Krishnamurthy, R. 2003. *Freeze-frame pictures: micro-diachronic variations in synchronic corpora* pp 15-31 in Jozsef Andor, Jozsef Horvath, and Marianne Nikolov (eds), `Studies in English Theoretical and Applied Linguistics', Lingua Franca Csoport, Pecs, Hungary; ISBN 963-641-994-9

Freeze-frame pictures: micro-diachronic variations in synchronic corpora

Ramesh Krishnamurthy, Cobuild, University of Birmingham, England
ramesh@cobuild.collins.co.uk

(based on a paper given in Seminar 6: *Corpus Linguistics and Computational Lexicography* at the ESSE 4 Conference, Debrecen, Hungary, September 5th-9th 1997)

1. Language Change

A living language is constantly changing, in order to adapt to new human situations, new human needs, and new human ideas. In the past, such changes were in general noted and studied only after a long period of time had elapsed, by historians of the language.

For example, in the front matter of the Times English Dictionary (1st edition, 2000), an article entitled *The Making of English* by David Brazil charts the progress of English from a language spoken by five or six million people in the British Isles in 1500 AD to a world language which is the mother tongue of over 300 million people all over the world in 2000 AD, and spoken by millions more. Brazil mentions words introduced by the Angles, Saxons and Jutes from the fourth century AD on which are still core vocabulary today: *man, child, eat, drink, sleep, love, hate, land, harvest, crops, he, him, her*. Christianity and Latin were responsible for words such as: *priest, monk, hymn, altar, master, grammar, plaster and fever*. The Vikings brought: *skirt, shirt, husband, ugly, call, want, they, them, their*. The Norman invasion gave us words from Norman French (*cattle, warden*) as well as their Central French cognates (*chattel, guard*). In more recent times, words have arrived in English from diverse sources all over the world: Canada (*igloo, anorak, kayak*), the Caribbean islands (*calypso, reggae, yardie*), Australasia (*kangaroo, dingo, wombat, outback, cobber*), Asia (*nirvana, fatwa, tandoori*) and Africa (*gnu, mamba, tsetse, veldt*).

Changes in language have not always been welcomed. The Roman poet Horace complained in his *Ars Poetica* (1st century BC): “As for me, why should I be criticized if I add a few words to my vocabulary, when the language of Cato and Ennius enriched the speech of our fathers and produced new names for things?” (DeWitt, 1961). In modern times, some language communities try to regulate the changes by means of Academies and laws, but even when they succeed, their efforts exercise only a temporary restraint. Although English has not had an Academy as such, lexicographers (e.g. Johnson (1755) and Murray (1884 -1928)) and style gurus (e.g. Fowler (1906)) have sometimes tried to inhibit the process of change. In recent years, the Campaign for Better English, Prince Charles, Professor John Honey and others have tried to limit or reverse some of the changes currently taking place in English. However, as Horace recognized long ago

(DeWitt 1961): “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.”

However, the main focus of this paper is not on our attitudes to changes in language, but the identification and measurement of the changes.

2. Freeze-frames

In the not-too-distant past, films were shown only in cinemas. An enthusiast might see the same film several times during its initial run, but most cinema-goers would see a film only once. Subsequent opportunities to see that film were rare, for example during occasional revivals, film festivals, or retrospective seasons.

The advent of television has allowed more frequent viewings of films. The invention of the VCR now enables us to see a film - or indeed any video or television broadcast - as often as we wish. And the pause button makes it possible to isolate and inspect a single frame - a freeze frame. Thus, the film that was originally a rapid sequence of visual actions viewed only once is now also available as a large collection of discrete single-frame images. The new DVD technology offers the same functionality.

Similarly, in language studies, we have progressed from being able to hear an oral text only once, to reading written texts, and then via printing and audio-recording technologies to re-reading or re-hearing texts whenever and however often we want. In the briefer history of electronic language corpora, we have begun to shift from viewing a corpus only as a unified whole to the facility to inspect texts produced within the same small periods of time, a year, a month, or even a day.

3. Diachronic and Synchronic

The words *diachronic* and *synchronic* are not in widespread general use, although they have been in existence for a long time. The Oxford English Dictionary (1971) describes *diachronic* as a nonce-word, and its earliest citation is from 1857. The word *synchronic* is labelled rare, and it is first attested in 1833. The 323 million word Bank of English corpus of 1996 had only 7 citations of each term, 5 for each from the same academic lecture recorded at the University of Keele. The other 2 citations for *diachronic* were from a book entitled Labyrinths by Maurice Berger, and the other 2 citations for *synchronic* were from the Guardian (5th August 1995) and the New Scientist (18th April 1992). The updated and expanded 418 million word Bank of English corpus of October 2000 has not significantly increased the evidence: there are still only 8 citations for *diachronic* and 9 for *synchronic*.

R.H. Robins introduces and defines the terms as they are used in linguistics (Robins 1989:5): “The terms ‘synchronic’ and ‘diachronic’ are in general used to distinguish respectively linguistic statements describing a stage of a language as a self-contained means of communication, at a given time, during which it is arbitrarily assumed that no changes are taking place, and

statements relating to changes that take place in languages during the passage of years." In a footnote to this paragraph, Robins adds: " 'Synchronic' and 'diachronic', like a number of other basic terminological distinctions in linguistics, are Saussurean in origin (51, 114-43). De Saussure gave us the term 'état de langue' to refer to a stage of a language at a particular period; thus Chaucerian, Johnsonian, and contemporary English are each different 'états de la langue anglaise'."

