*

CORPUS DEVELOPMENT AT BIRMINGHAM UNIVERSITY

1.0 INTRODUCTION

The English Department at the University of Birmingham has been working with text corpora for many years. Until recently, the state of computing technology made this a laborious process, but since about 1980 the growing sophistication of hardware and software for text-processing has eased the task somewhat and we have been able to make more rapid progress in the area of corpus development.

2.0 COMPUTING HARDWARE FACILITIES AT BIRMINGHAM

University machines which have been available to us in recent years are as follows:

- a Kurzweil Data Entry Machine (KDEM) supplied by Turnkey Software Ltd
- an ICL 1906A mainframe computer, with a GEORGE 4 operating system
- a DEC 2060 mainframe interactive computer
- a DEC PDP *11/34A* minicomputer, with 256kb RAM, a UNIX system, and 134 Mb of Winchester disk storage
- several Heath Z89 microcomputers, with 64k RAM and a CP/M operating system
- several CIT 101 VDU terminals to the PDP.

Of these, the last three sets of equipment are housed within the Cobuild project.

In the course of 1983, the ICL mainframe machine has been replaced by:

- a Honeywell 4 x DPS8/70M Processor, with the Multics operating system, (Release 10.1), and the MRDS database management system.

3.0 COBUILD

The underlying emphasis of work in Birmingham has, for some considerabe time, been the study and processing of essentially raw text. Recently this has found its full expression within a new section of the English Department known as the COBUILD project. COBUILD actually comprises a number of projects in computational linguistics. The largest of these is concerned with the building of a dictionary of current English, and is a very substantial enterprise, employing a large team of specialists in lexicography and computing.

All aspects of the work of COBUILD are related in some way to corpus analysis.

For the dictionary project it is fundamental, and it was this project which provided the stimulus for the development of what is now known as the Birmingham Collection of English Text, a large corpus of 'general' English.

Under the auspices of COBUILD, a second type of corpus has also been constructed. This is a smaller, more specialised, sample corpus, known as the TEFL Side Corpus. It is planned to build a series of specialised corpora in this mould over the coming years.

Q.O THE BIRMINGHAM COLLECT/ON OF ENGLISH TEXT

This is a body of written text and transcribed speech which currently amounts to over twelve million words and continues to grow. The long term plan is to treat it as a 'monitor' corpus (Sinclair, 1982) which can be manipulated to reveal insights into the state of the language at a given time. The process will involve the continual replacement of old data by new, so that the changing store of text can always reflect current linguistic behaviour.

Whilst this is an exciting prospect, it will be a little while before resources are adequate to allow us to embark on such an undertaking. As things stand, it is not justified to discard data after the effort currently required to obtain and process it. In the meantime, however, corpus data has to be made accessible to researchers, and the short term strategy has therefore been to extract an interim 'sample'-type corpus from the larger body of text. This static sub-corpus, known informally as 'the Corpus', since it is the entity with which researchers most regularly work, is made up of some 6 million words of written text, and 1.3 million words of transcribed speech.

For the remainder of this paper, the full corpus will be referred to as 'the Collection', and the sub-corpus as 'the Corpus'.

q.l THE PURPOSE OF THE COLLECT/ON

The Collection is intended to provide raw language data for a variety of purposes, some of which are already known and others which are expected to suggest themselves as it is explored. The COBU I LD use of the Collection has already been touched upon. At the moment, the 7.3 million word Corpus is under detailed analysis at the level of lex is, involving the observation of collocational and syntactic patterning, and of semantic, stylistic and pragmatic features, as they are revealed in the concordances. For this purpose, a corpus of 7.3 million words seems to furnish adequate data for lexical items in most frequency bands, and the general nature of the language content is proving largely appropriate.

Analysis at the level of discourse is a continuing preoccupation with colleagues within the English Department, and with this in mind the Collection was composed of long extracts and complete texts of approximately 70,000 words in length. Recent postgraduate research, notably work by Dr Martin Phillips on collocational patternings, has shown that provision to be justified. There is a growing demand for access to concordanced information on complete texts, and we are now processing our stock of text remainders so that the Collection contains complete works. I n so doing, we follow in the footsteps of colleagues who were involved in creating the Leuven Drama Corpus (Geens et al., 1975).

