1 x 1.3m spoken batch;
b) Merged Concordances (Main Corpus: 7.3m)

Additional Concordances (for items below freq 49 only;
from Reserve Corpus: 11m written; microfiche only)

1m Sub-corpus (online; left and right-sorting, longer contexts)

- ***Dictionary design, editorial policy, trials***

OHP02 – NOTES

1. The British Council libraries' borrowing lists were consulted so that the corpus would reflect the material being consumed by advanced students of English all around the world.
2. As the corpus was initially intended for EFL lexicography, no drama, poetry, or children's language was included. These are highly marked genres that would not serve as a practical model for students or teachers of English.
3. Scanning, keyboarding, and transcription are listed in order of cost (e.g. in the early 1990s, they cost £1000, £3000, and £25,000 per million words respectively). Hence spoken data is under-represented and will always be so, until a cheaper method of acquiring it is found.
4. Dictionary design, etc will not be considered today, as the focus is on the corpus methodology, not the lexicographic practice.

OHP03 – Brown corpus design model

SPOKEN 300	**DIALOGUE 180**	**PRIVATE 100**	**direct, distance**
		PUBLIC 80	**class lesson, broadcast discussion, broadcast interview, parliamentary debate, legal cross-examination, business transaction**
	MONOLOGUE 120	**UNSCRIPTED 70**	**spontaneous commentary, unscripted speech, demonstrations, legal presentation**
		SCRIPTED 50	**broadcast news, broadcast stories, broadcast talks, speeches)not broadcast)**
WRITTEN 200	**NON-PRINTED 50**	**NON-PROFESSIONAL 20**	**student untimed essay, student exam essay**
		CORRESPONDENCE 30	**social letters, business letters**
	PRINTED 150	**INFORMATIONAL (LEARNED) 40**	**humanities, social sciences, natural sciences, technology**
		INFORMATIONAL (POPULAR) 40	**humanities, social sciences, natural sciences, technology**
		INFORMATIONAL (REPORTAGE) 20	**press news reports**
		INSTRUCTIONAL 20	**administrative/regulatory, skills/hobbies**
		PERSUASIVE 10	**press editorials**
		CREATIVE 20	**novels/stories**

OHP03 – NOTES

1. The Brown corpus design model (of the 1960s) continues to influence corpus creators even today, e.g. the ICE corpus (London), and FLOB and FROWN corpora (Freiburg).
2. It is basically an "a priori model", which assumes that we know what language consists of. The proportions specified are arbitrary, but give a pseudo-scientific veneer to the model.
3. In practice, the model is inflexible and non-pragmatic, e.g. no allowance is made for problems of cost (e.g. spoken), availability, difficulty of obtaining permissions (e.g. business letters), new data-types (e.g. email), etc. The model may work for small, one-off, static corpora, but is not very suitable for large, dynamic corpora.
3. The BNC and the Longman Corpus have amended the model by adding various components:
a) a "random" selection from a catalogue of written texts
b) a demographic sampling of participants in the spoken data

OHP04 – Bank of English – Corpus Building Stategies

- **annual updates**
- **increase overall size**
- **increase sources/text-types**
- **replace older material with newer**
- **adjust disproportionate sizes of subcorpora**

BANK OF ENGLISH: INCREASE CORPUS SIZE

YEAR	million words (tokens)
1987	18
1993	121
1994	167
1995	211
1996	323
2000	418
2001	448

OHP04 – NOTES

1. Cobuild is perhaps an example of "a posteriori" design, more suitable for a large, dynamic corpus: we look back periodically at what we have (and what people have asked for, what has become available, gaps to be filled, etc), then try to adjust accordingly in our subsequent data acquisition.
2. Our main aims are:
a) annual updates: this enables us to keep the data up-to-date, which is important for lexicography.
b) increase overall size: as Collins native-speaker and bilingual dictionaries also increasingly make use of the corpus, we need information about a lot more words.
c) increase sources/text-types: e.g. to increase the varieties of English included, and include more data types.
d) replace older material with newer: another aspect of keeping the data up-to-date (e.g. replacing 1990 data with 2001 data).
e) adjust disproportionate sizes of subcorpora: e.g. subdivide larger subcorpora (e.g. British Books into Fiction and Non-Fiction; British Magazines into General Magazines and Special Interest Magazines); and focus on increasing smaller subcorpora.