However, these terms are not actually in widespread use even in linguistics books, and rarely feature in their indexes. Svensen (1993:18) discusses them in terms of dictionaries: "2.4.1.2 Synchronic and diachronic, historical and contemporary: A dictionary can be synchronic, describing the language as used during a limited period, or it can be diachronic, describing the way in which the language has developed during a longer period of its history. A dictionary can also be historical or contemporary. Examples of synchronic contemporary dictionaries are the Longman Dictionary Of Contemporary English and Collins COBUILD English Language Dictionary, which focus on the language of the present day. The diachronic historical dictionaries include, first of all, the great national dictionaries: the Oxford English Dictionary, the Grimm's Deutsches Wörterbuch, Trésor de la langue française, Woordenboek der Nederlandsche Taal, and so on; see also Merkin 1983. Many monolingual contemporary dictionaries try to cover these various aspects by introducing also an etymological component (see Chapter 15)." Apart from a slightly infelicitous conflation of ideas which leaves the reader wondering whether *synchronic* and *diachronic* are merely synonyms of *contemporary* and *historical*, the two terms are at least distinguished from each other.

Stubbs (1996) does not include the terms *synchronic* and *diachronic* in his index. But he does give several examples of both synchronic and diachronic linguistic studies: "the Brown and LOB corpora contain only texts from 1961" (p. xviii); "Scannell (1986) discusses the development of documentary programmes on BBC radio from the 1920s to the 1940s" (p. 13); "Jucker (1992) provides a detailed analysis of a sample of British newspapers based on ... a well defined corpus. The corpus comprises samples from a few days in 1987-8, of all eleven British national daily newspapers".

However, the examples cited by Stubbs also serve to point to a problem in the terminology for corpus linguistics. I would call the LOB, Brown, and Jucker corpora *synchronic*, yet LOB and Brown cover a whole year, whereas Jucker covers only "a few days". The Scannell study has to be labelled *diachronic* because it specifically deals with the 'development' of documentary programmes from the 1920s to the 1940s, yet I would want to call the original 18 million word Cobuild corpus (also referred to as the Birmingham Collection of English Text) *synchronic*, although it contains texts from a similar span of time, i.e. roughly 20 years, 1960s-1980s (Sinclair 1987:33).

4. Synchronic and Diachronic corpora

It seems to me that the terms *synchronic* and *diachronic* have a slightly different localized operational value or functional significance when they are applied to corpora. The terms are relevant not so much with reference to the period of time within which the corpus texts were produced, but rather to the way in which the texts can be accessed. If the corpus can be accessed only as a single entity, then it is functionally *synchronic*, whether the component texts were produced on the same day, within the same year, or even within the same century, because there is no possibility of studying the development of language during that day, year or century. If the corpus texts are held in such a way that texts from a particular period of time can be accessed as a separate and discrete group, then the corpus is functionally *diachronic*. We can compare April texts with November texts, or texts from the first decade of the century with texts from the final decade. Crucially, we can observe and comment on language change.

Thus, the Brown and LOB corpora cannot be labelled *synchronic* simply because we know that they contain texts exclusively from 1961. They can be labelled *synchronic* or *diachronic* for corpus linguistic purposes only after we discover whether texts from January can be inspected separately from February texts - or Spring texts from Autumn texts. The exact subdivisions are not important, the fact of subdivision or the possibility of subdivision is the key criterion.

In her survey of electronic corpora, Edwards (1992) labels only the Helsinki Corpus as *diachronic*, out of all the various data resources she describes (even though the London-Lund Corpus is listed as containing data from the "1960s and early 1970s", and the Lancaster Spoken English Corpus has data gathered between 1984 and 1987). Without making the point explicit as I have done above, she clearly makes the same distinction: Helsinki is *diachronic* because it contains texts "from periods at roughly 100-year intervals beginning in 850" and (more importantly, in my opinion) because "It is used for variational study of the development of English", that is, texts from the different periods are accessible independently.

5. Cobuild corpora: synchronic access only: 1980 to 1995

The Cobuild project started in 1980, and data collected until 1986 was merged to form the Birmingham Collection of English Text (BCET), a corpus of approximately 18 million words. The amounts and proportions of data from different periods are given in Table 1.

Table 1: Birmingham Collection of English Text (BCET): 1986

Corpus	pre-1960	1960-69	1970-79	1980-86
BCET (1986): 18m words	1.3m = 7.5%	2.7m = 15%	9m = 50%	5m = 28%

In 1991, a new corpus-building initiative began collecting data for the Bank of English (BoE) corpus, which reached 120 million words in 1993, 167 million words in 1994, 211 million words in 1995, 323 million words in 1996, and 418 million words in 2000. The principles behind the BoE updates are (wherever possible) to increase the size, increase the variety of sources, and maintain currency by replacing older data with newer data. Table 2 shows the composition of successive Bank of English corpora from 1993 onwards.