Q.2 SELECT/ONAL CRITERIA FOR THE COLLECT/ON

The intention was to build a store of text which reflects a broad spread of natural language usage in current English. This led to a decision, in October 1980, to impose the following constraints on text choice:

Texts should reflect current language usage.
For texts in book format, a post 1970 publication date was preferred, although some earlier works could also qualify for inclusion. Ephemera had to be post 1979.

Texts should, generally speaking, have been and continue to be widely read.
The argument here was that such texts were more truly representative of the language because they were influential in its evolution. In the case of major novels, this criterion overrode the one of recency, and allowed the inclusion of one or two older works where they still figured in required reading lists for native- and non-native-speakers. *Lord of the Flies* was one such example.

Texts should be couched in 'normal, adult, educated native-speaking English' .
We were committed to describing a wide range of linguistic activity, but set the limit at dialectal variation, children's language, subnormal speech,

non-native-speaker dialogue, and other intriguing but irrelevant language varieties.

- Poetry should not be included, although imaginative prose was acceptable.

In our judgement, poetry was essentially unrepresentative of mainstream linguistic behaviour, whereas fiction held too central a position in the realm of the written word to be neglected. Fictional writing can of course also be highly creative and idiosyncratic. A few texts showing this tendency were allowed into the Collection, but only where they fulfilled other selectional criteria.

- Drama should not be included, since the artificial dialogue of which it large ly consists was not our object of study.

Invented dialogue also featured in many of the chosen works of fiction. .In these cases, however, practical considerations generally led us to leave it in the text.

- Technical language was to be excluded.

In technical areas where the jargon has filtered into everyday use a selective exception was made. Such areas include economics, the arts, modern technology and psychology. Technical topics discussed in lay terms, as was the case in some spoken material, were of course acceptable.

Selectional criteria for the spoken component of our Collection were largely as for the written, and are implicit in the points above. To be explicit but brief, we wanted to cover the whole range of current, 'normal', but adult, 'educated', native-speaking English modes and registers of speech. As far as authenticity went, we were prepared to accept material which was less than spontaneous, but it had to be unscripted.

As the Collection progressed, its content was monitored so that obvious gaps could be filled. Attention was focused on the following areas:

age of author.
Authors were to be sixteen years of age or above at the time of writing.

sex of author.
There was considerable discussion about the proportion of female writing which should be included, but a minimum of 25% was finally agreed on. Feminist literature was included.

ethnic group.
This was only recorded where it characterised the style of writing.

language variety.
In terms of percentages, it was decided to include about 65-70% British, 25-30% American and 5% other varieties. The language in the Collection was to be predominantly British, since we were better equipped to deal with the analysis of this variety, but the American English component had to be sufficiently large to allow us to perceive major differences between the two varieties.

medium.
It was intended to include books, newspapers, magazines, brochures and leaflets, printed and handwritten letters and transcriptions of speech.

genre and topic.
A breakdown of the Corpus (see 4.2.1 below) in these terms can be found in Appendices 2 and 5.

4.2.1 The Corpus

The assembling of large text corpora is a complex matter. A large corpus grows unevenly, and it is impossible to maintain a precise balance among its various components at intermediate stages in its development.

Nevertheless, in November 1981, as the Birmingham Collection was still accumulating, we needed to make a representative extract available to COBUILD lexicographers. This required us to select for concordancing a restricted list of
titles which still retained something of the breadth of language offered in the original Collection. Our selection amounted to approximately 6 million words of text, which we judged to be sufficient for the needs of the lexicographers, and this was concordanced.

In the course of 1982, the balance of data in the Collection changed, and with it our perspectives. It now seemed appropriate to supplement the Corpus with fresh text, and its original 6 million word content was increased to 7.3 million words.

In future work, we now understand that the proper strategy is to aim initially for a lower target length in order to allow for the necessary process of supplementation.

An analysis of the content of the 7.3 million word Corpus, and statistics relating to its balance, can be found in Appendices 1-5.

4.3 COMPARISON BETWEEN THE COLLECTION AND THE LOB AND BROWN CORPORA

When the Collection is mapped against the text categories and subcategories of the BROWN and LOB corpora, a considerable degree of overlap is revealed, but there are also differences in content and weighting which are consistent with the different priorities involved.

