DOCUMENT RESUME

ED 365 127 FL 021 703

AUTHOR Benson, Philip
TITLE Hong Kong Texts on Hong Kong: Developing Computer Text Corpora at Hong Kong University.
PUB DATE Sep 93
NOTE 7p.; For journal in which this paper appears, see FL 021 693.
PUB TYPE Reports - Descriptive (141) -- Journal Articles (080)
JOURNAL CIT Hong Kong Papers in Linguistics and Language Teaching; v16 p117-122 Sep 1993

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computational Linguistics; Computer Assisted Instruction; *English (Second Language); Foreign Countries; Higher Education; Indexes; Language Research; *Language Usage; *Language Variation; Newspapers; Vocabulary
IDENTIFIERS *University of Hong Kong

ABSTRACT

This paper reports on a paper to develop two computer text corpora, called Hong Kong Texts on Hong Kong," at the University of Hong Kong English Centre. These corpora are intended to serve as a resource for teachers and researchers who need information on English usage in published text in Hong Kong. The aim of the report is to disseminate information on the project, and to initiate discussion on how the corpora might be used and developed. The paper is in three sections. Section 1 gives some background information on corpus development and text analysis in language research. Section 2 introduces the two corpora under development in the English Centre. Section 3 raises some issues concerning the use of the corpora and their future development. (Author/JL)

Hong Kong Texts on Hong Kong: Developing Computer Text Corpora at Hong Kong University

Philip Benson

Introduction

This paper reports on a project to develop two computer text corpora, called *Hong Kong Texts on Hong Kong*, at the University of Hong Kong English Centre. These corpora are intended to serve as a resource for teachers and researchers who need information on English usage in published text in Hong Kong. The aim of this report is to disseminate information on the project, and also to initiate discussion on how the corpora might be used and developed.

The paper is in three sections. Section 1 gives some background information on corpus development and text analysis in language research. Section 2 introduces the two corpora under development in the English Centre. Section 3 raises some issues concerning the use of the corpora and their future development.

Computer text corpora and corpus linguistics

A computer text corpus for language research (a 'linguistic corpus') is a collection of machine-readable texts which have been combined and formatted so that linguistic features can easily be analysed using text analysis software.

The history of large-scale linguistic corpora goes back 30 years to the 1 million word Brown Corpus of American English (Kucera and Francis, 1967). A comparable corpus of British English, the LOB corpus (Hofland & Johansson, 1982) was produced in the early 1970s at the Universities of Lancaster, Oslo and Bergen, to be followed by the 0.5 million word London-Lund corpus of spoken English (Svartvik & Quirk, 1980). The 1 million word Kolhapur corpus of Indian English (Shastri 1985) was the first of its kind in Asia, and the International Corpus of English currently has teams working on corpora in various parts of the world including Hong Kong, India, Philippines, and Singapore (Greenbaum, 1991).

Corpora which aim to provide evidence on the state of a language variety at a particular moment in time have tended to follow strict guidelines. The LOB corpus, for example, was divided into 15 'genres', and within each genre titles were selected randomly from library catalogues. Fixed length (2,000 word) samples were then chosen at random from the titles selected. How far such techniques can ensure genuine 'representativeness' is open to question, however (Sampson, 1991, Sinclair, 1991), and there is currently a trend towards much larger, and possibly looser corpora. The projected size of the British National Corpus (Leech, 1991) is 100 million words, and the *COBUILD* 'Bank of English' is comparable. Clear (1987) and Sinclair (1991) have also raised the idea of a 'monitor' corpus, which would effectively be unlimited in size since it would consist of a continuous flow of text passing through 'filters' designed to capture changes in the language.

The ideal size of a corpus designed to represent the state of a language variety at a particular moment in time probably cannot be specified because of the difficulty of quantifying the object to be represented (a language variety). It is clear, however, that perceptions of what constitutes an acceptable sample are determined mainly by the technical limits of the time. Nowadays, corpora of 1 million words

or more can be developed, stored and handled on an average desktop PC, and Sinclair's (1991, p.18) maxim, which states that "a corpus should be as large as possible, and should keep on growing" seems to be reasonable at a time when revolutionary changes in the storage and handling capacities of computers are upon us.