Table 2: Cobuild Bank of English Corpus: contents and dates

Bank of English Corpus	**1993**	**1994**	**1995**	**1996**	**2000**
	120m words	**167m words**	**211m words**	**323m words**	**415m words**
Sub-corpora					
American Academic textbooks	-----	-----	-------	---------	6m (1990-96)
American Books	16m (post-1985)	10m (1987-91)	19.4m (1987-94)	32.66m (1987-95)	32m (1987-95)
American Ephemera	-----	-----	-------	1m (1995)	3.5m (1995-96)
Wall St Journal	6m (1989)	6m (1989)	6.2m (1989)	-------	------
American newspapers	-----	-----	-------	8.58m (1989-94)	10m (1989-96)
American Radio (NPR)	10m (1990-91)	21m (1990-93)	22.3m (1990-93)	22.26m (1990-93)	22m (1990-93)
American Spoken	-----	-----	-------	--------	2m (1994-97)
BBC World Service	20m (1990-91)	18m (1990-91)	18.7m (1990-91)	18.52m (1990-91)	18.5m (1990-91)
British Books	31m (post-1985)	27m (1985-92)	27.8m	42.13m (post-1990)	43m (post-1990)
British Ephemera	1m (1991-92)	1.5m (1991-93)	1.9m (1991-94)	4.72m (1991-95)	4.5m (1991-96)
British Magazines	5m (1992)	28m (1992-93)	30.1m (1992-93)	30.14m (1992-93)	44m (1992-2000)
British Spoken	4m (1991-92)	8.5m (1991-93)	15.5m (1991-94)	20.18m (1991-96)	20m (1991-96)

Economist	3m (1991)	7m (1991-93)	8.7m (1991-94)	12.13m (1991-95)	15.5m (1991-99)
Guardian newspaper	-----	12m (1993)	12.6m	24.26m (1995)	32m (1995-99)
Independent	5m (1990)	5m (1990)	5m (1990)	19.45m (1990-95)	30m (1990-99)
New Scientist	-----	3m (1992-93)	4.2m (1992-93)	6.09m (1992-95)	7.9m (1992-99)
Times newspaper	10m (1992)	10m (1992)	10.4m (1992)	20.95m (1995-96)	30m (1995-2000)
Today newspaper	10m (1991-93)	10m (1991-93)	18.1m (1991-93)	26.61m (1992-95)	26m (1992-95)
Australian Regional newspapers	-----	-----	10.3m (1993-94)	33.38m (1994-95)	34m (1995-99)
Sun and News of the World	-----	-----	-------	---------	31m (1997-2000)

The exact proportions of data from each year were difficult to ascertain for the BoE from 1993 to 1995, because during this period an administrative/bibliographic database was being designed, created and tested, into which the publication/recording date and other information about each text was gradually being input. But whereas it was always possible to establish the date of a particular text (assuming it was known) in the Cobuild data, it was not possible to select all texts from a particular date, partly because the information was either not encoded in the corpus at all, or was encoded in different ways for different texts. The newspaper and radio subcorpora were the most easily datable, as the data was supplied to us in periodic batches, added to the corpus in periodic batches, and date information was already encoded by the suppliers in the header information for each article.

Table 3: Bank of English corpora: proportions of easily datable data

Corpus	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**	**1992**	**1993**	**1994**
BoE (1993): 120m wor **(c. 75m easily datable**					6m =8%	24m =32%	18m =24%	25m =34%	2m =3%	
BoE (1994): 167m wor **(c. 161m easily databl**	1m	2m =1%	3m =2%	2m =1%	13m =8%	34m =21%	22m =14%	37m =23%	47m =28%	
BoE (1995): 211m wor **(c. 200m easily databl**	1m	2m =1%	3m =2%	2m =1%	12m =6%	33m =16%	21m =10%	50m =25%	61m =30%	15m =8%

It may be worth noting here (as the term has some potential bearing on the discussion of synchronic and diachronic corpora) that some people refer to the BoE as a *monitor corpus*. In fact, in my view, the original concept of a monitor corpus seems to have been slightly different (Sinclair 1982; Clear 1986): a batch of stable data (e.g. 100 million words of the Times newspaper for 1990) would be analysed for word frequencies, collocations, etc, and comparable new batches (e.g. monthly batches of the Times from 1991) would be analysed and compared with reference to the stable batch. New words, changes in frequency or collocation, etc could be noted and retained, but data which merely added more examples of known linguistic facts could be discarded. Eventually, say at the end of 1991, a new stable batch of 1991 data could be matched in the same way with monthly batches from 1992. This was roughly the procedure adopted by the Aviator project at the Research and Development Unit for English Language at the University of Birmingham, although the term *monitor corpus* seems to have been replaced by *dynamic data* in their relevant publications (Renouf 1993; Collier 1993; Blackwell 1993).

Sinclair (1994:11) describes a monitor corpus thus: "…a corpus of constant size… which would be constantly refreshed with new material, while equivalent quantities of old material would be removed to archival storage. The composition of the corpus would also remain parallel to its previous states. … The language would flow through the machine, so that at any one time there would be a good sample available, comparable to its previous and future states. Such a model … added another dimension to contemporary corpora - the diachronic." Table 2 shows that while

the Bank of English is a *dynamic* corpus, it is not strictly a *monitor* corpus: its size increases, its composition changes, and archival materials are not easily recovered for comparison.