One COBUILD emphasis lay in capturing current usage in the areas where innovation could be expected, so that our lexicographers' intuitions would be supported where they might otherwise be hazy. This led us to favour national newspapers, magazines, popular non-fiction and the newer phenomenon of 'faction'. Attention was also given to the presentation of topical technical issues in layman's terms. In our Collection, the disciplines listed in category 'J' of the BROWN and LOB corpora are broadly covered, but at a more popular level than is there stipulated. In line with our desire for topicality, we also favoured fictional and non-fictional treatment of such issues as race, war, and women's position in society.

Emphasis was also placed, as explained before, on processing long extracts of text. This meant that it was not economical to represent heterogeneous topic areas, such as 'farming', 'sport', 'hobbies', 'animals' and 'pets', since they threw up large numbers of specialised books, but relatively few examples of general treatment. They are therefore only selectively represented in book format at the moment, although newspapers and magazines also make their contribution in this area.

In terms of genre, our Collection differs from BROWN and LOB in that it excludes western fiction, but includes the academic novel, and a wider variety of handbooks and guides.

The spoken component of the Collection has of course no BROWN or LOB equivalent. Of the categories in the London-Lund corpus, most are represented, but in different proportions. The majority of our texts are of the discussion and interview type, and surreptitiously-recorded data does not feature.

4.4 SELECTION AND ACQUISITION OF DATA FOR THE COLLECTION

Having established in general terms the types of language we were looking for, we set about the task of identifying and acquiring actual texts. For the written component of the Collection, this began in December 1980, with an investigation of various sources which were likely to yield information about books and periodicals which could be regarded as representative of certain areas. We appealed to our own experience and to that of colleagues and postgraduate students, both native- and non-native-speaking. We studied the weekly best-seller lists and annual reviews published in British newspapers. To gain a broader international view, we circulated British Council librarians around the world to discover which reading matter was most popular in their libraries. The results of this questionnaire identified the all-time winners among foreign readership as being 'How to get your child into British Public School' and 'The London Telephone Directory', but otherwise the findings were largely as we expected.

On the basis of the information available, a first selection of titles was made. The next step was to obtain the appropriate copyright clearance from the publishers concerned. This was a tedious process, taking an average of six weeks, and in the case of some magazines requiring a separate approach for each article in a given issue. The effort was rewarded, however, by a large measure of success. Only two rejections were received, from agents whose clients had forbidden all reproduction of their work.

The procedure we followed in selecting and obtaining materials for the spoken Corpus was somewhat different. With a general idea in mind of the kinds of language we wanted, we approached various institutions which we felt to be potential sources of suitable data, and then made our selection out of what they were able and willing to supply. From the British Council in London we received a series of transcripts of informal conversation. The BBC offered us regular batches of transcribed material from three radio discussion programmes, which dealt with current affairs, finance and the arts respectively. These were unscripted broadcasts, with a degree of rehearsal in some cases.

We approached universities in the UK. Various Education, Psychology, Linguistics and English departments offered us data, but it was usually of the 'non-standard' variety. Then Sussex University came up with a collection of taped interviews made for local radio, in which university staff sought to explain to the layman the nature of their specialism and its relevance to the community as a whole.

Within the confines of our own university we found suitable transcribed material buried in theses, and were also able to use some of the large store of tapes of unscripted conversation which Professor Sinclair had recorded in the sixties. By the time the Corpus was extracted, there was a reasonable range of spoken data available in the Collection, and it was growing rapidly. It did not, at that point, however, include any spoken American English.

4.5 THE PROCESSING OF TEXT IN THE COLLECTION

4.5.1 Conversion to *Machine-readable form*

There were *two* methods of text processing at our disposal. One was to employ the newly-purchased and innovatory KDEM machine, a type of optical scanner. The other was the standard technique of keyboarding. We used both methods.

4.5.1.1 The KDEM

Our original hope was that the KDEM could handle all our material, since it had the capacity to process text significantly faster than a keyboarder could. In the event, the majority of the books were processed by the KDEM, but virtually all other material was keyed.

