Zambezia (1993), XX (ii).

RESEARCH REPORT

CORPUS OF ZIMBABWEAN ENGLISH AT THE UNIVERSITY OF ZIMBABWE COMPUTER CENTRE

W. E. LOUW

Department of English, University of Zimbabwe

and

JOSEPHINE JORDAN

Department of Psychology, University of Zimbabwe

Abstract

A corpus of Zimbabwean English comprising Dawson's Structures and Skills in English *and Grant* et al., English for Zimbabwe: an English course for secondary schools *is available on computer tapes at the University of Zimbabwe Computer Centre. Also on tape is the dictionary file which lists in alphabetical order all the words contained in the books, together with a frequency file, which lists the words in order of use. The corpus provides a readily accessible source of lexical items encountered in Zimbabwe secondary schools and demonstrates the employment of these lexical items in grammatical structure and idiom.*

BETWEEN OCTOBER 1987 and December 1993 a partial corpus of Zimbabwean English was captured and installed at the University of Zimbabwe Computer Centre. The corpus contains two sets of secondary school English Language textbooks, each comprising four volumes. The series are Dawson, *Structures and Skills in English* and Grant et al., *English for Zimbabwe*.

The corpus is housed on a 2 400 foot magnetic tape which occupies 3,5 megabytes when fully loaded and has 592 994 words (tokens) of running text. On the same tape are two lists for each book: a dictionary file and a frequency file. A dictionary file lists, in alphabetical order, all words in the corpus, otherwise known as types. The frequency file provides a list of frequencies in descending order. Examples of dictionary files and frequency files are given in Table I.

Even today language corpora are not a very widespread phenomenon. The reason for this is that, until recently, corpora have had to be developed by keying in material rather than by electronic loading. There are no more than four well-known corpora of English language in the world, the oldest of these being the Brown University corpus (Francis and Kucera, 1982) and the LOB (Lancaster, Oslo, Bergen) corpus (Johannson, 1978). Neither of these

two exceeds one million words. There is a third covering Indian English, the Kolhapur corpus (Shastry, 1988). The most recent, and by far the largest, corpus is the COBUILD (Collins Birmingham University International Language Database) which comprises 7,3 million words of running text in the main corpus, and which can be extended to 21 million words of running text using the reserve corpora. The COBUILD corpora took from 1978 to 1984 to load, mostly by Optical Character Recognitions (OCR) techniques using the KDEM (Kurzweil Data Entry Machine) at Birmingham University. One and a half million words of the COBUILD corpus comprise the greatest existing corpus of spoken text (Sinclair, 1987b). This sub-corpus had to be keyed in, as voice recognition loading is not yet sufficiently sophisticated to load spoken text reliably.

Table I

EXAMPLES OF WORD FREQUENCIES & DICTIONARY FILE LISTINGS OF *ENGLISH FOR ZIMBABWE BOOK I.*

Rank	*Word*	*No. of Occurr.*	*Rank*	*Word*	*No. of Occurr.*	*Entry No.*	*Word*	*No. of Occurr.*
1	the	2 930	39	write	191	1	a	1 463
2	to	1 520	49	words	176	2	aardwolf	3
3	a	1 463	59	sentences	153	3	aardwolfs	1
4	in	1 325	62	following	150	4	abaluhya	4
5	of	1 163	63	read	150	5	abbreviated	2
6	and	1 057	92	paragraph	90	6	abbreviation	3
7	you	1 057				7	abbreviations	5
8	is	760				8	abc	1
9	it	577				9	abel	1
10	he	552				10	ability	4

The first column indicates the frequency file rank of the word in the list.
The second column constitutes a continuation of the ranking of the list in the first column.
The final column indicates the frequency of the first ten words of the dictionary file.
Occurr. = Occurrences

The Zimbabwean corpus differs from the above corpora in that it is a specialized corpus drawn from the major English language textbooks encountered in Zimbabwean secondary schools. All the corpora mentioned above have the standard language as their main concern and draw their texts from many carefully sampled provenances. This was especially the case with the COBUILD corpus as the object of COBUILD was to produce the first computer concordanced dictionary of English (Sinclair, 1987a).

