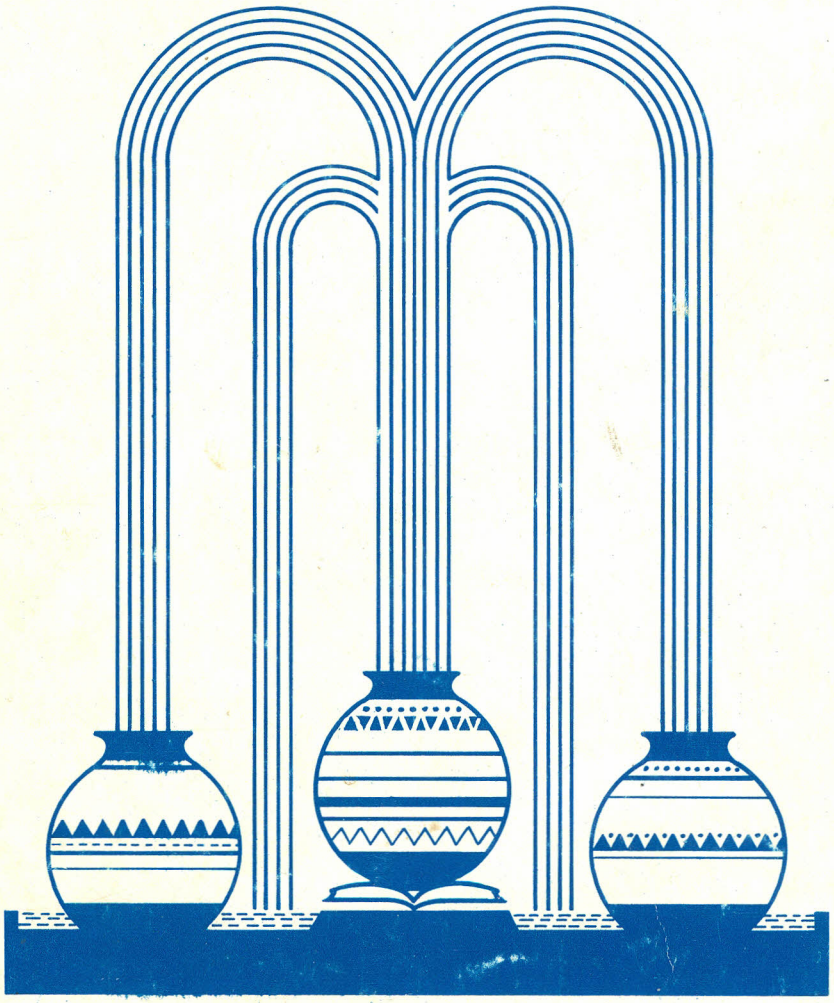
CHEMCHEMI

International Journal of Arts and Social Sciences

VOLUME 1 DEC. 1999

ISSN 1563 – 1028



Journal of the Faculty of Arts
Kenyatta University

COMPUTER-BASED CORPORA: PROBLEMS OF COLLECTION AND INTERPRETATION OF KENYAN TEXTS IN ENGLISH

Eunice A. Nyamasyo*

Abstract

Computer-based corpora as sources of language material for description is a relatively new concept in linguistics in Kenya, if not in Africa generally. The collection of relevant and / or appropriate text samples: spoken, written, or otherwise, is therefore fraught with a number of difficulties. The linguist is faced with a range of problems in, firstly, processing any collected material and secondly, making correct interpretations of the said data. The computer, a recent innovative tool in language-based research, requires the researcher to have both data inputting and processing skills. Text samples as the basis of data are obtained from various sources some of which require special permission to access. Once acquired, text samples vary in origin and characteristics hence raising issues of interpretation. Notwithstanding, the Kenyan sample is an essential component of the International Corpus of English (ICE) as a source of data for the description of the present-day English language.

