RAMESH KRISHNAMURTHY

CHANGE AND CONTINUITY AT COBUILD (1986–1996)

1 Introduction

The COBUILD project was set up jointly by the University of Birmingham and Collins Publishers (now HarperCollins) in 1980 with two main aims: to collect and analyze a large computerized corpus of contemporary English language, and to publish the results in a range of reference and pedagogical books for learners and teachers of English as a Foreign or Second Language.

1.1 Change

The range and scale of the changes at COBUILD during the past decade have been enormous. Our physical location has moved from a converted semi-detached house to an open-plan office block in the University Research Park. Our personnel has increased from 12 full-time staff to 22. Our organizational status has changed from an externally-funded research unit at the University to a limited company and sub-division of HarperCollins. Our primary orientation has shifted from academic research to commercial publishing. Our use of computer technology has intensified greatly, from part-time use of the University mainframe to running our own network of Unix workstations.

1.2 Continuity

However, change only represents part of the picture, and it is equally important to recognize the elements of continuity in COBUILD's development. The basic philosophy, of collecting large amounts of data, analyzing the data thoroughly, and presenting the results in a user-friendly way, remains at the core of everything we do. Although many of the original team have left, about half a dozen staff have been involved in the project for over a decade. The change in organizational status has not diminished the close relationship

between Cobuild staff and academics at the University. The note of continuity is particularly evident in our publications. For example, most of the innovative features of the original COBUILD Dictionary (1987) have been retained in the new edition (1995), although many of them have been refined or enhanced.

2 Data

2.1 Size

The 1987 COBUILD Dictionary was based on an initial detailed examination of a 7.3 million word corpus (six million words of written texts, 1.3 million words of spoken texts). This analysis was subsequently enhanced with reference to a 20 million word corpus prior to publication of the dictionary. By 1995, the corpus (now called the Bank of English) had grown to over 211 million words, and provided the evidence for the 1995 edition. During 1996, we anticipate that it will exceed 300 million words.

2.2 Vintage

The 20 million word corpus consisted mainly of data from 1975-1985, whereas the 211 million word corpus data originated largely from 1985-1995. So the new corpus was, of course, more up-to-date and reflected many linguistic and real world changes.

2.3 Corpus And Subcorpora

The 20 million word corpus was stored as one single entity. The 211 million word corpus is held as 16 subcorpora, distinguished by source or text type. This allows finer tuning of the analysis, including contrastive studies of different genres of language, such as informal speech and broadcast speech, broadsheet newspapers and tabloids, etc. We hope to add two new subcorpora during 1996.

2.4 Integerization

Another major difference is not apparent to the user, but has had a substantial impact on the speed of the corpus retrieval programs: the 20 million word corpus was stored as characters, and therefore even the simplest search program had to match each character of the search word with each character of the corpus word. The 211 million word corpus is held as integers (i.e. each word is encoded as a number), so the search program now only has to match one number with another.

2.5 Wordforms

The 20 million word corpus gave us information about c. 250,000 wordforms. However, many of these do not constitute valid candidates for dictionary entries. Most proper names need to be excluded. Regular inflected forms (such as plural forms of nouns, comparative forms of adjectives and adverbs, and inflected forms of verbs) that show no semantic, syntactic, or stylistic deviations from the base form, will be subsumed under the entry for the base form. On the other hand, multi-word items (such as phrasal verbs, noun compounds, idiomatic phrases) serve to extend the final inventory. The 211 million word corpus represented a tenfold increase in overall corpus size, but yielded only a two-fold increase in the number of wordforms (c. 500,000). The proportions of proper names and regular inflected forms remained roughly the same.