BANK OF ENGLISH: INCREASING CORPUS SIZE

1. This has also been influenced by projects in the past: e.g. the BBC dictionary required us to acquire and use BBC World Service and NPR radio broadcast data.
2. During 1987-96, because of changes in data input methods and the large increase in volume, data acquisition became increasingly out-of-house:
a) news data was acquired mainly on magnetic tape, requiring only a little processing effort (some character and file format conversion by an out-of-house agency): Times, BBC World Service, National Public Radio (Washington, USA), Today , Economist.

b) spoken data was transcribed by an out-of-house agency, who also started contributing corpus data acquired from their other clients.
c) magazines and ephemera were keyboarded by an out-of-house agency.
d) a large number of books were acquired on typesetter's magnetic tapes, and converted by an out-of-house agency.
e) however, the book data still needed considerable manual effort in-house, so we reverted to optically scanning books in-house using MACs and OmniPage.
3. The 2001 corpus has just been released. Details will be given later.

OHP05 – Bank of English: Corpus Composition

1993	1994	1995	1996	2000
American Books				**American Academic Textbooks**
British Books				
American Radio				
BBC World Service				
British Ephemera			**American Ephemera**	
British Magazines				
British Spoken				**American Spoken**
Economist	**New Scientist**			
Independent	**Guardian**	**Australian Newspapers**		
Times				
Today				**Sun and News of the World**
Wall St Journal			**American Newspapers**	

OHP05 - NOTES

BANK OF ENGLISH - INCREASING SOURCES/TEXT-TYPES

1. We work on the principle that the data in any subcorpus has to have some degree of homogeneity.
2. Subcorpora are kept separate for purposes of comparison in lexicography and research:
e.g. American vs British (Books, Radio, Ephemera, WSJ vs Economist),
e.g. "Genre" (Books, Radio, Newspaper, Ephemera, Spoken)
e.g. Spoken vs Written
e.g. Broadsheet vs Tabloid newspapers

3. The gap in corpus growth from 1996-2000 was due to internal HarperCollins restructuring.
4. In the future, we might: create a British Academic Textbooks subcorpus (to match the American Textbooks); split British and American Books (which are very large) into Fiction and Non-Fiction subcorpora; split British Magazines into General and Specialist subcorpora; get American Magazines (to match British magazines); get more technical journals (e.g. Economist, New Scientist – add journals from Computing, Medicine, etc); get national newspapers from more countries.