Introduction

A corpus is a sufficiently large body of naturally occurring language material collected and designed for particular research purposes. Such purposes may include a description of the structure of the grammar in the texts; a stylistic analysis of the textual material; a determination of the sociolinguistic variation of the language in question as observed in the texts; a comparative analysis of the grammar in the texts in the corpus with those in other corpora. Sebba (1991) points out that the advantage of a corpus, particularly a computer corpus in language research is that 'something is known or can readily be found about their linguistic and statistical properties.'

A computer corpus is a machine-readable version of a corpus which may have initially been collected and stored either in printed or spoken form. This may be illustrated by the two pioneer machine-readable corpora: *The Brown University Corpus of Standard American English* (Francis and Kucera. 1979), generally known as *The Brown Corpus* and *The London/Oslo-Bergen Corpus of British English*, completed in 1978 (Johansson. 1978) popularly referred to as *The LOB corpus*. These have been extensively used in the detailed and more systematic analysis of the structure of present-day English grammar and in the production of a variety of grammars (e.g. **The Comprehensive Grammar of the English Language** by Quirk *et. al.* 1985).

The two corpora have also provided the databases for other areas of research in, or with the language. For example, in machine-translation and machine-aided translation, language material from the two corpora are being used in experiments to develop a translation language to be used to 'comprehend' other languages occurring and spoken in the enlarged common European market. The textual material in the corpora is also employed in the development of language analysis computer programmes (Garside *et. al* .1987). They have acted as models to the compilation of other corpora of English such as *The Kolhapur Corpus of Indian English* (Shastri. 1982); *The Australian Corpus*

of English (Peters .1987); and to a large extent, the on-going *International Corpus of English* (ICE project being undertaken in a number of English-speaking countries.

The ongoing ICE project has involved the collection of both spoken and written texts in Englis produced from 1990 to date in countries in which English is spoken natively (e.g. Britain; th United States of America) and from those in which English occurs as a second or even third languag (e.g. Kenya; Tanzania; Zambia). Foreseen as a source of a modern text-oriented approach to th description of the English language obtained from different regions, ICE will offer a systemati data base from which descriptions of the various varieties of the language may be done. Mor particularly, the compilation which aims at a representative sample of language in use will provid the scope from which comparative linguistics and other types of linguistic analysis may be carrie out using computer-based analytical tools. So far textual material from a broad range of genre identified as *text categories*, have been derived from both written and spoken sources from divers areas of language use. These include such sources categorized as: Informational; Instructional (e.g administrative/regulatory; skills; hobbies); Persuasive; Biographical.

The 'Learned' Text Category

Identified as Category J in both the Brown and the LOB corpora respectively, the *learned te* category for ICE is classified under the broader Informational text category. The text extracts there in the Kenya sample have been obtained from a range of printed documents mainly in the literar and scholarly domain. These so far include, for example, 2,000 word long extracts from Ph.D research proposals and theses; M.A/M.Sc. proposals and dissertations; timed student essays; a well as untimed student essays. There are also recordings of spoken English obtained from radi broadcast lessons and discussions, and from television interviews and discussions on topics from range of specialist areas.

Issues in Corpus-Text Collection in the Kenya Sample

The Project team

Members of the Kenyan team, and of the East Africa team in general are not all from the sam region. For example, the team leader as well as the data collection co-ordinator for the East Africa sample are based in Germany. The centre of operations may therefore be said to be Germany. The are two other members working from Kenya. The distances between members of the team make th co-ordination and smooth consultations between the participants quite problematic.

