100 Million Words of English: The British National Corpus (BNC)*

Geoffrey Neil Leech

1. A National Corpus Project

In the United Kingdom, we have recently started a project to compile a British National Corpus (BNC): a computer corpus of 100 million words of British English, written and spoken. My purpose here is to describe the design, compilation, and foreseen uses of this corpus. I hope this will be of interest to this present audience, since other countries, perhaps especially Korea, may learn from our experience of building a national corpus of the native language. Lessons can, of course, be learned not only from the successes, but from shortcomings of our work.

The British National Corpus (BNC) project is a collaboration between commercial and academic partners. The leading partner is Oxford University Press, and two other major publishers (Longaman Group Ltd and Chambers) are also contributing on the commercial side. On the academic side, the collaborators are the Oxford University Computing Service, and Lancaster University. The British Library (our national library) also takes a role in such areas as archiving the corpus material and making it available. The project is supported approximately 50 percent by the British Goverment (Department of Trade and Industry, and Science and Engineering Research Council). It began in January 1991 and will continue for three years. The corpus is intended to be representative of a very broad range of English language use in speech and writing.

* Fuller details of the background to current English language corpora are given in Leech (1991).

2. What will be the Uses of the Corpus?

Before a corpus is compiled, the question which many people ask is: What on earth will it be used for? However, previous experience suggests that the uses to which a corpus can be put are far more numerous and varied than the corpus compilers could have imagined. From a recently-pubilshed 40-page bibliography of publications making use of computer corpora of English (Altenberg, 1991) it can be seen that existing computer corpora have been used in many applications such as the following:

linguistic theory; probabilistic language modeling
computational linguistics; natural language processing
machine translation; language understanding systems
grammar; syntax, morphology, parsing
lexicology, lexicography, word formation
semantics
pragmatics
discourse analysis, conversational analysis
language variation; spoken and written language
language and gender
text typology
dialectology
literary stylistics
speech technology; speech science; phonology; prosodics
historical linguistics
language acquisition
psycholinguistics
orthography (punctuation, etc.)
applied linguistics; language testing

It can be argued that, whatever the aspect of language being studied, the resources of computer corpora can be of help; and no doubt, future uses of the BNC will go beyond the list above. In section 5 below I will elaborate on some of the major uses of the corpus as foreseen by its compilers.

3. The Background of Previous and Current Corpus Compilation

Since the development of computer corpora has only recently impinged on the consciousness of mainstream linguistics, it may help to place this topic briefly in its historical and contemporary context. The first computer corpus, to all intents and purposes, was the Brown Corpus (compiled at Brown University under the direction of Nelson Francis and Henry Kucera, and completed in 1964). The Brown Corpus consists of c. 1 million words of various types of texts, and is limited to written American English. In the 1970s, a British counterpart of the Brown Corpus was compiled at Lancaster (in the U. K.) with help from Oslo and Bergen (in Norway). This corpus, completed in 1978, is therefore known as the Lancaster-Oslo/Bergen Corpus (or the LOB Corpus for short). The LOB Corpus, like the Brown Corpus, consists of c. a million words of written language. It should be mentioned that a research communities throughout the world, and can be automatically searched and processed in various ways valuable for research. These are obvious advantages of a corpus of linguistic material which is stored electronically, rather than on paper.

The first computer corpus of spoken English was the London-Lund Corpus, consisting of c. 500,000 words of speech by British speakers. The corpus is trancribed in a detailed prosodic transcription, and has proved invaluable for many studies of the spoken language.

By the 1980s the technology of computerising texts had advanced, and there was widespread use of automated text input devices known as *optical character readers* (now generally termed "scanners"). In addition, much text already existed in machine-readable form (for example, electronic text produced through modern printing technology, or through word-processing). The first corpus to benefit from such technological advances was the Birmingham Collection of English Text, compiled under John Sinclair, with support from the publisher Collins. The best-known outcome of this corpus has been the *Collins COBUILD English Language Dictionary* (1987), the first English dictionary to be based systematically on material derived from a computer corpus.

Since then, the compilation of English language computer corpora and other electronic text collections has gathered pace: for example, Longman (the pubilsher), with Lancaster (the university), has recently compiled the

30-million-word Longman/Lancaster Corpus, containing American, British, and other varieties of written English. There are now many different initiatives, the latest (and in some ways the most ambitious) of which is the Linguistic Data Consortium, described by Mitchell Marcus during the present conference. Those who worked on the Brown and LOB Corpora thought just one million words an exceedingly large amount of data to put together. But for a newer generation of corpus makers, the 100 million words of the BNC does not seem excessive. At the same time, it must be remembered that this immense increase in the "going rate" for corpus size is due entirely to the electronic availability of vast quantities of *written* data. Spoken data is still in short supply, as it has to go through the laborious human process of transcription before it can be computerized. Thus there is an enormous imbalance between the amount of written and spoken corpus data available: something which is reflected in the composition of the BNC, of which only 10 million words at the most are likely to be of speech.