The analysis of text corpora is carried out using text analysis software, which typically produces word lists, frequency counts, concordances, and statistical information of various kinds. *Microconcord, Micro-OCP* (*Oxford Concordance Program*) and *TACT* are examples of commercially-available text analysis packages which are capable of handling large text corpora on a desktop PC[1]. Computer corpora will generally support any kind of inquiry in which information can be inferred from the graphic forms of words, phrases and structures, and their relationships with each other. More complex inquiries (into syntax, semantics and discourse analysis) may be supported by tagged corpora, in which tags have been inserted to represent features of text which cannot be inferred from the graphic forms of words alone. Tags are sometimes used, for example, to discriminate between homographs, to indicate parts of speech, or to link semantically related items.

Corpora must normally be formatted and marked up in order to ensure that ambiguously represented features such as titles, paragraph breaks, page breaks, and so on are explicitly and consistently represented. Unfortunately, different types of software tend to support different types of formatting, markup and tagging. Corpus preparation and 'housekeeping' can therefore account for a large proportion of development time. The *Text Encoding Initiative* (TEI), based on *Standard General Markup Language* (SGML), represents a step in the direction of standardisation based on the philosophy that electronic texts should be interchangeable and support research aims other than those of the designer (Burnard, 1991).

Leech (1991, p.10) has listed fields of study in which corpus-based studies have been recorded: linguistic theory, computational linguistics, grammar, dictionaries, the study of meaning, discourse analysis, conversation analysis, language variation, speech technology, speech science, historical studies of language, child language acquisition, psycholinguistics, applied linguistics and orthography. The exhaustiveness of this list suggests that corpus linguistics is less a specialised field in its own right than a method of inquiry which has implications for a wide range of studies. The systematic and objective character of corpus-based evidence constitutes a challenge to traditional methodologies in fields which have relied heavily on introspection or random observation, and corpus linguistics has already had a decisive influence on lexicography (Sinclair, 1991) and computational linguistics (Sampson, 1991), where possession of a corpus is rapidly becoming a *sine qua non*. It remains to be seen whether the influence of corpus linguistics will be as strong in other fields.

The Hong Kong Texts on Hong Kong Corpora

Hong Kong Texts on Hong Kong is the name used for two computer text corpora under development in the English Centre at the University of Hong Kong. Both corpora are designed with the same aim in mind, to furnish textual evidence for the description of English in Hong Kong. The first (the 'academic' corpus) contains extracts from books and papers published in Hong Kong on Hong Kong topics, while the second (the 'newspaper' corpus) contains Hong Kong news reports from the *South China Morning Post*. Because the texts included in these corpora come from publishing houses which are known to maintain a high standard of English in their publications, the corpora can be seen from two points of view. On the one hand, they can be seen as corpora of standard English usage illustrated by Hong Kong contexts. On the other hand, they should also provide evidence on distinctive local features of usage which will be of interest to anyone concerned with the accurate description of English in Hong Kong.

The development of the corpora described here began in January 1993 and it is anticipated that the first stage of development will be completed by December 1993. At that stage, the academic corpus

will consist of around 1 million running words of text, and the newspaper corpus around 1.5 million. At the time of writing, some 800,000words of the newspaper corpus have been prepared, and are being used in trial runs. (Some initial results are mentioned in Section 3 of this paper.)

The academic corpus has been developed with the assistance of four major publishers of Hong Kong texts in English: the Centre of Asian Studies (University of Hong Kong), Chinese University Press, Hong Kong University Press and Oxford University Press. These publishers have given permission to store substantial extracts from a variety of sources in a database used for research within the English Centre. In some cases publishers have been able to provide text in machine-readable form, but a great deal of text has to be scanned electronically. The availability of text in electronic form has not been considered a valid criterion for text selection since the aim is to cover as wide a range of texts as possible. The time required for scanning text and correcting scanned files is therefore a major factor determining the timing of the project.