In the Kenyan context, the use of computer in language-based research is still relatively a ne area. As such there are relatively few people who are familiar with current trends in language-base research involving computer corpora and computer-based analytical tools. The computer is sti seen as a tool more applicable in the numerically-oriented natural and physical sciences than in th social sciences. The concept of a corpus, a computer corpus in particular, as a basis for linguist analysis has hardly any following. This is because, in the first place, computers and the accompanyin hard/software for research purposes are still relatively few and / or expensive. For example, th price range of an IBM compatible personal computer with the most essential software in Keny currently lies between sixty-thousand (Ksh. 60,000/-) and one hundred and ninety-five thousan

(Ksh. 195,000/-) Kenya shillings. This is unlike the situation in the U.K. where the price ranges between two hundred and fifty (£250) and one thousand (£1,000) sterling pounds (Kshs.102,000/-). There is also very little innovation in thinking regarding other applications of such equipment. Secondly, there are relatively few people who are computer literate and have any typing skills. Such people therefore may have very little enthusiasm to contribute the required manpower for the realization of the corpus.

Other equipment besides the computer, used in the collection of the textual material, such as radio-cassette recorders, tend to be viewed with a lot of distrust by many potential data collectors or providers of text extracts. There is evidence of suspicion and lack of interest with regard to the ideas regarding the processes involving the compilation of a representative corpus as well as the specific contents of the corpus. Some linguists and scholars from other related fields still consider the computer as a secondary rather than a primary facilitator to their work, and do not recognize its relevance to their work. This is more so due to the fact that computer-based approaches, at one point or another, require the researcher to have hands-on experience of the computer and the language data therein: skills which many researchers do not have.

The acquisition of the text samples for the corpus and their entry into the computer memory require time and patience. It is also regarded as a never-ending task with minimal possibilities of realization. This is complicated by the fact that many types of published documents seen as possible suitable sources of text extracts are located in special sections of public, special, or private libraries. These require the researcher to either gain membership to the particular library or to make special arrangements before they can access and retrieve the desired document. Where such has been gained, there may be lacking the necessary information retrieval techniques such as a working photocopier, a microfiche reader, or even an optical scanner.

Identification and Acquisition of Text Extracts

Sources of Text Extracts

There are several public, private and special libraries in Kenya in which are placed different types of documents from which may be obtained text extracts for the corpus. For the 'learned' text category, in particular, the academic libraries located at the five public universities1 provide an ample source of texts. However, a large number of such sources are held in their written form. Furthermore, the libraries hardly have updated accession or acquisition lists. And even if so, most recent acquisitions remain out of reach to users for considerable lengths of time. The acquisition of extracts from such sources in the end require special arrangements. Alternative sources of such materials may be through contacting either the authors or the publishers. But this brings with it other difficulties such as those involving the rights of control over texts; the price of the document containing the sample to be extracted; and even issues of copyright laws.

Some of the known large commercial and industrial institutions in Kenya such as the East African Industries (EAI) have well stocked and maintained libraries in which may be found a wide range of published documents. There are also specialist libraries in international research institutions such as the International Centre for Insect Physiology and Ecology (ICIPE); the International Centre for Research in Agro-forestry (ICRAF); and the African Medical Research Foundation (AMREF) all of swhich are located in Nairobi. These usually have well organised accession lists and title/

author catalogues showing up to date publications in their areas of specialization. However, to determine the nature of what is held in these libraries and to acquire the relevant extracts from the documents require, for non members of the respective institutions, not only special permission to use the library but also time and money to be able to obtain the selected extracts.

The Central Bank of Kenya library, for example, contains a wide range of important documents from a range of activities in the field of banking. There are also documents on research and project reports. Some of these documents contain confidential and/ or sensitive information to the banking sector in the country. These require special written authority from the senior personnel in the bank which, in most cases is obtainable very infrequently. A similar situation also occurs in libraries in the large commercial and industrial institutions mentioned above.

Retrieval Procedures

Most locations for the documents from which texts may be extracted are not yet digitalized. For example, despite the fact that the five university libraries have an inter-library loan system none of them has installed a machine-based network. Therefore, to identify what is available in such libraries, the researcher has still to manually go through the card catalogues. If this fails, one has to do the search for the required document from the open shelves.

