The British National Corpus
Michael Rundell

A major new resource for the language industry

FORGOTTEN BLUE WATERS

For a few weeks in the autumn of 1994, the political columns of the British press contained dozens of references to the phrase 'Clear blue water'. The context was that Britain's 'beleaguered' Conservative government (another buzzword of the time) was engaged in an internal debate about whether to appeal to the centre ground, or whether to follow a more radical agenda and - as the protagonists of this view put it - put 'clear blue water' between themselves and their political rivals. Now, if you were to assemble a corpus from newspaper text of the period, the frequency statistics would give the impression that clear blue water was an exceptionally common phrase in English - far more common, say, than in the clear, once in a blue moon, or keep your head above water. It would not matter how large your corpus was. Indeed, the more newspaper text you poured into it, the more powerfully the statistics would argue their case. But the statistics would be wrong, for within a couple of months, 'clear blue water' was all but forgotten.

FUNDAMENTAL PRINCIPLES

All of which usefully illustrates some of the pitfalls of corpus development. If a corpus is to form the basis of an accurate description of a language, it must of course be large enough to yield statistically valid evidence for the 'regularities' of that language. But size in itself is no guarantee of reliability. However large the corpus, it cannot he used to make reliable generalisations about the way a language works unless it also makes a serious attempt to be representative. These are the fundamental principles that underpin the design and development of the British National Corpus, or 'BNC', a collection of written and spoken British text that is both large enough and balanced enough to form the basis for an authoritative description of contemporary British English.

THOUSANDS OF SOURCES

The BNC project, which was completed in 1994 after a three-year development period, is a major collaborative venture backed by the UK government and designed to assemble a 100-million-word sample of modern English. The 'BNC' written component consists of 90 million words of text, and has been carefully designed to provide samples from the whole repertoire of discourse types. It includes material from thousands of sources, ranging from literary novels to teenage magazines, newspapers to university textbooks, government leaflets to romantic fiction, taking in letters, junk mail, and advertising copy along the way. The result is an exceptionally rich and well-balanced picture of the language, and the sheer range of the corpus
safeguards it from the dangers of 'skewing' the negative effects on the quality of data that can result when too high a proportion of the material is drawn from just one or two text-genres.

**SPONTANEOUS SPOKEN ENGLISH**

While the written corpus applies an established methodology to the collection of large volumes of text (the 30-million-word Longman/Lancaster Corpus, for example, was assembled in the late 1980s on broadly similar principles), the spoken component of the BNC represents a revolutionary breakthrough. For a whole variety of reasons - practical, financial, and methodological - the spoken language has been seriously underrepresented in modern corpuses. Admittedly, broadcast speech (from radio interviews, news programmes, talk shows and the like) has been collected in reasonable quantities. But the problem is that it is always to some extent scripted, or at least pre-planned, and consequently - as research has shown (Svartvik 1992) - 'radiospeak' is actually closer to written text in its linguistic features than to spontaneous face-to-face conversation. What was missing until now was a really significant quantity of ordinary conversational English, and this is the gap that the BNC's spoken corpus aims to fill. The data was gathered by 'wiring up' a carefully-selected group of volunteers, who wore Walkman recorders for a two-week period and captured on tape every conversation they were involved in. The result is by far the largest database of spontaneous spoken English ever collected (see further Crowdy 1993).

**CATERING FOR THE USES**

The British National Corpus, then, with its carefully-balanced range of text types and its uniquely authentic spoken component, marks a major new development in corpus building. In the very near future it will be made available to researchers throughout the European Union working in every branch of information technology and the language industry, computational linguists, phonologists, students of stylistics, ELT coursebook writers, and discourse analysts to name just a few. But the first users of the corpus are the compilers of dictionaries, whose inquiries are for the most part lexically-oriented, that is, they need to be able to ask, and get satisfactory answers to, questions about individual words and phrases.

These are just a few of the kinds of questions the corpus can help to answer:

- **what is the plural of walkman, or of mouse in its computer sense?**
- **is it only journalists who use slam and rap to mean 'criticise'?**
- **are we right to label quest as a literal word, or is it more widely distributed than this?**
- **what are the most frequently used meanings of words such as sacrifice, sacrilege, and sin?**
- **how extensive is the influence of American English on British idioms For example, how far do British speakers use mad to mean angry, and do they use expressions like take a rain check**
and *get to first base*?

- what exactly are the semantic differences between *fix*, *mend*, and *repair*?

- what is the relative frequency of the various complimentation patterns used with verbs such as *decide*, *remember*, and *suppose*?

- are well-known idioms like *kick the bucket* and *raining cats and dogs* ever really used?

**INVESTIGATING ENGLISH PHRASEOLOGY**

For over ten years now, computerized corpuses have been informing the dictionary-making process, and the evidence they provide has led to dramatic improvements in our description of English. The arrival of the BNC marks a further major advance. In the first place, the volume and quality of the data enables lexicographers to answer their everyday inquiries with more confidence and more authority than ever before. Even more importantly, a large corpus reveals recurrent patterns that might not have been apparent in earlier, less comprehensive databases. This is especially true in the case of English phraseology, an area in which there is a growing interest (see for example Kjellmer 1991, Nattinger and DeCarrico 1992). Here the BNC supplies striking evidence showing the extent to which writers and speakers (especially speakers) rely on pre-assembled chunks' of language to encode a wide variety of very frequent concepts.

**'PREFABRICATED' PHRASES**

Take the following extract from the Spoken Corpus.

*You can always get some more I would think. Ah well, as he said, well he didn’t say, but I mean it’s just a question of putting them in the post.*

We see here, first, some familiar (and some not so familiar) fluency devices: 'I mean', 'ah well', and the less well-known (but actually very common) 'I would think'. Then there are expressions used to convey the speaker's own evaluation of the situation: 'you can always' implies 'this is something you could do if you have to', and 'it's just a question of' means 'it's easy - this is all you have to do'. (Similar expressions, like 'let's face it' and 'say what you like' are also well-attested). Over and above all this, there is impressive evidence to support our growing understanding of the way that most common concepts in the language can be - and frequently are - expressed through 'prefabricated' phrases rather than through single words, so for example, we often talk about getting 'something to eat' rather than just 'food', and we tend to say that someone is 'not particularly clever/not all that bright' rather than simply saying they are 'stupid'.

**IMPLICATIONS FOR THE FUTURE**

Even from this small sample of the kind of information the BNC is yielding, it should be clear
that the implications for everyone involved in English Language Teaching will be far-reaching. Pedagogical dictionaries have already benefited enormously from the new insights the BNC brings: the Longman Language Activator, for example, and the new edition of the Longman Dictionary of Contemporary English provide a more accurate and comprehensive account of spoken English than has ever been possible before. Pedagogical grammars and coursebooks will be the next to be enriched by the corpus, and in the years to come its benefits will be felt across the whole range of reference and teaching materials.

USEFUL REFERENCES:

Crowdy, Steve. 
"Spoken Corpus Design" 
*Literary and Linguistics Computing* 

Kjellmer, Goran. 
"A mint of phrases" 
*English Corpus Linguistics* 
(Karin Aijmer & Bengt Altenberg eds), 
Longman 1991, pp111-127

Nattinger, James and DeCarrico, Jeanette. 
*Lexical Phrases and Language Teaching* 

Svartvik, Jan. 
"Lexis in English language corpora" 
*Euralex '92 Proceedings* 
(Hannu Tommola et al. eds) 
University of Tampere 1992, pp 7-31.