Among other new initiatives underway at the present time is the International Corpus of English (led by Sidney Greenbaum, London), which will consist of parallel subcorpora, each of a million words, from more than 15 countries in which English is the major first language or second language. Another is the ACL Data Collection Initiative (Mark Lieberman) in the U.S.A., and the Bank of English, a "dynamic" or "monitor" corpus, to be based in Birmingham (John Sinclair). The Oxford Text Archive and the International Computer Archive of Modern English (ICAME) (based in Bergen) have been in existence for a number of years as agencies for the archiving and distribution of computer corpora and electronic text generally. Against this present background of enormous expansion, it is regrettable but inevitable that legal restrictions, particularly those associated with copyright and confidentiality, impede the public availabilty of most corpus material (the Brown, LOB and London-Lund Corpora being those which have been most widely used and distributed). The Longman/Lancaster Corpus is now available for distribution under agreement to academic users. One of the aims of the BNC project is to overcome these problems of availbility.

4. The Main Tasks in Compiling the BNC

The main tasks of corpus development can be listed as follows:

a. Corpus design
b. Acquisition and preparation of data
c. Corpus processing (e.g. adding and extracting linguistic information)
d. Making corpus material available to end-users

4.1. Corpus Design

The goal of designing a national corpus such as the BNC is to make it as far as possible representative of the full range of variation in the language. In practice, strict "representativeness" is very difficult to attain, and some believe the term cannot be meaningfully applied to corpora at all. However, efforts can reasonably be made to cover as broad a range of the language as possible in a balanced way. In keeping with this aim, the general design of the BNC is as follows:

a1 WRITTEN LANGUAGE COMPONENT: INFORMATIVE

PRIMARY SUBJECT FIELD (or DOMAINS)

Natural & pure science	Applied science
Social science	World affairs
Commerce & finance	Arts
Belief & thought	Leisure
Biography	

GENRE

Books	Miscellaneous (published)
Periodicals	Miscellaneous (unpublished)
Written to be spoken	

LEVEL

Specialist	Lay	Popular

DATE: 1975-present

a2 WRITTEN COMPONENT: IMAGINATIVE

GENRE

Narrative fiction	Playscript
Essay	Poetry

LEVEL

Literary	Middle	Popular

DATE: 1950-present

b1 SPOKEN COMPONENT: DEMOGRAPHIC SAMPL ING

Selection of 100-200 "subjects" who are native speakers of British English, sampled across:

- region
- age
- occupation
- educational/social background

b2 SPOKEN COMPONENT: LAYERED SAMPLING

Sampling across a range of discourse types:

Dialogue
- Private
  - Face-to-face: structured
  - Face-to-face: unstructured
  - Distanced
  - Classroom interaction
- Public
  - Broadcast discussion/debate
  - Legal proceedings

Monologue
- Commentaries
- Lectures/speeches
- Demonstrations
- Sermons

It should be clear that many practical considerations intervene between the ideal corpus one might wish to build if one had unlimited time and resources, and the corpus design which, in practice, one has to adopt. For

example, ideally one might like to have not only an equal quantity of spoken and written material, but a comparable classificatory breakdown of the spoken and written parts of the corpus. However, spoken and written language have to be treated differently both for reasons of expense and for reasons of sampling. It costs much more to collect a million words of spoken data than a million words of written data. Moreover, for sampling written texts, it is possible to use complete catalogues of publications in the U. K. during the relevant years. But there is no similar catalogue of all the conversations in which the population of the U.K. have engaged! So, to sample for the spoken (conversational) part of the corpus, quite a different method is unavoidably being adopted: namely a demographic method whereby individuals carefully sampled from the full range of the country's adult population are given a high-quality portable tape recorder, and instructed to record all the spoken discourse in which they engage over a given period of 2-7 days. This conversational part of the corpus (being compiled by Longman) is expected to yield uniquely rich and varied natural material for the study of the spoken language, even though the transcription will necessarily have to omit much of the detail found, for example, in the London-Lund Corpus.

4.2. Acquisition and Preparation of Data

Once the corpus has been designed in outline, the next step is clearly the acquisition of the textual data. In the case of the written language, the texts will mostly be acquired in electronic form, or else will be read into the computer via a scanner. In the case of spoken language, as already mentioned, it is necessary to undertake the laborious process of getting the texts recorded, edited, and transcribed. Clearance of the texts for copyright is another considerable task, which the corpus compilers are aiming to handle through national agreements.