OHP06 – Bank of English – Replace Older Material, Adjust Proportions

SUBCORPUS	1993			2001		
American Books	**16m**	**13.30%**	**1985 >**	**32.44m**	**7.23%**	**1990 >**
American Radio (NPR)	**10m**	**8.33%**	**1990-1**	**22.23m**	**4.96%**	**1990-3**
BBC World Service	**20m**	**16.66%**	**1990-1**	**18.60m**	**4.15%**	**1990-1**
British Books	**31m**	**25.83%**	**1985 >**	**43.37m**	**9.67%**	**1990 >**
British Ephemera	**1m**	**0.83%**	**1991-2**	**4.64m**	**1.03%**	**1991-6**
British Magazines	**5m**	**4.16%**	**1992**	**44.15m**	**9.84%**	**1992-00**
British Spoken	**4m**	**3.33%**	**1991-2**	**20.08m**	**4.48%**	**1991-6**
Economist	**3m**	**2.50%**	**1991**	**15.72m**	**3.50%**	**1991-9**
Independent	**5m**	**4.16%**	**1990**	**28.08m**	**6.26%**	**1995-9**
Times	**10m**	**8.33%**	**1992**	**51.88m**	**11.57%**	**1997-01**
Today	**10m**	**8.33%**	**1991-3**	**-------**	**--------**	**--------**
Wall Street Journal	**6m**	**5.00%**	**1989**	**-------**	**-------**	**--------**
Guardian	**-----**	**----------**	**---------**	**32.27m**	**7.20%**	**1995-9**
New Scientist	**-----**	**----------**	**---------**	**7.89m**	**1.76%**	**1992-9**
Australian Newspapers	**-----**	**----------**	**---------**	**34.94m**	**7.79%**	**1995-9**
American Ephemera	**-----**	**----------**	**---------**	**3.51m**	**0.78%**	**1995-6**
American Newspapers	**-----**	**----------**	**---------**	**10.00m**	**2.23%**	**1994-6**
Sun and News of the World	**-----**	**----------**	**---------**	**44.76m**	**9.98%**	**1997-01**
American Academic Textbooks	**-----**	**----------**	**---------**	**6.34m**	**1.41%**	**1990-6**
American Spoken	**-----**	**----------**	**---------**	**2.02m**	**0.45%**	**1994-7**
Strathy Canadian Corpus	**-----**	**----------**	**---------**	**15.92m**	**3.55%**	**1980-00**
Wolverhampton Business Corpus	**-----**	**----------**	**---------**	**9.65m**	**2.15%**	**1999-00**

OHP06 – NOTES

1. It is clear that we have doubled the number of subcorpora between 1993 and 2001.
2. The new subcorpora obviously contain more recent data.
3. The tabloid newspaper Today ceased publication in about 1995, so we have replaced it with the tabloid Sun and News of the World.
4. The ageing (1989) data from the Wall Street Journal has been replaced with more up-to-date material from several American Newspapers. We may recreate the Wall Street Journal subcorpus with new material later, or find an alternative source of American economic journalism.
5. Older books (British and American, 1985-1990) have been replaced with newer ones.
6. Newer data has been added to several subcorpora: American Radio, British Ephemera, British Magazines, British Spoken, Economist, Independent and Times.

7. Older data for Independent and Times (Independent 1990, 20 issues; Times 1995 and Times 1996 issues) has been replaced with recent material.
8. Altogether, for the 2001 corpus, 40m words were removed and 70m words were added (307 texts = c. 15% new material).
9. The Strathy corpus of Canadian English (various sources: books, academic papers, magazines, newspapers, and spoken) was acquired from the Strathy Language Unit, Queen's University, Kingston, Ontario; mostly 1980s and 1990s (NB there is one novel from 1937). Inclusion in the BoE will make their corpus more widely available, and we are also allowing Strathy researchers access to the rest of the BoE to do more detailed analyses of British/Canadian linguistic differences.
10. The Wolverhampton Business Corpus: Business English (23 websites from 6 "countries"; the first web-sourced corpus in the ELRA catalogue) was obtained from ELRA and is the first publicly distributed web-sourced corpus.

OHP07 – Bank of English: Type Distribution, High Frequency Types, Low Frequency Types

Bank of English: Type Distribution

YEAR	CORPUS	TOKENS	TYPES	HAPAXES	NON-HAPAXES
1993	121m	120,362,928	475,633	213,684	261,949
1994	167m		574,587		
1995	211m		638,901		
1996	323m		812,467		
2000	418m	418,449,873	938,914	438,647	500,266

High Frequency Types

CORPUS	18m	418m
the	1,081,654	22,849,031
of	535,391	10,551,630
and	511,333	9,787,093
to	479,191	10,429,009
a	419,798	9,279,905
in	334,183	7,518,069
that	215,332	4175495
s		4072762
is		3900784
it	198,578	3771509
for		3690466
i	197,055	3216005
was	194,286	3092967