The transfer of a required document through the inter-library loans system still depends on two forms of transport: the postal system and the road. This brings about possibilities not only of delays to the borrower but also chances of loss or damage to the desired document. And even after the document has been received, there are difficulties in actually obtaining the text extract as photocopying or other methods of extraction may be out of order or altogether lacking. Besides, there are also possibilities that a number of the documents from which the texts are to be extracted are not of the desired length (that is, 2,000 running words). This then creates the problem of getting another passage to make up for the total required length.

Developing a Machine-Readable Version of the Text Extracts

For the language data to be accessible to the computer it has to be transferred from its original form to the computer memory. Although there are currently a number of possible methods through which this can be done, the most popular one involves manual key-boarding at the computer face. This necessitates the availability of a computer that can be used consistently for the purpose; and for spoken texts, the availability of a good playback cassette recorder as well as a computer.

Problems of illegibility and incomprehensibility of text extracts abound. Such include difficult names of persons, events and places as well as the inability to transcribe appropriately the information in the extract due to the transcriber's unfamiliarity with the sociology of the context of situation. This in turn dims the interest or enthusiasm of the recorder or transcriber. Once the data has been entered into the computer memory there is further work of editing and post editing to ensure that it is only the original version of the extract that has been entered into the corpus. The number of extracts and the overall length in running words for each selected text are also of considerable importance to the compilation process. This is in relation to the interesting observation that the typing skill, in the first place, is not a very common skill, and secondly it is regarded as secretarial work in most cases done by women. Partly for this reason, a number of interested participants

hesitate to volunteer in the whole process of data entry. The copying of a two-thousand-word-long document may, therefore, take hours, if not days - a length of time which many people are not willing to offer.

Issues in the Interpretation of Text Extracts

The text extracts in the Kenya sample for the ICE project have been drawn from a wide range of disciplines. These in turn, definitely espouse different philosophical and/ or ideological orientations. Is it therefore possible to determine a common base from which any form of interpretation of the range of materials in the corpus may be based? And what form of interpretation is the linguist to make from this range of textual material? There are possibilities that much more information than that in the extract themselves and the accompanying bio-data may still be required for certain forms of analysis.

The 'Other Text'

Apart from the actual information in each extract, attempts have been made to present some detailed information about the source of the material. There is accompanying biographical information, for example, about the author in each extract. Much of such information has been selected and provided by the author and/or publisher of the document. Other sources of such extra material are through rough assessments (e.g. of the age; social background; type of occupation) of the author. However, it has been interesting to observe that quite a number of authors or potential sources of text extracts are very shy, hesitant, or extremely sensitive to talking about themselves or even allowing information to be obtained about them. Consequently, how adequate is the accompanying information, what I term: the 'other text', and with what level of confidence can the researchers accept it, to aid appropriate interpretation of the textual material in the corpus?

Kenya is a multiracial and multilingual country. A relatively large number of Kenyans over the age of 18 years are able to speak more than one language. The choice of which language to use for what purpose at what place and with whom is dependent on a number of factors. For example, Kiswahili, as the national language, is used in official/administrative mainly government-related interactions. English, the official language, on the other hand is most frequently used in scholarly/ 'learned' discourse as well as in official transactions. It is also now being used at certain social levels as a first language. The mother tongue, or first language is, to a large extent, used among friends, and mostly in the family/home domain. There are, therefore, several language choices open to the speaker or potential author of a text to be included in the corpus; English is just one of the many. Does the interpreter need to take into account all these other factors when focusing on a single English extract produced by (an) author(s) in a given text category or genre in the corpus? What does the interpreter do with forms of usage which differ from what forms the interpreter knows intuitively or are in grammar books?