Yet a further task is the preparation of the data for archiving and distribution. This involves the coding of all texts (spoken and written) in a particular standardized format. Up to now, electronic storage and interchange of natural language material has suffered from the lack of any commonly accepted and unambiguous system for identifying features of the original texts in the electronic record. Although normal standard alphanumeric

characters cause little trouble, there is a vast array of special symbols - for example, mathematical symbols, diacritics, non-roman character sets - which crop up in a few million words of written English texts, as well as special type faces, type sizes, and features of visual layout. In addition, each text or extract needs to be marked clearly with *header* information, to indicate its classification, prevenance, etc. An internationally agreed system for *marking up* electronic texts is now being devoloped under the Text Encoding Initiative, and the BNC is being coded and marked up in conformity with this system, so that users of the corpus, whoever they may be, will be able to reconstruct any required information about the texts and their formats. This work is being undertaken by our partners at Oxford.

4.3. Corpus Processing

The part of the work being undertaken at Lancaster is largely concerned with the next stage: that of corpus processing. A corpus of linguistic data is not necessarily well-adapted to use when it is in its "raw" orthographic form. Various kinds of linguistic information may need to be added: a process often referred to as *corpus annotation*. For example, for lexicographic purposes and many other applications, it is useful to have the corpus in a *grammatically-tagged* form, each word in each text being accompanied by a label or tag indicating its grammatical part of speech. (In this way, homographs and homonyms, such as *wind* (noun) and *wind* (verb), *set* (noun), *set* (infinitive verb), *set* (past tense verb) and *set* (past participle verb) are distinguished in the corpus.) Hence one major task we have to undertake at Lancaster is the grammatical tagging of the whole corpus. In addition, we will undertake other types of annotation of selected parts of the corpus. These include syntactic, semantic, and discoursal analysis.

In addition to this *insertion* of linguistic information, another type of corpus processing involves the *extraction* of such information. The second major task we have to undertake at Lancaster is to provide adaptable software for searching the corpus (or parts of it), and retrieving information or datasets in various forms which users will find valuable.

4.4. Making the Corpus Available

When the corpus is in other respects complete, there is one task which

must continue indefinitely: that of making the corpus (with its associated tools and annotations) available to users. The eventual aim will be to make the corpus as widely available as possible, given certain safeguards and conditions which have to be observed.

5. Some Applications of the Corpus

Finally, I return to the question of how the corpus will be used. In view of the large number of potential uses mentioned in 2 above, it will be convenient to focus on four areas of application which are likely to be important from the point of view of the corpus compilers.

5.1. Applications to Linguistic Research

There is a multitude of likely future applications in linguistic research. By way of example, I can only mention one or two fields in which I believe the BNC will provide new insight into, and knowledge of, language in general, and the English language in particular.

(a) *Lexis* The study of vocabulary is one area of linguistics for which a very large corpus is needed. In particular, this is required for the extensive study of collocational and idiomatic phenomena. Perhaps important light will also be thrown on the issue of productivity. Do we generate sentences and texts through the productive power of syntactic rules, or through the ability to stick together chunks of words which have previously been stored (see Pawley and Syder (1983))? I would be surprised if the study in depth of idiomatic chunks of language in corpora would not reveal that linguistic production is a mixture of these two modes.

(b) *Discourse Analysis* Already the London-Lund Corpus has provided a fertile resource for investigating phenomena of spoken discourse (e.g. use of conversational phrases, tag questions, turn-taking, pauses). But that corpus is unfortunately derived to a great extent from the discourse of the London area. The demographic corpus described in 4. 1. above will, in contrast, provide a much larger testbed for investigating conversational behaviour across a much wider ranger of language users.

(c) *Language Variation* In a number of important studies by Douglas

Biber and Ed Finegan (e.g. Biber 1988), existing corpora have been employed in the development of new techniques of analysis for text typology, using a statistical multi-feature/multi-dimensional methodology. The BNC, by providing a much larger and (in many ways) more balanced representation of the English laugage, will supply the means for developing research in language variation much further.

5.2. Applications to Reference Publishing

The involvement of leading dictionary publishers in this corpus speaks for itself. The corpus is bound to provide better information about the meaning, frequency, and use of words than has been available up to this point, and will therefore be a major resource for lexicography. Another area of linguistic publishing likely to become more and more thoroughly based on corpus evidence is that of reference grammars: this trend has already been demonstrated in recently-published grammars, notably Quirk et al. (1985) and the *Collins COBUILD English Grammar* (1990). Other types of corpus - based reference books - such as guides and thesauruses - may soon follow this tendency.