Low Frequency Types

CORPUS	18m	121m	323m
No of types with n tokens			
1-9	**203,490**	**376,307**	**784,300**
10-99	**33,058**	**64,594**	**80,930**
100-999	**8,796**	**26,154**	**35,930**
1000-9999	**1,550**	**7,406**	**12,930**
10000-99999	**156**	**1,352**	**2,930**
100000-999999	**18**		**320**
1000000>	**1**		

OHP07 – NOTES

1. This OHP shows in detail the value of increasing the corpus size.
2. As is evident, the proportion of hapaxes (*hapax legomenon*, i.e. one-off occurrences of wordforms) remains constant whatever the size of the corpus, at just under 50%. Hapaxes are obviously not adequate information for creating dictionary entries, for which many examples are needed.
3. At the other end of the scale, the high frequency types (a *type* is a wordform, a *token* is one instance or individual example of that wordform in the corpus) do not change significantly whatever the size of the corpus. Some minor changes in rank do occur (e.g. *to* from 4th to 3rd; *s*, *is*, and *for* do not appear in the 18m list above, but no doubt occur slightly further down).
4. There is a great increase in the number of low frequency types (e.g. the number of types with a frequency of 100-999 increases from 8796 in 18m words to 35930 in the 323m words corpus). This is very important evidence, especially for native-speaker and bilingual dictionaries.

OHP08: Bank of English: rank and frequency of selected types

CORPUS	18m		418m	
	RANK	**FREQ**	**RANK**	**FREQ**
been	**48**	**48,068**	**47**	**1,019,904**
people	**75**	**26,057**	**72**	**610,679**
how	**94**	**20,906**	**104**	**393,586**
going	**129**	**14,924**	**147**	**288,607**
away	**150**	**12,168**	**225**	**185,260**
house	**176**	**9,890**	**206**	**198,592**
widely	**2,500**	**660**	**2,486**	**17,804**
prey	**5,000**	**280**	**9,211**	**3,185**
fulfilment	**10,000**	**107**	**15,122**	**1,506**
balloon	**15,000**	**58**	**9,011**	**3,298**
compromises	**20,000**	**37**	**16,395**	**1,327**
scenic	**25,000**	**26**	**15,651**	**1,429**
fungal	**40,000**	**11**	**25,633**	**628**
peyote	**70,000**	**4**	**58,153**	**129**

OHP08 – NOTES

1. Taking a selection of types from the 18m corpus frequency list, I have compared their rank and frequency with the 418m corpus.
2. Some items change very little (*been* rises from 48th to 47th, *people* rises from 75th to 72nd).
3. Others change much more: *widely* rises by 14 positions, and *how* drops by 10 positions, *going* by 18, *house* by 30, *away* by 75; *prey* drops by 4211 positions, and *fulfilment* by 5122; but *balloon* rises by about 6000 positions, *scenic* by nearly 10,000 positions, *peyote* by 12,000 and *fungal* by about 15,000 positions.
4. Of course, larger corpora will always have more actual examples (*tokens*) of most words: this is certainly true of all the items in the selected list.

OHP09: Bank of English: New Technology – New Words

	Number of corpus occurrences	
	1985	1995
camcorder	**0**	**1214**
cyborg	**2**	**31**
email	**0**	**39**
gopher	**2**	**35**
helipad	**0**	**27**
hypertext	**0**	**13**
imaging	**7**	**463**
keyhole surgery	**0**	**30**
laptop	**0**	**184**
microsurgery	**0**	**50**
mobile phone	**0**	**455**
palmcorder	**0**	**86**
palmtop	**0**	**25**
satellite dish	**0**	**236**
smart card	**0**	**68**
teleworker	**0**	**46**
videophone	**0**	**144**
virtual reality	**0**	**458**
videoconferencing	**2**	**38**

OHP09 – NOTES

1. This OHP shows the influx of new words in English in just one domain (technology) in one decade.
2. This process is obviously a continuous one, and provides lexicographers with problems of which new words to include in a new edition of a dictionary, and which words to omit.
3. For example, included for the first time in the 3rd edition of the Cobuild dictionary (2000) were: *affinity card, air ambulance, animal testing, applet, assistant referee, bad hair day, cooktop, daylong, docusoap, magic bullet, set-top box, 24-7,* and *WAP*.