6. Cobuild corpora: early diachronic research: comparing BCET and BoE

The data in the BCET was only accessible as a unified whole in terms of vintage, as was the data in the BoE corpora of 1993, 1994, and 1995 (although as Table 2 shows, all BoE corpora were held as a number of subcorpora in terms of source, i.e. Times data was held as a subcorpus, BBC World Service data as a subcorpus, etc, which enabled comparisons between written and spoken data, British, American and Australian data, broadsheet and tabloid newspapers, etc). Therefore, any diachronic comparisons made in research until 1995 tended to simplify the vintage issue and imply that the BCET data was all from 1986, the BoE (1993) corpus data was all from 1993, the BoE (1994) corpus data was all from 1994, and so on (whereas the tables above show that in fact the BoE (1993) corpus consisted mostly of data from 1989-1993, the BoE (1994) corpus data was mostly from 1989-1994, etc).

While teaching Corpus Lexicography on an MA course at Birmingham University in 1993, I prepared materials comparing evidence from the BCET and the 100-million-word sample from the nascent Bank of English, focussing on the exploitation of the proverb *A new broom sweeps clean*. The term *new broom* had been attested in the BCET as referring to a person: e.g. *the new broom may still be sweeping around*, and *some new broom's bright idea*. In the new data for the BoE, I noticed that *new broom* also referred to policies: e.g. *Heseltine's new broom has yet to show its bristles*, and that other variations such as the separation of *new* and *broom* and the omission of *new* were now evident: e.g. *…the new chief has wielded the broom through the business …as the Kinnock broom swept clean in Walworth Road …Bishko's rationalisation broom has swept through the 253-strong operation.*

In a 1995 session of the same course, I showed how the variable phrase *<VERB> yourself into a corner* (meaning to place yourself in a difficult situation by your own foolish actions) had extended its range of verbs since 1986. The BCET evidence for *talk, work,* and *paint* had been joined in the BoE data by *pen, box, back, force, put, barricade, dig,* and others.

In Krishnamurthy (1995) I made various comparisons between the 16.78 million word written component of the BCET (referred to as CORPUS 1) and the BoE (1993) 120 million word corpus (CORPUS 2): "The top ten items in the frequency lists for CORPUS1 and CORPUS2 are ... It is interesting to note that 8 out of the ten words in the lists are the same. However much we increase the size of the corpus, the commonest words seem to stay in roughly the same relative position ...". And later, "Of the five hyphenated items in the poem, only 'looking-glass' is well-attested in the corpora: 13 occurrences in CORPUS 1, 46 in CORPUS 2". Note that in this research it was the change in size, rather than vintage, that I drew attention to (although some of the differences may in fact have been due to the change of vintage).

Speaking on the topic of Language Change at a conference in Portugal in 1995, I used Cobuild data with more explicitly vintage-oriented focus (rounding the dates and figures for convenience: BCET (1986) of 18m words becomes 1985 and 20m words, BoE (1995) of 211m words becomes

200m words). Table 4 shows that increasing the corpus size by a factor of about 10 did not alter the rank or proportional frequency very much for the most frequent words in the language: *the* is top of the frequency list for both corpora, and occurs roughly once every 20 words in both; similarly, *it* is 9th in the BCET list and 10th in the BoE (1995) list, with a frequency of 180,000 in BCET and 1.6m in BoE (1995), and occurs roughly once every 100 words in both corpora.

Table 4: Word Frequencies

1985		**1995**	
20 million words		**200 million words**	
the	1023506	the	9900535
of	503284	of	4579352
and	475864	to	4340863
to	448378	and	4101538
a	388354	a	3752112
in	311996	in	3223675
that	190007	that	1879061
was	186792	's	1719686
it	180642	is	1636192
I	170090	it	1598918

Table 5 shows that language change, when viewed in terms of new words, can be identified even in a short period of time like 10 years, and items that did not occur (or occurred very rarely) in 1985 corpus data can be seen to occur (or to have increased significantly in frequency) in 1995 corpus data. The tenfold increase in corpus size during this period of time might account for increases in frequencies up to a factor of 10 or so, but cannot explain why words like *cyborg* and *gopher* have increased in frequency by more than 15 times, or words like *imaging* by nearly 70 times. This list of words from the domain of technology was culled from a much larger list of all the words in the BoE (1995) corpus that had not occurred in the BCET.

Table 5: New Technology: New Words

	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

In Krishnamurthy (1996) I made the comparison between BCET and BoE (1995) explicitly in relation to Cobuild dictionary entries: "CCELD does not have an entry for this item (*overstrung* meaning 'nervous, tense, excited', found in other dictionaries), because there was no evidence for it in the 20 million word corpus. In fact it is also totally absent from the 200 million word Bank of English, so its omission was amply justified." As corpus sizes increase even further, negative evidence of this kind becomes increasingly valuable, and is much harder for corpus sceptics to discount.

