Trawling the language: Monitor corpora

Jeremy Clear

The term *monitor corpus* was coined by John Sinclair in a brief resumé of the state of the art of corpus linguistics (Sinclair 1982). The paper discusses some features of the design and implementation of a mainframe-based software system to handle such a monitor corpus. A system of this kind is the obvious direction for future corpus-based lexicography. I shall confine my discussion to corpus analysis of general English, though the principles of sample and monitor corpora apply equally to the study of special varieties of a language. This research was begun as part of the COBUILD project in computational lexicography which was supported by Collins Publishers.

1. The characteristics of sample corpora

Sample corpora are static entities. There are no doubt many such corpora in existence, and most lexicographers will be familiar with the Brown, LOB or London-Lund corpora. These examples illustrate very clearly the characteristics of a sample corpus. There are two phases involved in the work: text collection followed by analysis. Johansson (1980), Francis (1979), and Svartvik and Quirk (1980) have published on the important aspect of sampling, and corpus based study of this kind typically follows some basic guidelines.

Fixed size. The length of the corpus is fixed at a certain number of tokens.

Balanced sampling. If the corpus is meant to be representative of the language and if it is only a few millions of words in length, then it is important that samples are chosen carefully and are of controlled size. The aim is to achieve a cross-section of genres, language varieties, dialects, etc.

Accurate representation of surface form. The sample corpus aims to represent as accurately as possible in machine-readable form the relevant features of the substance of the sample. This usually means that punctuation, spelling, paragraphing, etc. are laboriously and expensively checked and corrected. Speech transcripts are often standardised and exhaustively checked.

The analysis of the corpus is not fully independent of the text gathering unfortunately, and manual pre-editing of the computer text may be required to suit the specific needs of the analyst. Nevertheless the prepared corpus can in principle be used as the basis for many studies of widely differing aspects of language. Brown and LOB were designed to fulfil this function and they have been very valuable aids for a large number of researchers. Some of the features of the analysis of sample corpora are these:

Reanalysis of the same data. The corpus is a constant data sample to which varying methods of analysis may be applied and the results justly compared.

The sample is dated. Brown and LOB consist of samples from English published in 1961. The sample corpus is inevitably synchronic in orientation.

2. The characteristics of monitor corpora

At Birmingham University we are proposing to establish a monitor corpus – different in kind from the recognised sample corpora. The outline of my proposal is that text should continuously *flow* through the computer system, with a steady input of new text being subject to analysis by standard software and the results of the analysis directed to an online database. The corpus may be represented diagrammatically as in Figure 1. The need for sample corpora remains, of course, and so a "sump" of some 10–20 million words is held at the end of the process. This sump may change in a number of ways. Old texts may be replaced by more up-to-date ones; the range of genres, subject areas, varieties of English, etc. may be adjusted; a complete subset of the sump may be drained off to form a smaller, more specialised sample corpus, and so on. I am not convinced, however, of the need to store the whole corpus in perpetuity, and my proposal is that the raw stream of text once monitored should be discarded. Luckily the waste will be nontoxic, non-radioactive, and may be safely buried in anyone's back yard. It would be absurd for entymologists studying beetles to keep in the laboratory every single beetle that has ever been looked at; if they want to see

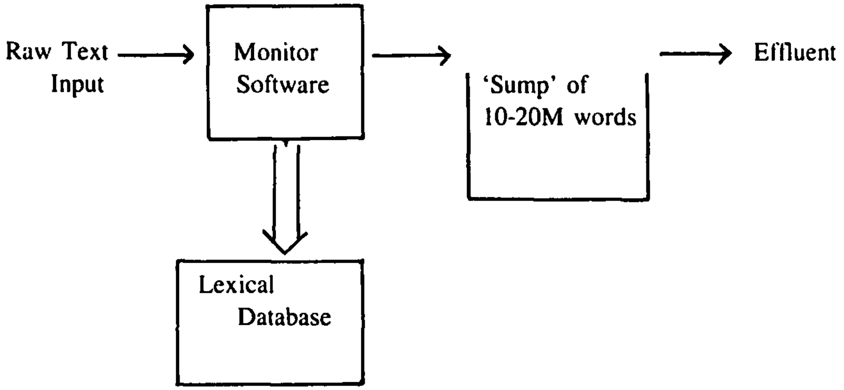

Fig. 1

more beetles then there are plenty of them in the wild. So it is with words: only the rare and interesting specimens need to be kept in matchboxes.

Machine-readable text is now available in abundance and can be poured into the system at a rate of, say, 1 million running words per month. Some of the features of a monitor corpus are:

More information about language. The benefit for lexicography in particular is clear. It is widely recognised that the observed frequency of occurrence of very many English words, familiar to most if not all adult native speakers, is well below 1 per million. In a corpus of over 7 million tokens, there were just over 7000 types with a frequency greater than 50. Evidence of lexical behaviour drawn from a corpus of fewer than 10 million words will be sparse for all but the core vocabulary of the language.