In the academic corpus three main criteria apply in the selection of a text:

1. It must be published in Hong Kong for Hong Kong readers.

2. It must be available from Hong Kong bookshops.

3. It should deal with current social issues in Hong Kong.

The cut-off dates for selection are 1980 and 1993 (date of publication). To ensure a range of texts, it was decided to limit the contribution of each author to around 10,000 words. This limit has been applied flexibly, however, because it was also decided to include whole texts (papers or book chapters) whenever possible. Within these criteria the aim has been to collect as much text as possible.

Since the texts included in the corpus have been written for publication in Hong Kong, they might in fact be described as 'popular academic' since they are generally written for a readership which extends beyond fellow academics. Topic areas covered include law, politics, sociology, economics, education, urban planning and the environment (no attempt has yet been made to group the texts into 'genres' within the corpus). Law is prominent as a topic, but this reflects its penetration into other topic fields (for example, the political future of Hong Kong). Not surprisingly, '1997' has a tendency to crop up as a topic in virtually any of the texts selected.

The newspaper corpus has been developed with considerable assistance from the *South China Morning Post*, who have provided in machine-readable form more than 2,500 Hong Kong news reports published in 1992-3. As a result this corpus has progressed more rapidly than the academic corpus. The newspaper corpus also covers a much wider range of topics than the items in the book corpus, including, for example, crime reports and reports on entertainment and social events.

Both of the corpora contain samples of writing from both native and non-native speaker writers. I have not attempted to separate these categories, although this might be possible if detailed biographical information on authors were collected. Some care would need to be taken in assigning texts to authors however, since the work of non-native speakers may well be checked by native speakers before publication, while slips might appear during typing of material written by native speakers. These difficulties aside, we can also assume that authors follow more or less strict rules and conventions of the genres within which they are writing. All this suggests that the corpora will not yield valid or interesting information on the native/non-native speaker distinction. On the other hand, as representations of the language which a society has before it, we might expect the corpora to yield interesting data on models of English which are prevalent in Hong Kong.

One question which arises is how representative the corpora are of models of English in Hong Kong. It should perhaps be made clear that there is no intention to suggest that these corpora represent 'Hong Kong English' as a whole. Similarly, it must be recognised that Hong Kong readers of English have access to a wide range of imported models of English other than those represented in the two corpora. Nevertheless, the two text-types selected represent a significant proportion of locally published text, given that for certain text-types (e.g. creative writing in English) locally written and published text is close to zero. It is not known exactly what each corpus represents in terms of percentages of the text which is theoretically available for its text-type. I would estimate that these percentages are relatively high, when compared with other corpora such as LOB and Brown. In the case of the academic corpus, at least one extract is included from each book which was identified from publishers' catalogues as being potentially relevant, and I would estimate the representation of authors to be more than 50% of those who have published locally since 1980.

Teaching, research and development

In this last section, I would like to discuss some of the ways in which the corpora described might be used in teaching and research, and how they might be developed beyond the current project. A major concern of the project has been the accessibility of locally-relevant corpus data in Hong Kong, and the intention is to make these corpora as widely available as possible within restrictions of copyright. At present this means that researchers will be encouraged to use the completed corpora by arrangement with the English Centre on condition that they are not copied or distributed.

Although it is likely that these corpora will be too unwieldy for direct classroom access (cf. Ma in this volume), a number of applications in ELT are envisaged. These include production of vocabulary lists for locally relevant teaching material, and generation of examples of words, phrases and structures in contexts which are relevant and accessible to Hong Kong students. The value of Hong Kong texts for ELT in Hong Kong can be argued on a number of grounds. Familiarity of content is recognised as an important starting point for new learning, and the presentation of new items within accessible contexts should have positive effects on the learning process. Locally relevant text may also motivate learner interest in English, and a strong case can be made for the use of local texts as a means of developing learner awareness of the roles and functions of English as a language of Hong Kong.

A further argument for the use of a Hong Kong text corpus in teaching and research is that there is strong evidence that many English words are used differently in local text and imported text. In initial trials on 800,000words of the newspaper corpus, in which concordance output on selected words has been compared with descriptions in the *COBUILD* dictionary, some interesting data was revealed. In an investigation of the words *local, locally, localised* and *localisation* (Benson, 1993), it was seen that there were extensive differences between the description of these words in the *COBUILD* dictionary and their use in the corpus. These differences covered both the senses of the words and the relationships of derivation among them.