2.6 Frequency

The increase in corpus size should not be considered solely in terms of number of wordforms. The frequency of occurrence of each wordform is also extremely significant for lexicography. However, the increased frequency is of little benefit to the lexicographer in the analysis of the very common words, because they do not vary a great deal in corpus frequency rank or in linguistic usage, but merely reflect the tenfold increase in overall corpus size:

20m corpus (1987)		211m corpus (1995)	
Word	Number of occurrences	Word	Number of occurrences
the	1,081,654	the	11,611,078
of	535,391	of	5,359,185
and	511,333	to	5,180,130
to	479,191	and	4,941,561
a	419,798	a	4,537,660
in	334,183	in	3,796,752
that	215,322	that	2,226,871
it	198,578	it	1,954,556
i	197,055	is	1,940,162
was	194,286	for	1,794,630

The top 8 items are exactly the same, except that 'to' and 'and' have changed position. Similarly, the replacement of 'i' and 'was' by 'is' and 'for' is not very significant, and merely represents minor changes of position. However, even here, a word of caution is required. As we shall see below, even frequent words must be constantly reviewed, or subtle changes in usage will be missed.

Most wordforms lower down the frequency order, for which the 20 million word corpus gave insufficient information, are usually much better attested in the 211 million word corpus, and more precise and detailed information can be obtained from the larger corpus.

Very low frequency items are of little use in any corpus, as they do not provide enough evidence on which to base any lexicographic statements (for example, about half of the wordforms in a corpus occur only once: i.e. over 125,000 wordforms in the 20 million word corpus, and over 250,000 in the 211 million word corpus).

3 Corpus Access And Retrieval

3.1 History

This area has changed most radically in the past decade. Until 1985, we looked at paper printouts of concordances from the 7.3 million word corpus. This meant that only one person could inspect a particular set of concordances at any given time, and that concordances could be easily misplaced, lost, or damaged.

Each page of printout contained 56 concordance lines with a context of 100 characters that were sorted to the right of the keyword, with text references at the left of each line.[1] This gave us reasonable information about the immediate right-hand collocations (nouns modified by an adjective keyword, prepositions governed by verb keywords, etc), but non-adjacent collocations were difficult to spot, and left-hand collocations were even more difficult to identify.

If a particular concordance line contained insufficient context, the relevant context could only be obtained by tracking down concordances for other words in the line, a time-consuming affair

[1] Looking Up. Ed. by J. M. Sinclair, HarperCollins Publishers, London, 1987. p. 36.

(especially if those concordances were being used by another lexicographer, or had been lost or misplaced).

In 1986, each lexicographer was supplied with a set of all the concordances on microfiche, and this made such searches much easier. Fiches were less easy to misplace, lose, or damage. The fiches also added data from the 20 million word corpus, but only for words of lower frequency in the 7.3 million word corpus.

Also in 1986, a one million word corpus was made available online. This automatically provided longer contexts and the facility to re-sort concordance lines by the word to the left of the keyword.

3.2 Lookup

The 1996 Bank of English corpus is always used online (although printouts can, of course, still be obtained if needed) and has a tremendous array of display options. All the main access and retrieval facilities are integrated into a single computer program called 'lookup'.

3.2.1 Frequency And Subcorpus Distribution

The 211 million word corpus is held as 16 subcorpora. So whenever we investigate a word or phrase, 'lookup' first gives us some information about frequency and subcorpus distribution. This enables us to see whether a word is more commonly used in speech or writing, in British or American English, and so on.

For example, here are the figures for 'actually' (see Appendix for an explanation and description of the subcorpora) :

Corpus	Total Number of Occurrences	Average Number per Million Words
spoken	21131	1360.5/million
npr	6704	301.1/million
bbc	3944	210.7/million
scbooks	763	190.9/million
mags	5220	173.2/million
american	3360	172.8/million
britbooks	4026	168.1/million
guardian	1449	115.0/million
newsci	429	102.7/million
indy	490	97.3/million

econ	815	93.2/million
times	934	90.2/million
today	1622	89.6/million
wsj	559	89.0/million
oznews	866	84.2/million
ephem	148	78.6/million

This clearly suggests that 'actually' is used more in speech than in writing.