The objective in loading the corpus of Zimbabwean secondary school materials was to provide a resource of interest to academics, language teachers and curriculum planners, psychologists, writers, publishers, and those interested in specific areas, such as research in reading. Typical research objectives might be investigations into the mismatch between these materials and authentic Zimbabwean spoken and written text; comparison of key language terms in these books with, for example, the same terms in the Zimbabwe Hansard, *The Herald*, news bulletins, interviews and the like; comparison of the English curriculum and its coverage by the Zimbabwean textbooks with those for example, held in a one-million-word sub-corpus held by Renouf within COBUILD's research division.[1]

This report is designed to afford Zimbabwean researchers in many fields some insight into the contents of the corpus and the techniques of corpus linguistics, both of which may be of interest in their disciplines.

CORPUS DEVELOPMENT

As funding for the project was too limited to provide a general corpus of Zimbabwean English drawn from all textual provenances, spoken and written, it was decided that the first step in Zimbabwean corpus development should be the creation of a specialized language corpus. The series included in the corpus were chosen from the textbooks prescribed by the Ministry of Education in 1986 for Forms One to Four in secondary schools. For a full description of the sampling techniques for general corpus development, see Renouf (1987).

The appearance of the selected textbooks was not very promising for the successful loading on to the computer using OCR techniques. Both sets of materials were printed on local Mutare bond paper which affords insufficient contrast for OCR scanning. In addition the materials were broken up in two ways. Firstly, they contained many illustrations which disrupt the digitizing of the text. KDEM software is self-educating and runs more quickly as it becomes used to the fonts involved. Each time the scanner encounters illustrative material, this process has to be reinitiated.

The second way in which the texts were broken up is that pages were printed in two columns. This means that each column has to be masked off by the KDEM operator during loading. If this is not done, the KDEM 'eye' runs linearly across the page and incorporates the text for both columns into a single span. Newer KDEM software avoids this problem.

The texts were loaded at Birmingham University and edited within MULTICS screen edit to correct the numerous scanner errors. Editing took

[1] Antoinette Renouf, English Language Research Unit, Birmingham University. Since the corpus was installed it has been used successfully to support research into the difficulty of vocabulary tests (Jordan, 1989) and the use of personality words in Zimbabwean English (Jaynes, 1991).

the form, where possible, of global replacement of consistently misread fonts with the correct font or, where there was no consistency, strings from the alphabetical list had to be called up and corrected piecemeal. This process took in excess of 70 hours, working at off-peak times when computer reaction is at its best.

Once the 'dirty' copy had been cleaned up by editing, it was possible to draw word-frequency and alphabetical lists for both sets of books. This would disclose how effective the scanning and editing processes had been. The list generation processes operate on the machine-readable text once it has been converted into a string of single items each occupying one line or record. In UNIX operating systems (available at the University of Zimbabwe Computer Centre), the 'uniq' facility provides an intermediate stage for the generation of these lists and affords the option of word-frequency counting. It is a simple procedure to reorganize the list in alphabetical order or in order of descending frequency.

A tape labelled CORP01 with the full texts and their dictionary and frequency files is available at the University of Zimbabwe Computer Centre or from the authors. UNIX manuals are available from the manager.

GENERAL CHARACTERISTICS OF CORPORA

Until the advent of computational linguistics, there was a generally held belief that the words of a language fall into two separable and monolithic categories: that there are *full* words and *form* words (a distinction which can be traced as far back as Aristotle). Full words are generally words that contain some obvious semantic content, such as a referent, e.g. 'tree' or 'house', or are verbal in character such as 'fly' or 'run'. Form words were believed to be mainly grammatical in function, e.g. 'the', 'a', 'if' 'then', etc. The machine-based retrieval of full words from large corpora brought with it the surprise that, although these forms appeared to the intuition to have full lexical status in all of their uses, the more frequent the term, the more those uses shaded towards grammatical function. This phenomenon of 'washing out' of meaning has been described as progressive delexicalization by Sinclair (1987c). The most illuminating example offered by Sinclair is the term 'take,' for which there is only one intuitively recoverable full meaning but some forty-six progressively delexicalized meanings cited in the COBUILD dictionary (Sinclair, 1987a, p. 1489). At the far end of the progressive delexicalization scale, the term 'take' is readily interchangeable with grammatical terms of form word status. For example, it is not clear whether the status of 'take' in 'take a look at this book' is that of a full word or form word under the traditional labelling, given that it can be replaced by 'have a look' or omitted altogether as in 'look at the book'. Thus the distinction between full and form words is more in the nature of a continuum, or what linguists call a 'cline',

and a separate judgement can be made about each form of 'take', e.g. 'to take a bus' will be coded as more full than 'to take a bath'. The latter use forms the basis of much poor humour in pantomime and situational comedy, for example, as in the response 'where shall I take it to?'