Better statistical foundation. Linguists who have worked on the statistical behaviour of words will be only too aware of the difficulties which arise when they have to work with very small observed frequencies. Current research is tending towards language models which take into account probabilities, and a monitor corpus offers the chance of acquiring adequate figures.

A diachronic perspective. Eventually, analysis will be able to furnish detailed evidence of the diachronic aspects of language.

More effective use of computer facilities. Since text is processed sequentially, the load of processing is more easily managed. If really massive text samples must be analysed, and if computer hardware and software develops at the present rate, then most computer centres will be unable to handle processing of sample corpora using conventional methods.

3. Gathering text samples

The proposal of a monitor corpus is partly motivated by the increasing availability of English text in machine-readable form. In 1965, the cost and effort involved in digitising one million words of English was a major consideration for Kucera and Francis. In 1986, text can be obtained in machine-readable form much more easily. At Birmingham we have already exploited the 1970s technology of the Kurzweil Data Entry Machine to convert printed text into computer files. This process continues, and the KDEM has been steadily enhanced to improve throughput. In the 1990s the information technology boom will certainly ensure that documents are stored and transmitted in digital form. Unfortunately, the speed of technological advance has left us with an ethical and legal confusion over the ownership of information, which is hindering the acquisition of text. Typesetter tapes, online databases, electronic journals, news agencies, electronic mail systems: these are just some of the sources which can feed a monitor corpus.

The sheer bulk of the incoming text will make it prohibitively expensive to do a manual proofread, and so the emphasis at first will be on quantity rather than quality; I would rather have substantial evidence than a standardised, edited but inadequate sample. Manual pre-editing of the input to the monitor corpus must be minimal, so I adopt a number of guiding principles:

1. Text will be accepted in a wide range of formats, depending on the source. The pre-edit software attempts to ignore the "extra-textual" material (typsetter codes, racing results, chemical formulae, etc).
2. A certain amount of format checking can be done by the computer, to standardise use of single quote and apostrophe, full stops in abbreviations, line-end hyphenation, and so on.
3. The computer must await the human operator's decisions on the classification of the text type (genre, source, subject matter, etc).
4. Full use should be made of any available software which will assist in preparing a clean text. Even a crude spelling checker will improve KDEM output considerably (since the KDEM makes many "illogical" misreadings).
5. Attention must still be given to ensuring that the texts processed do not as a whole constitute a grossly unbalanced sample. It will be necessary to select and reject texts according to their appropriateness to the aims of the project.

4. The analysis of a monitor corpus

How is the analysis done? The mainframe computer acts as a filter, trapping data which is of interest to the researcher and letting through data which is not. It will no doubt be objected that we cannot predict *in advance* which aspects of language will be of interest to the linguist. This is of course true – it is equally true when we analyse sample corpora: but in the latter case we simply rescan the same limited stretch of language, while in the former we cast our net afresh into more language data.

The principle is that we know roughly what features of the corpus we want to record from the outset, but that as analysis proceeds new interests, unexpected patterns, or different theories may demand additional filters. The sump provides a limited "clawback" facility: text which has already been seen can be pulled back for re-analysis.

The monitor software gathers information from each text and stores it in a growing database. It is important to control the rate of growth of the database, so that it is possible to maintain an online access facility given the particular hardware limitations. Despite almost daily announcements of advances in computer technology, the accumulation of analysed data in a linguistic database will place a very heavy load on most computer installations. Experience has shown that the lexicographers' demand for more and better access to real language data always outruns advances in software design and hardware speed. The monitor corpus must be designed to make most effective use of current facilities. The continual flow of data to be processed distributes the load of computer processing so that maximum use can be made of CPU resources in off-peak, cheap shifts. The amount of information which can be derived from a corpus is very large, and only a small fraction of it can be recorded explicitly or implicitly in the database. Our interest in a corpus will be focussed on lexical items, in one way or another, so the lexical item will be the primary unit of organisation for the database in the first instance. The COBUILD lexical database already forms a skeleton structure which can be fleshed out with data collected automatically from the corpus. The monitor software is distinct from the automatic pre-edit phase, and is modular in design. The text processing tools which are included in the Unix operating system provide a model for the design of the monitor package. A number of different programs operate virtually independently, taking their input either direct from the text stream or from the output of another module. Each program should be designed in such a way that it can be slotted into the existing package with the minimum of disruption. I am working on the development of monitor software at Birmingham, and I will outline the stages of analysis which are envisaged.