Another example is 'robust':

Corpus	Total Number of Occurrences	Average Number per Million Words
wsj	128	20.4/million
times	156	15.1/million
econ	119	13.6/million
indy	62	12.3/million
guardian	155	12.3/million
newsci	50	12.0/million
mags	344	11.4/million
american	136	7.0/million
oznews	65	6.3/million
britbooks	147	6.1/million
ephem	10	5.3/million
bbc	95	5.1/million
today	84	4.8/million
scbooks	14	3.5/million
npr	70	3.1/million
spoken	30	1.9/million

This suggests that 'robust' is more often used in writing than in speech, and more often in economic texts and broadsheet newspapers than in tabloid newspapers and ephemera.

Multi-word items can be similarly investigated. A search for 'new+broom' yields the following information: it is more often used in British English than American English, and more in writing than in speech:

Corpus	Total Number of Occurrences	Average Number per Million Words
guardian	14	1.1/million
times	10	1.0/million
econ	7	0.8/million
today	10	0.6/million
newsci	2	0.5/million
indy	2	0.4/million
scbooks	1	0.3/million
spoken	3	0.2/million
npr	3	0.1/million
britbooks	3	0.1/million
mags	3	0.1/million
oznews	1	0.1/million
bbc	1	0.1/million
american	1	0.1/million
ephem	0	0.0/million
wsj	0	0.0/million

An interesting feature is brought to light when we look for 'different+than': we may have intuited that it is more favoured by American English than British English, but the data shows that it is also very frequent in British speech, though comparatively rare in British writing:

Corpus	Total Number of Occurrences	Average Number per Million Words
npr	218	9.8/million
american	53	2.7/million
spoken	40	2.6/million
wsj	13	2.1/million
scbooks	8	2.0/million
oznews	9	0.9/million
britbooks	18	0.8/million
mags	17	0.6/million
bbc	6	0.3/million
today	5	0.3/million

newsci	1	0.2/million
indy	1	0.2/million
times	2	0.2/million
guardian	2	0.2/million
ephem	0	0.0/million
econ	0	0.0/million

3.2.2 Concordances

Concordances from 'lookup' can be examined with considerable flexibility. In the following example, I have first isolated the lines from the 'spoken' subcorpus, then sorted by two words to the right of the keyword ('node' in current program terminology), so that it is easier to see whether 'than' is followed by a noun group or a clause.

< right.<F01> it would b+ it's	different than	<M01> You've got time then.>
< either actually.<M0X> Well <M0X>	Different than.	<M0X> it's in the grammar >
some American English speakers say	different than	# <M0X> Mm.<M0X> but >
< W+ <ZF0> we don't even recognize	different than	<ZF1> in <ZF0> in the <ZGY>
<derogatory <ZGY> <F0X> Yeah.<M0X>	Different than	<ZGY> view of <ZF1> the >
< I think a theory aspect is very	different than	a practical aspect.<M01> >
< No it's rather <F01> Er no.<M01>	different than	a convent. [laughs # <F01> >
< <ZZ0> Erm was your build-up any	different than	before?<F02> tuts] Oh yes. >
< hospital.<M01> Does sound rather	different than	Edmund Street <F01> Yes. >
< like different to different from	different than	erm again I think erm that >
< <M01> Yeah.<M02> it's so much	different than	from all the other models >
<know he said to me it would be no	different than	having a fag but it was. >
but you will hear different to and	different than.	I mean that's would seem >
< way of approaching it was quite	different than	if you're living with >
< was actually probably not very	different than	it is now only it was far >
< changed. It's all so much	different than	it used to <ZF1> be <ZF0> >
< in that union must have been	different than	it was by the time you >
speak <ZF1> is i+ i+<ZF0> is a lot	different than	it was <ZF1> ten y+ <ZF0> >
is my fault like. That's where I'm	different than	most of 'em.Er so <ZF1> I >
<<ZF1> is <ZF0> is like completely	different than	my idea <ZGY> what sexism >
<erm she thinks she's a little bit	different than	other people.<M01> Mm.<F01>
< Has that come up in B B C data?	Different than.	pause # <F0X> I don't know >
< think [pause] oh is a sticky rim	different than	suction?<F03> Well it's >
< well I thought er Brian you're	different than	that but er [pause] that's >
< <ZF1> I mean <ZF0> I mean er	different than	the rest of Europe I feel. >
< <M01> Er apparently he was much	different than	the previous vice >
solved by the Dutch child is a bit	different than	the problem to be solved by
I think the contemporary family is	different than	the traditional family and >
<It's P E it's just like something	different than	the lessons you have to sit