At first sight, a word-frequency list appears to divide itself between the form words (most frequent) and the full words (those which make up the tail). However, in a general corpus, such as the COBUILD corpus, research has discovered that the 2 000 most frequent words are those which suffer the most progressive delexicalization. The COBUILD dictionary is the first dictionary to make a detailed computer-assisted grammatical and lexicographical description of the terms. Furthermore, these 2 000 terms form the basis of a lexical syllabus developed by COBUILD and incorporated into language-teaching materials for the study of English. The researchers see these 2 000 words as being so powerful a teaching tool that the emphasis in language teaching can now be swung with confidence away from grammar and into vocabulary. Mastery of 2 000 terms, in all of their uses, will carry the fundamentals of English grammar with them. In the case of COBUILD there are, of course, no made-up examples and the new COBUILD language materials (Willis and Willis, 1988), in common with the dictionary, draw all their examples from authentic text, although in the case of Willis and Willis such texts are often elicited from informants who become authentic characters within the texts; a far cry from the John and Mary of made-up examples.

It is this observation which makes the Zimbabwean corpus of particular interest. The Zimbabwean corpus is the product of the intuition of Zimbabwean materials writers. It carries with it the expectation that the lexical forms set out will cluster around full intuitive meaning, rather than delexical meaning, and where sentences are made-up examples, for use in pattern practice drills, they will lack what Sinclair (1988) calls 'naturalness'. Sinclair offers the following three sentences for coding as natural or unnatural:

1. We searched.
2. We searched all night.
3. We searched all night for the missing climbers.

Of the three, coders identify sentence two quite readily as the most natural, and sentence three, the sentence most likely to feature in language-teaching materials, where the examples have been made up. Made-up examples are always self-contextualizing to a degree which is unnatural in authentic text.

Because made-up examples are never a product of genuine interaction, such interaction as does take place using them will generally take on a ritualized form. These rituals are readily discernible if one studies a word-frequency list from specialized corpora such as English Language Teaching (ELT). Very often, high up in the envelope of grammatical words, words such

s 'discuss', 'paragraph' and 'groups' will be prominent. In other words, with pecialized corpora, some words become frequent because of their subject natter or ritual use and consequently move up into the grammatical envelope, vhile others, because of the way the text is divorced from authenticity, nove down.

LINGUISTIC CHARACTERISTICS OF THE CORPUS AND ITS RESEARCH POTENTIAL

The specific characteristics of a specialized corpus will be immediately apparent to those with any experience in corpus linguistics. However, to the uninitiated, the characteristics of such corpora often hold surprises. For example, it comes as a surprise to teachers of English that the language of classroom and textbook management and organization features so prominently in the word-frequency lists. In the list partly presented in Table I, as one descends from the most frequent form 'the' ('the' makes up about four per cent of all texts in English), and reaches words like 'following', 'words', 'write', 'sentences', 'read', it is remarkable to reflect how often these words appear in the texts and yet how infrequently their meanings are taught directly in the classroom.

A case in point is the word 'paragraph'. Teachers often express surprise that textbooks can contain more than 90 occurrences of a word, the meaning of which we all take for granted. There would be an argument for incorporating definitions of such basic sub-technical terms into the textbooks themselves. The list for inclusion could, of course, be furnished from an analysis of the corpus.

Corpora are equally revealing in their application to other disciplines. For example, researchers in reading have at their disposal a readily accessible source not only of lexical items encountered in the secondary school but also of the involvement of those items in grammatical structures and idioms. Indeed, grammatical structures can be assembled in a profile form for the entire text under discussion: for example, the percentage of active declarative sentences in relation to more complex forms such as relatives and passives. Information of this kind involves not only an examination of frequency lists but also of the transitional probabilities of each item. A concordancing programme or 'grep' within UNIX can provide this facility. The information discovered in this way can be cross-compared in reading research with other texts that the researcher might wish to match against the corpus.

Contemporary ethnographers do not leave an encounter with the corpus disappointed. There is strong prima facie evidence, in purely numerical terms, of sexism which must have a pernicious and profound effect in the first four years of secondary school.

Quite apart from the fact that normal expository text produces a higher proportion of the form 'he' than the form 'she', no amount of reasoning in

that direction could justify the discrepancies set out in Table II. Furthermore, if the actual spans associated with these forms are sought (Louw, in press), the case is instantly decided. The most frequent non-grammar word collocate of 'girl', in Louw's research, is the word 'marry'. There is no comparable form for 'boy', but the most frequent non-grammar word collocate for 'boy' is 'wonder' emanating, in its frequency, from a long story, sexist in character, entitled 'Wonder Boy'.