7. Cobuild corpora: micro-diachronic research from 1996 on

Corpus-building has always taken place at Cobuild with the main emphasis on providing large amounts of up-to-date material for lexicographers to analyse, with a view to the compiling of dictionaries and other reference books. Less attention had been paid to the needs of academic researchers in terms of corpus administration and encoding. Also, as corpus-building is just one of many tasks performed by Cobuild staff, several people had been involved periodically, intermittently, and on a part-time basis in the construction of the corpora. Therefore until 1995, corpus text references had been invented and assigned on a rather ad-hoc basis. Alphabetic characters had been used in text references as a mnemonic for the subcorpus to which the text had been assigned, because knowing the subcorpus of origin was of great help to lexicographers in the selection of dictionary examples and the assigning of style labels for word senses. But the numbers used in text references were often arbitrary or inconsistent. In Krishnamurthy (1996), I showed how text references were displayed on the computer screen:

Table 6: Corpus Concordances with Text References

```
gua00009072   pollution from mass exploitation".<LTH> But human
tim00080892   his successors were exploiting a vast mass of
tim00040992     wicked state that exploits and represses the
masses
gua00009071  class and one set of exploiters. And as mass parties
npr00111092   a mass murderer who exploited Indians # The
npr00071290      artist to really exploit mass media images in
gua00009071     but the anonymous exploited masses, including
tim00050992      about capitalist exploitation of the masses?
```

If one knows the constituent subcorpora of the BoE, it is fairly easy to deduce that ‘gua’ means ‘Guardian newspaper’, ‘tim’ is ‘Times’, and ‘npr’ is National Public Radio (Washington USA). However, it is not clear
whether in ‘gua0009072’ and ‘tim00080892’ the numbers are arbitrary or meaningful. The latter might well mean ‘Times, 08/08/92’, but the former could not possibly mean ‘Guardian 00/90/72’. However, ‘npr00111092’ could again be ‘NPR, 11/10/92’.

Among the administrative changes made in June 1996, during the annual BoE corpus update, one was to incorporate the date information into the text reference number for all newspaper and radio subcorpora, in a standardized format. For example, the Times for 17th April 1995 would have the reference number N2000950417, a BBC World Service broadcast of 3rd September 1991 would have the text reference number S1000910903, etc, where the first five digits identify the data source (N2000 for the Times, S1000 for the BBC World Service, etc) and the last six digits represent the year, month, and day.

This principle could, with hindsight, have usefully been applied to all corpus texts, and this would have provided a robust and comprehensive date-selection facility for the whole corpus. However, the advantages were not apparent at the time. The facility to use the text reference number as a search key had been part of the Cobuild corpus retrieval software for many years, but had not been of significant value. However, once the text references had been systematized, it became a simple task to isolate occurrences of a word from texts published or broadcast in the year 1995, by specifying in a search request that 9 and 5 should be the 6th and 7th characters from the beginning of the line.

By a simple change in the text referencing system, therefore, it suddenly became possible to do more detailed micro-diachronic analyses. In the BoE (1996) corpus, about 200 million words were easily accessible by date.

Table 7: Bank of English Corpus (1996): proportion of data accessible by date

BoE (1996): 323m word	**1989**	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**
c. 200m words accessibl by date	4m =2%	24m =12%	18m =9%	23m =12%	6m =3%	14m =7%	90m =45%	14m =7%

The request for a series of journalistic articles relating to the Bank of English enabled me to pursue my micro-diachronic research in various ways, in different linguistic domains, and the facility to quote authoritative annual figures indicating small-scale changes in language proved very valuable. A humorous economics article in connection with Budget Day contained the suggestion that it might be possible to predict the Chancellor of the Exchequer's forthcoming measures by looking at the corpus. The terms *bull market* and *bear market* were selected, and their frequencies for the period 1989-1996 (the rough span of the easily date-retrievable data in the BoE (1996) corpus) were scrutinized. During this period, the expressions ‘bull+market|markets’ had occurred 2.3 times more often (283 to 123) than

'bear+market/markets'. Table 8 shows the occurrence rate for both expressions for each year from 1989 to 1996.

Table 8: Economics: *bull market* and *bear market*

bull+market|markets : 283 matching lines							
Distribution by date: occurrences per year							
1989	1990	1991	1992	1993	1994	1995	1996
57	13	23	8	6	19	76	41

bear+market|markets : 123 matching lines							
Distribution by date: occurrences per year							
1989	1990	1991	1992	1993	1994	1995	1996
29	5	7	19	12	3	14	10

At this stage, I used raw frequency figures, and made no allowance for the different amounts of data for each year.
I then looked at *bullish* and *bearish* (where the ratio of occurrence was even greater at 3:1) and concluded that the *bears* would fare better than the *bulls* in the Chancellor's speech, if economic balance was his aim.

This article prompted several enquiries from the financial press, who responded well to this type of statistical analysis, especially when it was backed up with graphs and tables, so I produced a more elaborate analysis which discussed 'hard indicators' (analysing the lexical items *consumer spending, inflation, house prices, negative equity, taxation, employment statistics, exchange rates, growth rates, share prices,* etc) and 'soft indicators' (*feelgood factor, consumer confidence, poverty, confidence, happiness, job security/insecurity, stress levels,* etc). However, one or two journalists began to take this material too seriously ('Surely you are not suggesting that a language corpus can be an economic predictor' asked one interviewer, prickling at the thought) and swallowed my tongue-in-cheek
explanation that, after all, there is an economic prediction system based on astrological data which is used by some stockbrokers, so why should corpus data not be used for economic forecasting? It later dawned on me that there might have been a better linguistic justification: we are constantly reminded by economic pundits that the economy is all a matter of confidence and perceptions and hunches, so surely the use of language must play a crucial role in 'talking up' or 'talking down' the economy.