OHP10 – Bank of English – Lemmatized word frequency list

the	**DT**	**17845179**	
be	**V**	**11989968**	
	VB	**be**	**1702992**
	VBD	**were**	**828676**
	VBD	**weren't**	**14366**
	VBDZ	**was**	**2423790**
	VBG	**being**	**274115**
	VBM	**am**	**68169**
	VBM	**'m**	**207711**
	VBN	**been**	**801065**
	VBR	**are**	**1407720**
	VBR	**aren't**	**20257**
	VBR	**art**	**80**
	VBR	**'re**	**246235**
	VBZ	**is**	**2970439**
	VBZ	**isn't**	**50691**
	VBZ	**'s**	**910625**
	VBZ	**was**	**13**
	VBZ	**wasn't**	**55201**
of	**IN**	**8321789**	
and	**CC**	**7582146**	
a	**DT**	**7060847**	
in	**IN**	**5826066**	
to	**TO**	**5132283**	

OHP10 – NOTES

1. I will not focus on corpus annotation, as Cobuild's stance on annotation is broadly negative.
2. The lemmatised list is just a reminder that when we look at type frequency lists (as in OHP07), we may sometimes be misled into making simplistic statements such as "The most frequent words in English are *the, of, and*, etc" when in fact the lemma *be* (consisting of 18 types) constitutes the second most frequent *lemma* (roughly equivalent to the lexicographic concept of *headword*) in English.
3. "Every wordform in the current corpus has been assigned a grammar tag such as JJ for adjective or VBG for the -ing form of a verb. The corpus retrieval software allows searches for structures consisting partly of lexical items and partly of tags. But the tags are only used as a heuristic tool at the beginning of grammatical analysis; they by no means constitute a grammatical description." (Clear, Fox et al, *COBUILD: the state of the art*, IJCL 1:2, 1996)
4. I append two more passages here to explain Cobuild's position on corpus annotation.

a) Error correction

"As the corpus grew we abandoned the attempt to do comprehensive manual checking and editing of the corpus data as it was acquired. The rate of acquisition of new text was growing much faster than our manpower resources. We came to the realization

that the value of new corpus data far outweighed the penalty of some mis-transcribed words, incorrect markup codes, lacunae, and other noise inherent in the data gathering. We decided that incoming data should be subject to a certain amount of automatic checking and correcting followed by manual editing where that was considered absolutely essential. Within the academic field of humanities scholarship precision, accuracy and exhaustive manual checking have become revered as icons. But the much greater potential for analysis and understanding that we gain from the 250 million words in the Bank of English compared with the one million word corpora of the Brown and LOB era reassures us that our strategy is justifiable." (Clear, Fox et al, *COBUILD: the state of the art*, IJCL 1:2, 1996)

b) Corpus markup

"Following from this principle is the decision to adopt a markup scheme for the corpus which is light and loose. This was an uncontroversial approach in the 1980s, before the advent of SGML, when any sort of markup was considered acceptable. Today there is a powerful movement towards the adoption of internationally agreed SGML-based markup schemes and the COBUILD approach seems quaintly unfashionable. However, we persevere with the use of markup codes which encode only the most obvious features of the surface representation of text, such as text divisions, headings and titles in writing and speaker change, long pauses, and inaudible sections in speech. This is what we mean by light markup. We also allow texts in the corpus to vary with respect to the features which are encoded, dependent on the ease or otherwise with which such features can be reliably and automatically detected. As the range of material included in the corpus extends, so too does the problem of consistent markup. Thus, we have found it simply not worth the manual effort of imposing markup standards upon texts which seem to us to be resolutely non-standard. New markup codes have been added as the need arises, other codes dropped or rationalized and over the years the encoding has stabilised to a great extent. This is what we mean by loose markup."