These are only some of the many research applications to which the Zimbabwean English Language corpus may finally be applied. The corpus will, doubtless, be enlarged into other provenances as OCR technology develops and, indeed, as Zimbabwean texts become available on floppy disks or CDROMs.

Table II

WORD-FREQUENCIES IN THE EIGHT BOOKS INCLUDED IN THE CORPUS

Frequency of selected personal pronouns			*Frequency of selected nouns*				
Text	**He**	**She**	**Text**	**Man**	**Woman**	**Boy**	**Girl**
1	552	180	1	110	42	44	37
2	702	259	2	105	63	29	17
3	828	161	3	136	39	24	21
4	626	199	4	62	42	9	15
5	441	174	5	75	10	18	13
6	674	77	6	60	22	7	4
7	692	71	7	91	21	3	7
8	539	69	8	101	20	25	10

Acknowledgements

The development of the corpus was undertaken with funds from the University of Zimbabwe Research Board and the British Council, with copyright permission kindly supplied by Longman Zimbabwe and College Press. The loading and analysis of the text took place at Birmingham University, for which special acknowledgement is given to Professor John Sinclair of the Department of English and Antoinette Renouf and Jeremy Clear of the English Language Research Unit.

References

DAWSON, D. 1984 *Structures and Skills in English: Book 4* (Harare, College Press).

DAWSON, D. 1985 *Structures and Skills in English: Book 1* (Harare, College Press).

DAWSON, D. 1986 *Structures and Skills in English: Book 2* (Harare, College Press).

DAWSON, D. 1986 *Structures and Skills in English: Book 3* (Harare, College Press).

GRANT, N. J. H. 1984 *English for Zimbabwe: an English course for secondary schools. Book 4* (Harare, Longman).

GRANT, N. J. H. and BIMHA, J. 1981 *English for Zimbabwe: an English course for secondary schools. Book 2* (Harare, Longman).

GRANT, N. J. H. and MAMUTSE, E. 1983 *English for Zimbabwe: an English course for secondary schools. Book 3* (Harare, Longman).

GRANT, N. J. H. and NDANGA, H. 1981 *English for Zimbabwe: an English course for secondary schools. Book 1* (Harare, Longman).

FRANCIS, W. N. AND KUCERA, H. 1982 *Frequency analysis of English usage: lexicon and grammar* (Boston, Houghton and Miffin).

JAYNES, K. 1991 'A bottom-up approach to personality testing' (Harare, University of Zimbabwe, Department of Psychology, Unpubl. dissertation).

JOHANNSON, S. 1978 'Manual of Information to accompany the Lancaster-Oslo-Bergen Corpus of British English for use with Digital Computers' (Oslo, Oslo University).

JORDAN, J. 1989 'Forging an environmental supports bank for vocabulary and verbal reasoning testing in Zimbabwe', *Psychology and Developing Societies, I*, 165–175.

LOUW, W. E. in press 'Computer assisted materials evaluation: content vs national policy', *English Language Research Journal*, III, 29–42.

RENOUF, A. J. 1987 'Corpus Development', in J. M. Sinclair (ed.), *Looking up : An Account of the COBUILD Project in Lexical Computing* (London, Collins), 1–40.

SHASTRY, S. V. 1988 'The Kolhapur Corpus of Indian English', *ICAME Journal, XII*, 15–26.

SINCLAIR, J. M. 1987a *Collins COBUILD English Language Dictionary* (London, Collins).

SINCLAIR, J. M. (ed.) 1987b *Looking up: An Account of the COBUILD project in Lexical Computing* (London, Collins).

SINCLAIR, J. M. 1987c 'Collocation: A progress report', in R. Steele (ed.), *Essays Presented in Honour of Michael Halliday* (Amsterdam, John Benjamins).

SINCLAIR, J. M. 1988 'Naturalness in Language', *English Language Research Journal, II*, 11–20.

SINCLAIR, J. M. and RENOUF, A. J. 1987 'A lexical syllabus for language learning', in M. J. McCarthy and R. A. Carter (eds.) *Vocabulary in Language Teaching* (London, Longman), 140–160.

WILLIS, J. R. and WILLIS, J. D. 1988 *Collins COBUILD English Course* (London, Collins).