like it's going to be really a lot	different than	the description it gave to >
< of observing the ocean is so	different than	the atmosphere. The >
< should he be getting treated any	different than	them. You know what I mean.>
< qualities on that one are very	different than	there.<M01> Well it might >
< fact that erm ice particles are	different than	water particles <M01> Yeah.>
s twenty-six and they are entirely	different than	what this younger one is. >
< then you know. It's something	different than	what we usually do like >
< <M01> But it was certainly	different than	what came afterwards wasn't
< corrupt it's just a completely	different than	what we're used to.pause] >
< development's probably something	different than	what East Anglia means >
< if you judge somebody who's	different than	you so much <ZF1> on <ZF0> >

In fact, concordance lines can be sorted by any word up to five words to the left or right of the node.

3.2.3 Collocation

The collocations that had to be obtained manually and somewhat haphazardly in 1986 can now be obtained at a keystroke, accompanied by robust statistical measurements. The top collocates can be displayed as a list. For example, this is a list of the top 24 collocates for 'continuity':

of	678	16.525324
and	451	10.481660
sense	54	7.104089
between	54	6.455226
change	44	6.237443
the	692	5.904582
ensure	35	5.818858
there	79	5.815418
stability	31	5.527369
is	157	5.373034
historical	28	5.228340
announcer	26	5.092955
policy	30	5.084427
maintain	24	4.805348
provide	22	4.411755
lack	9	4.000000
degree	16	4.000000
announcement	23	4.000000
closing	14	4.000000

jewish	8	4.000000
element	8	4.000000
discontinuity	7	4.000000
care	19	3.937419
in	234	3.824960

There are 1222 occurrences of continuity in the 211 million word corpus, and the list tells us that 678 of the 1222 occurrences have 'of' within four words of 'continuity', 451 have 'and' within four words, 54 have 'sense' within four words, and so on (see the second column). The third column is a statistic called 'T-score', which takes into account the corpus frequencies of 'continuity' and 'of' and indicates the statistical significance of the collocation when compared to random distribution.

The list does not tell us where the collocate occurs in relation to the node, before it or after it, one word or three words away, etc. So the collocates can also be displayed in a positionally specific graphic form (called 'picture' in the Cobuild system), up to six words on either side of the node.

For example, here is the 'picture' for 3 words either side of 'continuity', again based on 'T-score'):

there	to	and	NODE	of	the	also
a	of	the	NODE	and	care	world
to	is	of	NODE	in	change	history
sense	provide	a	NODE	announcer	than	foreign
opening	change	for	NODE	with	service	policy
closing	break	provide	NODE	is	consciousn	past
this	maintain	some	NODE	from	stability	and
no	ensure	historical	NODE	between	progressio	there
of	stability	maintain	NODE	rather	this	in
so	lack	ensure	NODE	or	its	from
by	degree	its	NODE	there	one	of
who	announceme	than	NODE	<t>	all	this
start	thread	no	NODE	has	our	life
provide	would	family	NODE	as	been	gulf
can	there	essential	NODE	but	support	continue
some	and	giving	NODE	important	their	training
would	an	needs	NODE	which	education	local
able	some	experience	NODE	here	work	church
makes	lose	both	NODE	point	us	between
important	need	represent	NODE	although	between	whole

which	believe	offer	NODE	tradition	important	council
social	keep	about	NODE	seems	social	which
history	form	want	NODE	otherwise	ancient	change