Word frequency statistics. The software breaks the input into graphic word forms and keeps a frequency table. Word forms in this indexed list are linked to the lexicographic entries in the database. Homographs are multiply linked to the headwords which contain the relevant word form in their inflected forms list. The frequency table also records the number of times the word form occurs in each text. Once a grapheme has occurred in more than 20 different texts, the record is summarised and condensed by recording the frequency per genre category or subject area. If 20 different genre frequencies are collected, then no further distribution figures are stored for this word. This illustrates the application of the principle of graded summary as a means of keeping the database of manageable size. At regular intervals the frequency table is scanned and a tabulated report is produced of type/token ratio, number of new word forms encountered, and statistics of the form produced by the Oxford Concordance Package. These reports may be printed onto fiche then dumped to archive tape every 5 or 10 million tokens. It is quite straightforward to add routines which report on significant deviations from an established pattern of word frequency and distribution. Words whose overall frequency increments erratically can be marked for special attention. Texts which show unusual patterns of word frequency may be put aside for an editorial check.

Keyword in context citations. These are obviously of interest to the lexicographer. Until now it has been COBUILD policy to produce complete KWIC concordances to sample corpora, but in future citations will be selected automatically. At first the selection will be crudely mechanical. For word forms with an average frequency greater than 1 per 1000, it is not desirable or necessary to store full citations and the computer will hold only a limited number. The number may be a constant or may vary with the word's relative frequency overall, and the citations might be selected randomly, or to reflect the distribution of the word across text types. Work is in progress at Birmingham to use statistical measures of collocational patterns to classify citations as typical or untypical. The machine will not perform as well as a human lexicographer at this task, but it does work faster for longer hours. The low frequency words will slowly accumulate citations until the pre-set ceiling is reached and a selection mechanism is applied. To save on expensive disk storage, a daemon process runs at an appropriate time of day to move the citations for low frequency words onto tape or an exchangeable disk pack. Figures obtained from our 20M sample show that over 60% of the word types occur fewer than 4 times, and if a cut off of 500 were applied, only 3200 keywords would need their citations pruned. If the database gets too big again, then we must introduce more severe restrictions – the excision of many proper nouns would reduce the bulk without losing anything too valuable – but local interests will dictate where cuts should be made.

Integration with the lexicographic database. The analysis of a monitor corpus can be seen as a natural development of the work which has already been done on compiling the COBUILD dictionary database. The database can now act as a simple "knowledge base" for the monitor software. Let me suggest practical ways in which the two complement each other.

The computer definition of a word is usually "a sequence of alphanumeric characters surrounded by spaces" (Hofland and Johansson 1982: 7). The monitor corpus software can consult the database and select special phrases for separate entries (*the Labour Party, first and foremost, video display unit,* for example).

Word frequency listings can be partially lemmatised by a simple automatic lookup procedure. These listings will yield very interesting new information on the frequency of lemma in comparison with the frequency of word-forms. Automatic homograph separation is still a long way off, unfortunately for the lexicographer.

The detailed information on the syntactic behaviour of each lexical item provides the basis for a reasonably accurate word-class tagging procedure. I have been using the most basic *n, vt, vi* labels extracted from a printed dictionary as a look-up list for a word-class tagging program and it has proved very successful. There is no doubt that the availability of more delicate syntactic information about each word will improve the results significantly. The basic word class tag can be attached to each concordance citation so that, for example, the concordance for *'record* can be displayed separately from that for *re'cord.*

Each lexicographic entry in the dictionary database can be linked directly to the KWIC concordances for every headword, so that a lexicographer working at a terminal can call up the relevant citations.

These proposals are based on the practical programming work carried out during a five year dictionary project using a corpus of over 7 million words. The statistical information gained from study of this sample enables strong predictions to be made about the behaviour of words in text, so that we can tune the monitor corpus to record in detail the rare and significant events and *summarise* that which is frequent and regular.

References

Francis, W. Nelson (1979), "Problems of assembling and computerizing large corpora", in: Bergenholz H./Schaeder, B. (eds.), *Empirische Textwissenschaft: Aufbau und Auswertung von Text-Corpora,* Königstein: Scriptor Verlag, 110–123.

Hofland, Knut/Johansson, Stig (1982), *Word Frequencies in British and American English,* Bergen: Norwegian Computing Centre for the Humanities.

Johansson, Stig (1980), "The LOB corpus of British English texts: presentation and comments", in: *ALLC Journal* 1, 25–36.

Sinclair, John McH. (1982), "Reflection on computer corpora in English language research", in: Johansson, S. (ed.), *Computer Corpora in English Language Research,* Bergen: Norwegian Computing Centre for the Humanities, 1–6.

Svartvik, Jan/Quirk, Randolph (eds.) (1980), *A Corpus of English Conversation,* Lund: Gleerup/Liber.