In the early stages of its operation, the limitations of the optical scanner became known. The major one was that it only functioned adequately for our purposes
on certain qualities of paper and print. When presented with loose-grained paper of the type found in many paperback books, where the print tends to run slightly, the KDEM was unable to distinguish characters which were no longer well-formed or totally discrete. On occasion, this inability to read characters even occurred with paper and print quality which to the naked eye seemed perfectly acceptable. The upshot of this discovery was that we were compelled to provide good quality, usually hardback, book copies. I n a few cases where hardbacks were simply not available, it was sometimes possible to coax the KDEM th rough the paperback version.

We also learned in the first weeks that transparent paper, such as is found in some books and in all newspapers, caused a problem of readability for the KDEM. As the scanner illuminated the text from below, it tried to read print on both sides of the page simultaneously.

Another reason for keyboarding not only newspapers but also magazines was the columnar layout of both. Although columns can be read by the scanner, provided it is given accurate specifications of column width and spacing, our need to maintain the integrity of articles conflicted with the habit of periodicals of splitting them across a number of pages. It was therefore uneconomical in preparation time to use the KDEM on these texts.

We further came to learn of other KDEM characteristics, such as its apparent problem with the reflective qualities of glossy paper, the unpredictability in its rate of throughput even under seemingly optimal conditions, and the apparent tendency for its error rate to increase in proportion to the number of hours it had been running. Our general observation was that the figures of throughput usually quoted for the KDEM, and those on which we based our original schedules, represent optimal rather than average performance.

It was, however, possible to compensate for the various delays through the introduction of a second terminal to the machine, and our overall experiences would still lead us to recommend the use of a KDEM where possible in text processing. The average performance of our early model is about twice the speed of keyboarding. The design is constantly being improved on, and with good staff it is an indispensable tool in text-processing today.

4.5.1.2 Keyboarding

Material unsuitable for KDEMing was, as said, keyed onto tape.

4.5.2 File transfer

Our Computer Centre wrote the required macro for the transfer of Corpus files to the ICL 1906A.

4.5.3 Concordanclng of the Corpus

The COCOA package was selected for concordancing in preference to other packages available at the time. The Oxford Concordance Package (OCP) was running but at that stage was untested on large files, whereas COCOA had at least been in use for many years. The CLOC package was felt to be too uneconomical in its consumption of random access disk storage for our large-scale purposes.

Traditional concordance packages are all wasteful of computer space and time. The usual strategy, and the one which we adopted, was one of file replication, by which a corpus of 6 million words produces an equivalent number of lines,
each of 120 characters. In the initial attempt to handle this amount of data we encountered two specific problems: one was that the Computing Centre was unable to provide the necessary disk space; and the other was that jobs requiring as much mill-time as ours did were automatically aborted.

Our first solution to this was to divide the then 6 million word Corpus into five batches and to write a macro to provide a restart facility to prevent overlarge jobs failing. Unfortunately this did not work because the filestore became exhausted before the restart could be implemented. We were then driven to take alphabetical slices through each batch, making a rough estimate of the size of each chunk of data in advance, so that it would be large enough to make optimal use of the space budget.

The concordance still took a long time, however, because introducing so many small jobs (250) actually exacerbated the problem of turnaround time. Two specific new problems were encountered: the first was that the 1906A automatic scheduling algorithm caused an increasingly low turnaround in response to concentrated demands on the processor. The second was that the COCOA package
required a dedicated diskpack and where, as in our case, there was only one diskpack available, only one job could be run at a time, whether large or small.

We finally achieved our objectives by resorting to a further series of strategies. With the co-operation of the Computer Centre we bypassed the 1906A scheduler; we set up a system of job-chaining, whereby each successful job automatically triggered submission of the next in line; and we hired exclusive use of the mainframe at weekends. To allow us maximum benefit from this extra time, the Computer Centre gave us access to two dedicated diskpacks. The turnaround at weekends was in the order of ten times that of weekdays, and after two weekends' processing the bulk of the concordancing was done.

4.5.4 Flchelng

The completed concordances were fiched by a commercial agency and available by April 1982.

4.5.5 Concordance fornKn

Concordanced data is presented in KWIC format, sorted by right context, and referenced to the left.

4.5.5.1 Interpretation of coding

The left-hand text reference consists of a series of characters and digits with the following values:

1	Text source:	G/T = General/Technical Corpus W/S = Written/Spoken component of the 000n = nth book in Corpus
2	Author's cultural identity:	BR = British AM = American, from USA AU = Australian AA = Anglo-American OT = Other (e.g. S African)
3	Country of publication:	BR = British AM = American, from USA AU = Australian AA = Anglo-American OT = Other (e.g. S African)
4	Page reference:	nnn

Thus GW00 43 br br 422

refers to page 422 of *The Mighty Micro* by C. Evans, which is book no. 43 in the written component of the COBUILD Corpus of general English.