We now discover that 'continuity' is (see the column to the left of 'NODE') often preceded by verbs such as 'provide, maintain, and ensure', that 'continuity' is often modified by adjectives such as 'historical, family, and essential'. From the column to the right of NODE, we see that 'continuity' is often followed by the prepositions 'of, in, with, from, and between', and exists in the noun compound 'continuity announcer'.

A single keystroke can take you from the list or 'picture' display to the concordance lines containing a particular collocate, from which you may want to select an example. For example, if we select 'historical':

< served as symbols of	historical continuity,	important for ethnic Malays, who
a remarkable example of	historical continuity	within a modern centralised >
that magic circle where	historical continuity	and intellectual intercourse cast
pride and identity, the	historical continuity,	the Palestinian determination >
< rehearse them daily:	historical continuity;	urban architecture and town- >
< the emblem of	historical continuity,	the seal on the unbroached >
of any violent break in	historical continuity.	The entrusting of this role to a
< role as an emblem of	historical continuity	in British life" while her powers
<role, 'as an emblem of	historical continuity	in British life and (as long as >
there is in fact a long	historical continuity	in migration movements in France.
< the ethos, if not the	historical continuity,	of the original Rosicrucian >
releases Weber from the	historical continuity	of his age, however stringently >
<in which he defended a	historical continuity	between the British Empire and >
< the perception of	historical continuity	in the prevalence of long-term >
is all part of the same	historical continuity	up to the Boer War and beyond, if
< Without feeling for	historical continuity	he assumed his own art was >
< role as an emblem of	historical continuity	in British life # Labour's deputy
6 7; 11:1-5), while the	historical continuity	was especially stressed by >
<demand this unity: the	historical continuity	between the Testaments; the >
< Forest an agent of	historical continuity	that binds the fifty-year-old >
<below and suggests the	historical continuity	of Satyagraha.<t> Tolstoy (Act I)

3.2.4 Expanded Contexts

Expanded context for any concordance line is available at the touch of a key. For example, here is the expanded context for the first line above (note that the text reference is always provided automatically - in case the user wishes to attribute the source):

no single sultan was recognised as sovereign over the others, the nine families agreed to take five-year turns serving as the nation's Agong, or King.
The sultans served as symbols of historical continuity, important for ethnic Malays, who make up about half of the country's population and are at odds with an ethnic Chinese minority that dominates business in Malaysia.
Corpus: oznews Text_id: <tref id=N0000005348>

3.2.5 Text References

Text references can be displayed for each concordance line, at a single keystroke, and full details for each text are available on request. Cobuild lexicographers mainly use them to make sure that the concordances are coming from a reasonable variety of source texts:

ozn N0000005348 of	historical continuity,	important for ethnic Malays, who
tim T0000080892 of	historical continuity	within a modern centralised >
npr N0000050792 where	historical continuity	and intellectual intercourse cast>
npr N0000270492 the	historical continuity,	the Palestinian determination >
mag N0000000505 daily:	historical continuity;	urban architecture and town- >
mag N0000000130 of	historical continuity,	the seal on the unbroached >
ind I0000041090 break in	historical continuity	The entrusting of this role to a >
ind I0000021090 of	historical continuity	in British life" while her powers>
ind I0000011090 of	historical continuity	in British life and (as long as >
bri B0000000850 a long	historical continuity	in migration movements in France.>
bri B0000000719 not the	historical continuity,	of the original Rosicrucian >
bri B0000000564 from the	historical continuity	of his age, however stringently >
bri B0000000560 a	historical continuity	between the British Empire and >
bri B0000000560 of	historical continuity	in the prevalence of long-term >
bri B0000000560 the same	historical continuity	up to the Boer War and beyond, if>
bri B0000000311 for	historical continuity	he assumed his own art was >
bbc W0000001090 of	historical continuity	in British life # Labour's deputy>
ame B0000001074 the	historical continuity	was especially stressed by >
ame B0000001074 the	historical continuity	between the Testaments; the >
ame B0000000428 agent of	historical continuity	that binds the fifty-year-old >
ame B0000000389 the	historical continuity	of Satyagraha.<t> Tolstoy (Act I)