A Remembrance Day article looked at the way in which words and phrases that had originated in war situations were being used in casual contexts, perhaps even trivializing the original

circumstances and offending service personnel. From a general survey of military concepts such as *attack* and *defence*, *victory* and *defeat* and their use in sports journalism, the focus switched to items such as *bikini, shell shock, go over the top, exocet,* etc and the phrase particularly associated in Britain with Remembrance Day: *lest we forget*.

The topic 'Are dialect words increasing in mainstream use?' highlighted the research problem raised earlier (see the reference to Krishnamurthy (1995)): dialect words that did not appear in BCET were evident in BoE (1996), but whether this was due solely to the increase in corpus size, or reflected a genuine increase in the rate of usage, was hard to judge.

The General Election of 1997 yielded a rich harvest. Some key terms from the political debates were very frequent in the BoE (1996) corpus but were not found in the BCET corpus at all. See Table 9.

Table 9: Politics: terms in the BoE (1996) corpus but not in the BCET

YEAR	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**
Data accessible by date	24m	18m	23m	6m	14m	90m	14m
Distribution by date: occurrences per year							
additionality	1	0	0	0	0	8	0
communautaire	1	1	3	8	1	5	3
eurosceptic(s)	1	0	8	3	28	320	85
fundholding/-er(s)	0	0	15	1	49	412	39
majorite(s)	0	1	3	7	3	23	2
majorism(s)	1	3	15	2	3	23	2
quangocracy	0	0	0	0	1	11	0
subsidiarity	15	24	93	16	8	72	10

Other political terms had occurred in BCET, but had dramatically changed their corpus frequency between 1986 and 1996. The concern to distinguish between words which had increased in frequency purely because of the increase in corpus size, and words which had genuinely increased in use, led to the introduction of a crude arithmetical algorithm. Table 10 shows a selection of the words analysed, including some words which were chosen because of social changes known to have taken place, e.g. the introduction of the TESSA savings schemes, the National Lottery (*lottery, rollover*), and social concerns (*dysfunctional*). The items range from *maastricht* (which occurred 91 times more than expected from a roughly 18-fold increase in corpus size) to *privatization* (which increased 3 times more than expected). Note that the Cobuild software changes all capital letters into lower case in frequency lists (e.g. Maastricht > maastricht, TESSA/Tessa > tessa).

Table 10: Politics and Society: words that greatly increased in frequency from BCET to BoE (1996)

Word	**A**	**B = (A x 18)**	**C**	**D = C:B**

	Actual frequency in BCET: 18m words	Expected frequency in a 324m corpus	Actual frequency in BoE (1996): 323m words	Ratio of actual : expected frequency for 323m words
maastricht	3	54	4941	91
dysfunctional	1	18	946	53
tessa(s)	3	54	2085	39
lottery	28	504	9522	19
underclass(es)	3	54	872	16
rollover(s)	2	36	218	6
cronyism	1	18	81	5
privatization	16	288	839	3

The word *sleaze* dominated the entire pre-election period, and I discovered it was a new coinage that was roughly coeval with the the advent of the Tory administration of Mrs Thatcher. The OED (1971) did not have the noun, although the adjective *sleazy* (originally from 'Silesia' and referring to a type of thin, flimsy, insubstantial cloth) is attested since the 17th century. Similarly, there were no occurrences of *sleaze* in the 1986 BCET corpus, but there were 17 occurrences for *sleazy* and its derivatives, used in its modern sense of 'shabby and disreputable' to describe places: *a sleazy area, a sleazy attic, a sleazy cafe, a sleazy-looking hotel*, etc.

Mrs Thatcher had come to power in the 1979 election, and *sleaze* was entered as a headword in the Collins English Dictionary (1986 edition), and even in the Collins Gem English Dictionary (1985-7), its inclusion in such a small dictionary proving that it had become a mainstream word. Since then, the word had experienced an astonishing increase in circulation within a very short time: from 0.6 occurrences per million words in 1990 Cobuild data, to 1.9 per million in 1992, to 8.7 per million in 1995. The use of "occurrences per million words" rather than raw frequencies in my research was yet another attempt to distinguish increased occurrence due to increase in corpus size from genuine increase in use. The analyses of sleaze and related words also progressed beyond mere frequency to include collocational profiles: *sleaze* collocated strongly with *Tory/Tories, government* (i.e. the Tory one), *scandal, greed, corruption*, etc. The 1997 General Election also prompted research on the names of the main politicians and parties, the party manifestos, etc using collocational methodology.

I had embarked on a piece of research on British and American English for a Wordwatch article on Cobuild's website. Looking closely at concordances for *have a bath* and *take a bath*, I accidentally discovered a new restricted meaning. In the business world, `taking a bath' obviously means losing a lot of money:
...those Japanese who took a bath in Bombay.
Shareholders have already taken a bath.
Investors announced that they were taking a bath.
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.
I continued this investigation in an article about Americanisms and the influence of American English on the British variety, obviously a diachronic issue. Unfortunately, the amounts of data analysable at this level of detail are relatively small in most cases (see Table 11): the American data accessible by date reaches a peak of 8.2 million words in 1992; the British data accessible by date drops to a low of 2.5 million words in 1993. And at these low volumes of data, the specific sources of the data (i.e. precisely which newspaper or radio material was accessible) may play a large part, may skew the analyses. Even worse, no American data at all was accessible by date for 1995 and 1996, so we can only guess what might have happened from the general patterns. However, the general patterns are, surprisingly, fairly clear.