Similar evidence emerged from an investigation of the words *converge* and *convergence*. The *COBUILD* descriptions of these items suggest that the verb *converge* normally takes a plural or collective subject (i.e. *ideas* or *people converge*), and that if the noun *convergence* is followed by a preposition, that preposition will be *of* or *between*. In the newspaper corpus, on the other hand, due to the prevalence of the form *convergence with the Basic Law*, it was seen that the verb *converge* often takes a singular subject, and that both *converge* and *convergence* are most often followed by the preposition *with*. The most prominent patterns in the newspaper corpus turned out to be '*A converges with B*' and '*convergence of A with B*' in contrast to *COBUILD*'s 'A and B converge' and 'convergence of/between A and B'. *Convergence* is of course a specialist term in the context of Hong Kong politics, but the interesting point to note is that the emergence of a specialist term is accompanied by a change in syntactic form.

It is anticipated that investigations of the two corpora will reveal more data of this kind. This kind of data may well be of interest to researchers in various fields. It might also be translated into the kind of information which would be of great interest to learners of English in Hong Kong.

The first stage of development of the *Hong Kong Texts on Hong Kong* corpora is scheduled to be completed within one year. Developments beyond that point are uncertain, but it is clear that there will remain a 'shortage' of electronic text in corpus form for the linguistic analysis of English usage in Hong Kong (the need for a corpus of spoken English immediately comes to mind). One proposal which I have raised is the development of a 'Corpus of English Usage in Hong Kong', which would be developed on a modular basis, possibly incorporating existing and developing corpora, and filling in gaps with new corpora, where data can obtained.

This proposal calls for a degree of cooperation among researchers and a commitment to the accessibility of corpora. There would be clear benefits in developing a local fund of knowledge and experience covering techniques and principles of text collection, preparation and analysis of corpora, use and development of software, handling of copyright problems and creating wider access to completed corpora. These benefits would be additional to the growth of an invaluable source of data on the various manifestations of English in Hong Kong, which would provide systematic and objective evidence in response to concerns over standards of English in the territory. Disseminating information on corpora currently being developed and encouraging their use by researchers might represent a first step towards this goal.

Notes

[1] *Microconcord* and *Micro-OCP* are published by Oxford University Press, and *TACT* is produced by the Center for Computing in the Humanities, University of Toronto.

References

Benson, P. (1993). *Electronic text analysis and the lexis of Hong Kong English: What can it tell us that we don't already know?* Paper presented at the Joint Seminar on Corpus Linguistics and Lexicology, Hong Kong University of Science and Technology Language Centre and Guangzhou Institute of Foreign Languages Department of English, June 1993.

Burnard, Lou (1991). An introduction to the Text Encoding Initiative. In D.I. Greenstein (ed.) *Modelling Historical Data*, Goettingen: Max Planck Institut für Geschichte.

Clear, Jeremy (1987). Trawling the language: monitor corpora. In M. Snell-Hornby (ed.) *ZuriLEX Proceedings*, Tübingen: Francke.

Sinclair, John M. (1987). *Collins COBUILD English Language Dictionary*. London: Collins.

Greenbaum, Sidney (1991). ICE: the International Corpus of English, *English Today*, 28, 3-7.

Hofland, K. and Johansson, S. (1982). *Word Frequencies in British and American English*. Bergen: The Norwegian Computing Centre for the Humanities.

Kucera, Henry and Francis, W. Nelson (1967). *Computational Analysis of Present-day American English*. Providence: Brown University Press.

Leech, Geoffrey (1993). 100 million words of English, *English Today*, 9(1), 9-15.

Sampson, Geoffrey (1991). Natural Language Processing. In Christopher Turk (Ed.) *Humanities Research Using Computers*, 126-136. London: Chapman and Hall.

Shastri, S.V. (1985). Towards a description of Indian English: a standard corpus in machine-readable form, *English World Wide*, 6, 275-278.

Sinclair, J.M. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Svartvik, J., ed. (1990). *The London-Lund Corpus of Spoken English*. Lund University Press.