3.2.5 Grammar Tags

The word class of the node can be displayed at any point during the search. For example, here are 24 concordance lines for 'report': as is evident, most are nouns (NN or NP), but two are (correctly) tagged as verbs (VB). The tagging is not always one hundred per cent accurate, but errors are usually easy to spot, or are ambiguous even for the grammarian.

NN	spanly obtained supplies, the	report	says. And it urges the United >
NN	brains have been damaged. A	report	from Dick Oliver of our Science >
NN	body, is sceptical about the mps'	report	It addresses itself to >
NP	newsnote for dawn reel amnesty	report	on jordan By Francis Mead </h> <t> NN
	Kashmir Valley and has sent this	report	on the continuing campaign there
NN	to the Cambodian conflict # This	report	from Peking by Simon Long:<h> DESK
NN	today's operation which the	report	said was witnessed by a >
NN	about it later.<t> The race	report	is just as intriguing. Senna had
VB	RUBINSTEIN <t> Sir: You	report	the malodourant results of the >
NN	Mikey Massive <LTH> THIS week's	report	by the Royal Commission on Criminal
NN	Middle East Watch says their	report	is an attempt to break through the >
NN	to Moscow and make some kind of	report	# And I asked Lithuanian officials >
NN	prices for April # March's	report	noted a gain of 2/10ths of a >
NN	successor # Don Gonyea filed this	report	<t> Don Gonyea reporting:<t> The GM
VB	of information.<M01> Mm.<F01>	Report	back to the group and then draw up >
NN	intrusion on others. <t> The	report	came as leaders of different faiths
NN	Monopolies and Mergers Commission	report	into beer supply. <t> Now Camra is >
NN	regulator returned a	report	because 'it wasn't heavy enough, it
NN	<hl> International -- Corporate	Report:	Pioneer Electronic Corp # </hl> >
NN	information. It suggests, in a	report	to be published tomorrow, that the >
NN	away # <t> Almost every recent	report	on residential child care has cited
NN	this week it published a	report	suggesting how this might work.<p> >
NN	arrangement with an employer.The	report	studied a number of investment >

Other online facilities include the ability to view data from one or more selected subcorpora only, and to edit concordance output and re-run frequency or collocation programs on the edited sample. All these facilities have helped the lexicographer to generate a much more accurate and reliable dictionary entry.

4 Dictionary Compiling And Editing

This is another area that has changed dramatically since 1986. The original analysis of the 7.3 million word corpus was written on paper slips and then input to a database by keyboarders, and printed out for

lexicographers to revise. In 1996, not only is the analysis conducted online, but the results are simultaneously entered into dictionary files on the computer.

The dictionary files are carefully and rigorously coded, so that all the information can be checked automatically by a validation program. It is also relatively easy to tell typesetters how to deal with the data.

Here are some extracts from the file containing the entry for 'robust':

[HWD]robust
[FRQ]2
[PRN]/r&o&!ub*%ust, r*o*!ub&%ust/

[MNM]1
[GRM]ADJ-GRADED
[SYN]sturdy
[ANT]
[DEF]Someone or something that is [CIT]robust[/CIT] is very strong or healthy.
[EXB]
[X10]More women than men go to the doctor. Perhaps men are more robust or worry less?...