All the selected items (chosen initially by lexicographers using their intuition) had a higher occurrence rate per million words in American data than in British data in 1990, i.e. they were used more commonly in American English then. The most extreme example is *downsize* and its inflections, which occurred 2.8 times per million words in American data in 1990, and 0.0 times per million words in British data. The difference in occurrence rates is very small for *fat cat(s)* (0.4 in American and 0.3 in British) and *a raft of* (0.8 and 0.5). The only exceptions are *feisty* (which was used 0.4 times per million words in both American and British data in 1990), and *politically correct* and *political correctness* (which both occurred 0.0 times per million words in both American and British data in 1990). None of the items was more common in British English in 1990.

In each case, we are able to see an increase in both American and British usage after 1990. In many cases, we are also able to see a fall in usage in later years. For example, *fat cat(s)* starts at 0.4 in American data, rises to 1.2 in 1992 an 1993, then falls back to 0.0 in 1994; and starts at 0.3 in British data, rises to 16.3 in 1995, and drops to 2.3 in 1996.

The peaks in usage rates occur in earlier years in American data than in British, e.g. *stash* and its inflections are already at their peak in American data of 2.8 in 1990 and decline steadily each year (but note a slight resurgence in 1994), whereas they do not reach their peak in British data until 1993-4. In the analyses, it is just about possible to discern a transition point, where the words begin to be used more frequently in British English than in American English (e.g. 1993-4 for *fat cat/cats*; 1994 -5 for *feisty*; 1990-1 for *a raft of* and *stash*).

Table 11: Variety: micro-diachronic comparisons of American and British English

Data accessible by date:							
	1990	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**
American English	2.5m	7.9m	8.2m	3.4m	2.5m	0	0
British English	19m	8m	14m	2.5m	11m	58m	13m
Expressions:	**Occurrences per million words:**						

fat cat(s)							
American English	0.4	0.6	1.2	1.2	0	----	----
British English	0.3	0.5	1.1	0.8	7.9	16.3	2.3
feisty							
American English	0.4	1.3	3.5	2.7	5.2	----	----
British English	0.4	0.6	1.1	1.2	3.5	2.5	2.8
a raft of							
American English	0.8	0.1	0.9	0.6	0	----	----
British English	0.5	1.9	0.9	2.8	1.2	2.9	3.1
stash(-es, -ed, -ing)							
American English	2.4	1.7	1.5	1.2	2.0	----	----
British English	0.8	1.8	2.9	3.2	3.2	2.0	1.7
politically correct							
American English	0	3.9	3.3	3.5	1.6	----	----
British English	0	2.5	2.6	7.6	6.9	5.9	5.8
political correctness							
American English	0	0.8	1.7	6.8	2.8	----	----
British English	0	0.6	1.2	4.0	4.0	5.9	8.3
downsize(-s, -ed, -in							
American English	2.8	2.4	5.7	5.0	1.2	----	----
British English	0	0.1	0.2	0.8	0.7	3.0	2.9
fast track							
American English	1.6	5.1	1.2	4.7	1.2	----	----
British English	0.8	7.2	1.9	11.2	1.5	3.6	3.5

The patterns for *politically correct* and *political correctness* follow the broad outlines described above, but there is also an interesting shift from the adjectival phrase *politically correct* to the nominalization *political correctness* (cf. the shift from adjective *sleazy* to noun *sleaze* discussed earlier). This shift occurs within each variety: in American data in 1993; in British data in 1996.

Table 12: Grammatical processes: from adjective to noun

	1990	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**
politically correct							
American English	0	3.9	3.3	3.5	1.6	----	----
political correctness							

American English	0	0.8	1.7	6.8	2.8	----	----
politically correct							
British English	0	2.5	2.6	7.6	6.9	5.9	5.8
political correctness							
British English	0	0.6	1.2	4.0	4.0	5.9	8.3

8. Conclusion

Language change is a topic eminently suitable for investigation using corpora. Identification and measurement of the changes is possible with current corpus retrieval software. One of the defining features of a corpus might be regarded as the ability to view all the data in the corpus using the same set of software tools. Accordingly, the terms *synchronic* and *diachronic* have been redefined for the purpose of corpus analysis, as dependent on the ability of the software to view corpus texts according to their date.

The diachronic research using synchronic corpora reported in this paper started with comparisons between the BCET and BoE corpora, roughly representing 1980s and 1990s English respectively. The progress to the micro-diachronic level depended on a simple change in administrative practices, in terms of incorporating text dates into the text reference numbers for certain data types (newspapers and radio broadcasts).

As the research technique progressed various correctives, such as arithmetical calculations of expected occurrences, and using 'occurrences per million words' rather than raw frequencies, were introduced into the analyses to cope with the disproportions in the amount of data available for each date division (in this case, a year). What began as a rather casual and secondary type of analysis, stimulated by journalistic and publicity motivations, has begun to take shape as a valid methodological insight into the observation and measurement of language change. Most significantly, the tools and techniques that would be required to bring the patterns of change into sharper and more critical linguistic focus have been explored.