However, it is more interesting to think of the ways in which reference publishing may develop in new directions as a result of corpus development. One type of development is likely to be in the area of frequency dictionaries and collocation dictionaries: here corpora provide evidence of a kind which has been virtually unobtainable from other sources. One may also point to probable new developments in the combination of electronic reference pubilshing (e.g. dictionaries on CD-Rom) with the distribution of electronic databases such as the BNC. The creation of a good electronic corpus-dictionary interface will soon be a priority.

5.3. Application to Natural Language Processing by Computer

In recent years the value of electronic text corpora has been increasingly recognized by those researching in natural language processing (NLP) and speech technology. The reasons for this are diverse, but one recurrent argument in favour of corpora is that they furnish an empirical means of testing out software concerned with such NLP tasks as parsing, language understanding systems, and machine translation. Another argument is that

these days NLP specialists are becoming more and more aware of the importance of probabilistic information about language, and probabilistic models of language processing. For these statistical applications, corpus data has to be analysed quantitatively on a large scale. Advanced corpus-based NLP applications typically require corpora which have been annotated (e.g. with parsing information), rather than corpora in their "raw" state.

5.4. Applications to Language Teaching

Unfortunately, educational applications tend to be poorly funded, in comparison with those in NLP (where nowadays research is often backed by governments and powerful commerical enterprises). Consequently most of the applications in this area are potential rather than actual. Nevertheless, we may look forward to a time when educational users may be able to make extensive use of corpora such as the BNC, in this respect benefiting from the spin-off from more powerful commercial and governmental interests.

Examples of educational uses (going from the more immediate to the more long-term applications) are:

(a) Data-driven language learning, i.e. making use of corpus-derived concordance data in the classroom, has been persuasively proposed (Johns and King (1991)) as a method of discovery learning in such areas as grammar and lexis.

(b) Sub-corpora consisting of special text types (e.g. law, science, technology) may be used as a basis of teaching materials in ESP (English for Specific Purposes).

(c) Frequency studies based on corpora can be applied to the granding of vocabulary and grammar materials, or of reading materials.

(d) Developmental corpora (of children's language) can also be used for grading purposes such as those of (c) above. Some of the material of the BNC will include children's language.

(e) Corpora of learners' language (e.g. the Longman-Birkbeck Error Corpus) can be systematically compared with a corpus of native-speakers' language, such as the BNC, in order to provide better understading of the

processes and needs of second language acquisition, as revealed in learners' productions.

(f) Bilingual and multilingual corpora will in the long run provide a far better basis for contrastive analysis and translation studies than has been available up to now. For example, parts of the BNC may eventually be compared with electronically-stored translations in other languages.

6. Conclusion

The last sections above reveal clearly that a major corpus development, such as the compilation of the BNC, is no more than one stage in an ever-widening build-up of corpus resources. This applies on the one hand to the need for "collateral corpora" such as translation corpora and corpora of learners' language, and on the other hand to the need for continual diachronic extensions and updatings of existing corpora. It may seem a discouraging thought that no corpus-building enterprise is ever complete or completely satisfactory: if for no other reason than that the language itself continues to evolve. On the other hand, each corpus-building exercise is a solid achievement - an enduring resource for future users - and a platform on which future corpus compilers and corpus users can build.

References

Altenberg, B. (1991) 'A Bibliography of Publications Relating to English Computer Corpora,' in S. Johansson and A.-B. Stenstrom (eds.) *English Computer Corpora: Selected Papers and Research Guide*, Berlin: Mouton de Gruyter, 355–96.

Biber, D. (1988) *Variation across Speech and Writing*, Cambridge: Cambridge University of Birmingham.

Leech, G. (1991) 'The State of the Art in Corpus Linguistics,' in K. Aijmer and B. Altenberg (eds.) (1991) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, London: Longman, 8-29.

Pawley, A, and Syder, F. H. (1983) 'Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency,' in J. C. Richards and R. W. Schmidt (eds.) (1983) *Language and Communication*, Lon-

don: Longmans, 191-226.
Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*, London: Longman.

ABSTRACT

100 Million Words of English: The British National Corpus (BNC)

Geoffrey Neil Leech

This paper is a brief decription of the British National Corpus (BNC) project, which is a collaboration between commercial and academic partners. The "100 million words of English" corpus is intended to be representative of a whole range of English language currently used in speech and writing. The main tasks of corpus development can be listed as: corpus design, acquisition and preparation of data, corpus processing and making corpus material available to end-users. The focuses on four main areas of application, i. e. linguistic research, reference publishing, natural language processing by computer and lanuage teaching.

Department of Linguistics
and Modern English Language
Lancaster University
Lancaster, LA1 4YT
England