"The low priority which COBUILD assigns to corpus markup follows from and is consonant with our sceptical view of the value of database structures. When a corpus of raw text is marked up with a dense, standard encoding the resulting entity can be regarded as a database, having a structure which, by virtue of the encoding, is explicit and mechanically recoverable. This seems to be regarded almost universally as beneficial for corpus linguistics, while at COBUILD we are unconvinced of the validity of such structures. A markup scheme which requires `sentences' to be delimited seems dangerous, since it is surely clear to anyone who has worked with a large corpus that the received notion of a sentence is deeply suspect. COBUILD's approach to corpus encoding is paralleled in our corpus analysis and lexicography. Received categories and concepts are resisted and tested very hard against the mass of data. Categories that prove their worth in the business of analysis are retained for the purpose of description, but are always subject to revision when necessary or appropriate. The corpus analysis software likewise errs on the side of caution in processing symbols rather than wordforms from the corpus. Collocations are calculated without prior lemmatization of wordforms, for example, which has served to alert lexicographers to the dangers of assuming that only lemmas are interesting. This myopic concern for the lemma is, unfortunately, deeply rooted in the professional practice of lexicography, so bound up is it with the problems of

accessing words from an alphabetically arranged printed book. COBUILD's approach to corpus analysis has done a great deal to open up questions about the status of the lemma, and many more such established and apparently natural concepts."
(Clear, Fox et al, *COBUILD: the state of the art*, IJCL 1:2, 1996)

OHP11 – Collocation

the	the	the	NODE	to	the	the
to	a	local	NODE	of	a	to
by	and	of	NODE	the	for	in
with	to	health	NODE	and	that	and
a	of	his	NODE	on	to	a
of	with	an	NODE	that	was	is
that	financial	services	NODE	for	is	of
it	s	its	NODE	in	and	that
and	has	with	NODE	s	they	on
on	have	their	NODE	<p>	from	s
as	his	education	NODE	said	not	but
for	national	palestinia	NODE	has	he	have
but	local	no	NODE	had	s	was
was	on	and	NODE	over	said	be
in	their	own	NODE	was	are	they
no	british	police	NODE	as	been	had
is	police	s	NODE	is	who	all
they	<p>	in	NODE	he	had	he
which	in	complaints	NODE	which	their	asia
president	our	monetary	NODE	would	most	monitori
economic	power	moral	NODE	it	govett	<p>
could	by	that	NODE	but	i	i
director	from	to	NODE	when	it	now

OHP11 – NOTES

1. The Cobuild corpus retrieval software (called "lookup") incorporates collocation tools.
2. Looking at a span of 3 to 6 words either side of the "node" word (in this case the node is the word *authority*), collocates can be displayed as lists, or (as here) as a table showing collocates for each position in relation to the node (i.e. in this OHP, the leftmost column shows the most frequent collocates 3 words to the left of *authority*, the second column shows the most frequent collocates 2 words to the left of *authority*, etc).
3. In this OHP the collocates are ordered by raw frequency, but they can also be ordered by statistical measures (T-score, MI-score).
4. It is also possible to see, for any collocate in any position, all the concordance lines containing that specific co-occurrence.
5. The display above shows that the NODE (in this case *authority*) is frequently followed by *to* (usually a to-infinitive), *and*, *of*, *in*, *the*, *on*, *is*, etc. Premodifiers are evident in the third column: *local*, *health*, *education*, etc.
6. In general, collocational strength is greatest in positions closest to the node.

OHP12 – Collocation (cont.)