The code [HWD] indicates the headword, [FRQ] the frequency band, [PRN] the pronunciations, [MNM] the meaning number, [GRM] the grammar, [SYN] the synonyms (if any), [ANT] the antonyms (if any), [DEF] the definition (with the headword highlighted by [CIT] and [/CIT]), [EXB] the beginning of a set of examples, [X10] an example selected from subcorpus number 10.

This coding system can also be used to do analyses of the dictionary text, for example to find all the definitions that contain a particular word, all the words with a particular grammar, etc.

5 A Brief Comparison Of The Original Cobuild Dictionary (1987) And The New Edition (1995)

5.1 Features

Most of the important surface features of the original dictionary have been retained: selection of headwords based on corpus frequencies, full-sentence definitions explaining usages in terms of linguistic context (collocations, structural patterns) and social context (style, register, pragmatics), use of real examples from the corpus, and so on.

However, in response to our continuing research and users' comments, some changes have been made:

5.1.1 Superheadwords

Long entries ('superheadwords') have been subdivided for easier reference either by semantic or syntactic grouping.

5.1.2 Grammar Codes

The grammar codes have changed from a mixture of formal and functional labels (e.g. V+O; V = verb = formal label; O = object = functional label) into a simple surface description consisting of only formal labels (e.g. V n; V = verb = formal label; n = noun or noun group = formal label). Also, the examples have been ordered so that they match the sequence of grammar codes.

5.1.3 Superordinates

Superordinates were provided in the 1987 Dictionary, but were not easily understood, and were usually already contained in the explanations, so they have been omitted from the 1995 edition.

5.1.4 Frequency Bands

Frequency information has been added for every headword: five black diamonds means that the headword is in the 'most frequent' band, no diamonds means that the word is in the 'least frequent' band. About 15,000 headwords have one or more diamonds.

5.1.5 Pragmatics

Although pragmatic information was an important feature of the 1987 Dictionary, many users commented that it was not clearly signalled, and was therefore easily overlooked. Therefore, in the 1995 edition, a box labelled PRAGMATICS has been introduced into the Extra Column.

5.1.6 Examples

A major feature of the 1995 edition is that all the examples from the 20 million word corpus have been replaced by examples selected from the 211 million word corpus. The examples are therefore more up to date and reflect current real world events and personalities.

5.2 Coverage

For the 1987 Dictionary, a rough test for inclusion was a minimum of 6 occurrences from 3 different source texts. For the 1995 edition, this was increased to around 15 occurrences. However, because of the increase in corpus size, the number of references has still increased from 70,000 (1987) to 75,000 (1985).

5.3 New Words, New Meanings

Many of the new entries in the 1995 edition reflect changes in technology ('camcorder'; 'multimedia') and in society ('PC' meaning 'politically correct'; 'ethnic cleansing'). Other words have become archaic ('baksheesh', 'golliwog') or have been replaced ('Red Indian' by 'Native American', 'Common Market' by 'European Union').

The 1995 edition includes a new noun use at 'take 42': "Someone's **take** on a particular situation or fact is their attitude to it or their interpretation of it". An even more recent glance at the corpus revealed a new business use of the phrase 'take a bath' meaning 'lose a lot of money'; so even common words like 'take' need to be kept under constant scrutiny.

The 1987 Dictionary listed 'the **abandonment** of a place, person, or thing' and 'the **abandonment** of a piece of work, plan, or activity', but the 1995 edition adds a third usage 'the **abandonment** of an idea or way of thinking'.

Similarly, the 1995 Dictionary covers new uses of 'sad' (If you describe someone as **sad**, you do not have any respect for them and think their behaviour or ideas are ridiculous; an informal use.), 'macro' (in computing), 'parked' (If you are **parked** somewhere...) and so on.

6 Back To The Future

6.1 Using A Corpus

In 1987, Cobuild was the only dictionary to be produced using corpus data. In 1996, all the other major dictionary publishers claim

to be using corpora, not only for EFL dictionaries but also for bilingual and native-speaker dictionaries.