The research in this paper has concentrated on frequency of occurrence of individual lexical items, compounds, and collocates. Researching changes in collocational profiles constitutes an even more complex task, and has not been reported here.

The same techniques could however be applied to look at changes in the occurrence rates of many other linguistic features, such as syntactic patterns. In the BCET data, examples were noticed for *persuade (someone) to (do something)*, a pattern which the verb *persuade* had assimilated from *tell*, *convince*, etc. This had also given rise to a few examples for the associated pattern *persuade (someone) into (a place/an action):*

only a little nudging and pushing to persuade them into the starting-stalls
it appears, had any duty to persuade inmates into such a course of
Freda so well that I'm sure I can persuade them into some suitable sort of
the animals were heavily goaded to persuade them into the Weinberg pen and

In the BoE (1995) data, we find confirmation of the pattern *persuade (someone) into (a place/an action),* but also evidence for the addition of a strong secondary pattern *persuade (someone) into (doing something),* which *persuade* has acquired by analogy with verbs such as *talk, cajole,* and *trick*:

former England bowler, returned to persuade Lloyd into a loose stroke.
that he would find it easier to persuade her into an occult order that
that iron hand gently but firmly persuade him into a nice, open apology to
spite of her parents" efforts to persuade her into a wig, Laura <FCH> still
the maverick politician might be persuaded back into ZANU. But it now looks
so much that now women have to be persuaded back into it. There are many
it is hoped Chief Buthelezi can be persuaded back into the negotiating
his quest for synthesis sometimes persuaded Yeats into a dogged forcing of
get away with too much, and then persuading them into your bed with high
in June that Coakley's was persuading clients into unadvisable
is a signal to the west to persuade Israel into freeing its Lebanese
he was able to coerce or cajole or persuade them into using in their best in
Irpen:<FCH> Some new acquaintances persuaded us into spending the summer near
the foreign secretary, reputedly persuades Thatcher into committing Britain

The amount of evidence is very small when compared with amounts available for the investigation of more frequently occurring linguistic features such as lexical items. However, with the continual increase in corpus sizes, this kind of investigation should also gradually become more feasible and more reliable.

References

Blackwell, S. (1993) From Dirty Data to Clean Language, in Aarts,J., Haan, P. de, Oostdijk, N. (eds.) E*nglish Language Corpora: Design, Analysis and Exploitation*, Rodopi, Amsterdam

Clear, J. (1986) Trawling the Language: Monitor Corpora, in Snell-Hornby, M. (ed.) *ZuriLex'86 Proceedings*, A Francke Verlag, Tübingen/Basel

Collier, A. (1993) Issues of Large-Scale Collocational Analysis, in Aarts,J., Haan, P. de, Oostdijk, N. (eds.) *English Language Corpora: Design, Analysis and Exploitation*, Rodopi, Amsterdam

Crystal, D. (1997) *English as a Global Language*, Cambridge University Press, Cambridge

DeWitt, N.J. (1966) *Horace - Ars Poetica*, Dell, USA

Edwards, J.A. (1992) Survey of Electronic Corpora and Related Resources for Language Researchers, in Edwards, J.A. and Lampert, M.D. (eds) *Talking Data: Transcription and Coding in Discourse Research*, Erlbaum, NJ (also available in electronic form via ftp from 128.32.211.5)

Fowler, H.W., and Fowler, F.G. (1906) *The King's English*, Oxford University Press

Johnson, S. (1755) *Dictionary of the English Language*

Krishnamurthy, R. (1995) The Macrocosm and the Microcosm: The Corpus and The Text, in Payne, J. (ed.) *Linguistic Approaches to Literature: Papers in Literary Stylistics*, Discourse Analysis Monograph 17, English Language Research, University of Birmingham

Krishnamurthy, R. (1996) Exploiting the Masses: the corpus-based study of language, in Zettersten, A. and Pedersen, V.H. (eds.) *Symposium on Lexicography VII*, Lexicographica, Series Maior 76, Max Niemeyer Verlag, Tübingen
Murray, J. (1928) *Oxford English Dictionary*, Oxford University Press, Oxford
OED (1971) *Oxford English Dictionary*, 1979 Compact edition, Oxford University Press, Oxford
Renouf, A. (1993) A Word in Time: first findings from dynamic corpus investigation, in Aarts,J., Haan, P. de, Oostdijk, N. (eds.) *English Language Corpora: Design, Analysis and Exploitation*, Rodopi, Amsterdam
Robins, R.H. (1989) *General Linguistics*, Longman, Harlow
Sinclair, J. (1982) Reflections on computer corpora in English language research, in Johansson, S. (ed.) *Computer Corpora in English Language Research*, Norwegian Computing Centre for the Humanities, Bergen
Sinclair, J. (1987) (ed.) *Looking Up*, Collins ELT, London
Sinclair, J. (1994) Corpus Typology, EAGLES document EAG-CWG/Corptyp (available at http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/node19.html#SECTION00090000000000000000)
Stubbs, M. (1996) *Text and Corpus Analysis*, Blackwell, Oxford
Svensen, B. (1993) *Practical Lexicography - Principles and Methods of Dictionary-making*, Oxford University Press, Oxford
Times English Dictionary (2000), HarperCollins Publishers, Glasgow