Collocates of *file*

file as VERB			*file* as NOUN		
collocate	**frequency as collocate**	**t-score significance**	**collocate**	**frequency as collocate**	**t-score significance**
to	669	18.45	*rank*	442	21.00
for	187	8.84	*fact*	178	12.72
bankruptcy	52	7.19	*tape*	150	12.14
complaint	45	6.69	*on*	398	11.24
against	54	6.64	*and*	934	11.08
suit	41	6.34	*single*	96	9.30
charges	39	6.14	*a*	801	9.06
will	66	5.60	*feature*	75	8.54
returns	30	5.43	*from*	227	7.68
lawsuit	28	5.28	*the*	1739	7.24
a	249	5.27	*archetype*	51	7.14

OHP12 – NOTES

1. OHP10 showed that conflating types can be beneficial in some analyses: we learned that *be* is the second most frequent lemma in English (whereas *of* is the second most frequent type).
2. OHP12 shows that in other analyses, it is useful to subdivide the concordances for a single type (in this case *file*) before analysis.
3. In this case, the subdivision is based on word-class (*file* as verb and *file* as noun are analysed separately), but interesting differences may also be discovered by analysing separately the concordances from spoken data and written data, or from American data and British data, from tabloid newspapers and broadsheet newspapers, etc.
4. Here we can see that *file* as a verb forms phrases such as *to file for bankruptcy, file a complaint against* (someone), *file a lawsuit*, etc.
5. Whereas *file* as a noun forms phrases like *rank and file, fact file, a file* (stored) *on tape, tapes* (kept) *on file, single file*, etc.

OHP13 – linguistic creativity and phrasal variation

```
        two beanshoots short of a spring roll
             a bishop short of a chess set
             51 cards short of a full deck
           five cards short of a full house
         several cards short of a full hand
         several cards short of a full deck
            two chops short of a barbie
           a few eggs short of a dozen
         two fishbones short of a Royal funeral
    several hatstands short of a cloakroom
a couple of kangaroos short of a full paddock
        several pawns short of a full set
              a penny short of a calculator
           two prawns short of a paella
           a sandwich short of a picnic
          one sandwich short of a picnic (2)
about five sandwiches short of a picnic
```

two sarnies short of a picnic
one spanner short of a socket set
two sticks short of a bundle
one swastika short of a Nuremberg Rally
a few tokens short of a pop-up toaster
several wigs short of a judges' convention

OHP13 – NOTES

1. One of the most striking features of corpus analysis is the high degree of linguistic creativity that human beings are capable of, as shown in the wide variations in common phrases (especially humorous ones).
2. Items such as this provide lexicographers with an almost impossible problem: at which headword should this phrase be included, and which of the variations should be shown as examples?

OHP14 – Bank of English – Future Possibilities

- **Collins Language Databanks**
- **Increasing International Sources**
- **Increasing Technical Sources**
- **Increasing Web-sourced Data**
- **Monitor Corpus**
- **Software Developments**
- **Change the Administrative Database**

OHP14 – NOTES

Possible changes over the next year or two include:
1. Collins Language Databanks - allowing public access to corpora in several languages using the same software, which reflects the increasing use of corpus methodology in bilingual dictionaries.
2. Increasing International Sources - to better represent `global English'.
3. Increasing Technical Sources - medicine, science, technology, computing, etc, for the "English for Special Purposes" aspect of EFL, and for native-speaker dictionaries.
4. More Web-sourced Data - changes in technology are changing corpus data acquisition methods.
5. Monitor Corpus - to keep up with the increased speed of language change, especially in English; needed by both native-speaker and bilingual dictionaries.
6. Software developments - for lexicography in particular, but also for corpus linguistics in general.
7. Changes to the Administrative Database, which holds information about all the texts in the Bank of English - eventually allowing users to create their own corpora by selecting individual texts (e.g. women's writing only; autobiography only; telephone dialogues only; 1999 texts only; etc).