6.2 Using Full-Sentence Definitions

In 1987, Cobuild was the only EFL dictionary to use full-sentence definitions. In 1996, several other dictionaries have adopted this practice.

6.3 Using Authentic Examples From The Corpus

In 1987, Cobuild asserted the importance of using authentic examples directly from the corpus. In 1995, most other major EFL dictionaries claim that their examples are to some extent at least 'corpus-based'.

6.4 publishing the data

COBUILD has created a reputation for innovation and creativity. Several years ago, we began to get requests from users who wanted to see the data on which our publications were based. So we published edited concordance data in our Concordance Samplers for teachers to photocopy and use in the classroom. We also released a small corpus (called the Word Bank) on our COBUILD on CD-Rom. Collocation data for the top 10,000 words was released on another CD-Rom.

6.5 Extending Online Use

Now we are offering an online and email service (COBUILD Direct) for people who want to use our corpus data on a regular basis. We have a Website on the Internet (http://www.cobuild.collins.co.uk) where users can access a demonstration model of our corpus, look at our dictionary and other publications, and participate in various competitions and language activities. We encourage feedback from COBUILD users by letter or by email (e-mail: editors@cobuild.collins.co.uk) on all our publications and services, and take this feedback into account when planning our future activities.

7 Conclusion

I would like to conclude with two quotations which I think are particularly relevant. The first specifically refers to 'graphemes' and

'phonemes', but the general point seems to back up Cobuild's confidence in the use of corpora:

"It is sometimes objected that there is no one-to-one correspondence of grapheme and phoneme, or that the graphemes show irregularities or inconsistencies of use. This is true; but there is consistency in the inconsistencies, and in sufficiently large samples the inconsistencies are leveled out. As the number of chances becomes larger and larger, the effects of each single event become less and less important and tend to cancel out; the final consequence approximates the average with great accuracy." (Joshua Whatmough, 'Language: A Modern Synthesis', Mentor Books, New American Library, New York, 1956, pp 180-1).

The second quotation is taken from Arthur C. Clarke's address to the American House of Representatives Committee on Space Science and Applications on 24th July 1975, and talks about how accurately we can predict the future:

"It's a cliche that we often tend to overestimate what we can do in the near future - and grossly underestimate what can be done in the more distant future. The reason for this is very obvious, though it can only be explained with a certain amount of hand-waving. The human imagination extrapolates in a straight line, but in the real world, as the Club de Rome and similar organisations are always telling us, events follow a compound-interest or exponential law. At the beginning, therefore, the straight line of the human imagination surpasses the exponential curve; but sooner or later the steeply rising curve will cross the straight line, and thereafter reality outstrips imagination.

How far ahead that point is depends not only on the difficulty of the achievement, but also upon the social factors involved." (Arthur C. Clarke, To the Committee on Space Science, ch 23 of The View from Serendip, Victor Gollancz, London, 1978).

Appendix

Bank of English: Summary of Composition
April 1995

Source	size (millions of words)	number of texts	percentage of total
BBC World Service	18.7	(500)	8.88%
Independent	5.0	49	2.38%
Times	10.3	79	4.89%
National Public Radio (Washington)*	22.0	729	10.45%
Economist	8.7	28	4.16%
ephemera	1.8	837	0.86%
New Scientist	4.1	92	1.95%
Australian news	10.2	141	4.85%
Spoken (general)	15.5	1571	7.36%
Today	18.1	540	8.60%
American books and newspapers*	19.4	273	9.22%
British books	27.9	406	13.25%
Guardian	12.6	137	5.99%
Magazines (general, popular)	30.0	760	14.25%
Wall Street Journal*	6.2	15	2.95%
TOTAL	210.5	6156	

* = US English

Additional corpora:

Business English (journals, textbooks)	3.0
Academic writing	1.5
GCSE textbooks	1.0