Corpus Linguistics

The use of the computer, corpora, and computer-based corpora in particular, provides a relatively new approach to language-based research in Kenya, if not in other countries in Africa. As pointed

out by Leech (1990:2) 'machine-readable corpora provide quite a new philosophical approach to linguistics as it provides a new kind of knowledge as well as a new way of thinking about language'. How is it possible to amalgamate 'this new way of thinking about language' with other older, more familiar ways? To those not yet initiated into the computer revolution and the age of computer-based language research, does the computer-based corpus offer a convincing source of language data? By using a computer corpus do they have enough room to employ their analysis strategies in testing any hypothesis they may have about the language?

Corpora: Competence or Performance?

A computer corpus focuses immediate attention on the behavioural aspects of the given language. This is in the form of naturally occurring spoken or written discourse. However, since the early 1960s, there has been a strong influence by linguists working within the generative grammar tradition on preferences for language data obtained from the intuitions of the native speaker of a language; that is, data obtained from what has been described as the competence of the native speaker rather than from performance. Performance-based language data, it has been argued, is plagued with 'deviant' forms: possibly, the results of momentary lapses in the memory of the speaker or perhaps the speaker's physiological make-up. Such material, it is argued, does not reflect the true characteristics of the language in question. According to this perspective, where does the language data obtained from non-native speakers of the given language, such as Kenyan speakers of English, fit in and what stand does the interpreter take from corpora with such background?

According to Chomsky's (1965) competence/performance dichotomy, what he has recently refined into the notions of 'internalised' language (I-language) for competence and 'externalised' language (E-language) for performance (Chomsky .1988:45), corpora falls within the E-language domain. It can however, be strongly argued that information obtained from performance in a given language can provide adequate linguistic data with which to determine the nature of the given language. However, to the interpreter of such data, what amount of such material would be considered adequate and 'representative' to determine, or talk about, the characteristics of the given language?

Summary

Language-based research has a lot to contribute to and gain from the computer revolution. Existing computer-based corpora such as the Brown and LOB corpora respectively, and the ongoing International Corpus of English being developed offer more systematic approaches and models in the study of language than ever before available. Evidence of the efficiency and effectiveness of the analytical tools offered through computer programmes for language-based research should woo the uninitiated: diehards of the old methods in linguistic analysis, to incorporate the computer, corpora, computer-based corpora and other electronic means and sources of language data and means of communication in their research programmes. In this way, it will make it possible for them and their respective institutions both to acquire computers and other digitalized means of handling language material as data as well as research assistants and researchers who will expand research in human language. In itself too, the computer has many additional exploitable applications such as word processing; other kinds of information storage and, with the emerging powerful versions of the machine and of software in the market, computer to computer communication via the Internet Web

This will in turn provide solutions to some of the problems encountered, e.g. by the participants in the Kenya sample of the ICE Project, and make future corpus compilation an achievable task.

References

Chomsky, N. 1988. ***Language and Problems of Knowledge***: The Nicaraguan Lectures. Cambridge. Mass: MIT Press.

Francis, W. N. and H. Kucera. 1971. "Manual of Information to Accompany a Standard Sample of Present-Day American English." Providence. R. I. Brown University, Department of Linguistics.

Garside, R.; G. Leech and G. Sampson (eds.) .1981. **The Computational Analysis of English**. London: Longman.

Johannson, S., G. Leech and H. Goodluck .1978. "Manual of Information to Accompany the Lancaster/Oslo-Bergen Corpus of British English for Use with Digital Computers." Oslo: Oslo University, Department of English.

Quirk, R.; S. Greenbaum; G. Leech; and S. Svartvik. 1985. **A Comprehensive Grammar of the English Language**. London: Longman.

Sebba, M. 1990. "The Adequacy of Corpora in Machine Translation." **Applied Computer Translation** 1: 15-28.

Shastri, S. V. 1980. "A Computer Corpus of Present-Day Indian English." **ICAME News** 4: 9-10.

---

* Dr. Eunice Nyamasyo is a Senior Lecturer, Department of English, Kenyatta University, Nairobi, Kenya.