W. B. The distinction between '2' and '3' above is made to alert corpus analysts to possible inconsistencies in spelling, syntax and other language features arising from the editing of foreign works according to particular publishing house style. An example of this is where a British publishing house will partially an

glicise the spelling in an American work, while leaving the syntax and idiom intact.

4.6 WORK SUBSEQUENTLY CARRIED OUT ON THE CORPUS

4.6.1 Enlarging the Corpus

As was explained in 4.2.1, the Corpus was 6 million words at the first processing, but later grew to 7.3 million words. On the second occasion, the written data was processed as before; the spoken data was concordanced on the PDP by means of a concordance program written for the purpose by ~.ie COBUILD Computer Officer. This particular batch of output was again in a KWIC-type format, although slightly different from the COCOA version. The enlarged Corpus was available on fiche from August 1982.

4.6.2 Concordance merging

The enlarged Corpus consisted of six separate batches of data. For ease of access it was decided to merge these. The merging was achieved by adapting the new PDP concordance program for use on the Computer Centre's new Honeywell mainframe.

Because the program was so straightforward and the Honeywell machine so powerful, it was possible to create one giant word index to all 7.3 million words and to sort that into alpha order. We could exploit the speed and power of the Multics Sort/Merge package for this phase of the merge, thereby saving programming and machine time. The concordance could then be generated from the index. The merged Corpus was available on fiche from November 1983.

To render the output more manageable, we did not concordance the 50 most frequent words of the Corpus. These will be dealt with separately at a later stage, but for present purposes they are still available in the original batched version.

4.6.3 Interactive concordance access

The PDP concordance program mentioned above has been developed to generate interactive concordances on demand. It allows the user to ask for a concordance to any specified word or words and to specify how much context is required. It also incorporates a simple selection algorithm so that control can be exercised over the number of available citations to be printed.

11.6.11 Hard-copy library

Lexicographic work of the COBUILD type cannot be done solely on screen, but also requires access to concordance data on paper. To meet this need with maximum efficiency, a hard-copy library has been set up, containing a complete set of Corpus concordances. The same data is of course still available on microfiche.

5.0 A *SPECIALISED SAMPLE CORPUS: THE TEFL CORPUS*

5.1 INTRODUCTION

We learnt a great deal in building the main Corpus. Expertise was gained in the handling of bulk language data; new text processing software was developed; corpus management skills were acquired. By 1982, we were ready to exploit these resources in the production of a series of smaller, more specialised, corpora. Our first choice was a corpus of the language of the EFL world, consisting of a number of course books used in the teaching of English as a foreign language, and informally referred to as 'the TEFL Side Corpus'.

5.2 USES FOR THE TEFL CORPUS

Our purpose in assembling it was to provide computerised access for the first time to the type of language which many learners of English undergoing formal instruction traditionally encounter. More spp.cifically, our objectives were as follows:

5.2.1 to achieve a detailed awareness of the lexical items and phraseology which are common to the majority of EFL course books, and therefore to some extent available to recent generations of formally-trained language learners. Wherever one wishes to communicate in English directly with a learner, it is valuable to' have as much information as possible about his/her familiarity with the vocabulary one intends to use.

5.2.2 to study the instructional language, or metalanguage, found in EFL books, in terms of rubric, page heading and instructional prefaces. It is a paradox not necessarily acknowledged in TEFL circles that such language represents a significant proportion of the total content of a course whilst not being treated explicitly in the teaching programme. It is not taken into consideration in devising a controlled language input to courses, and its placement does not conform to the received notions of language grading which otherwise underly the content selection procedure for such books. This is explicable when one considers EFL teaching practice hitherto. Learners are still not generally trained to adopt the initiating or managerial roles traditionally held by the teacher or the course material in the classroom situation. Nor are they taught to use language as the means of analysis of language.

5.2.3 to investigate the nature of the constructed language commonly found in EFL courses. The academic interest here lies in identifying how this language differs from its natural counterpart; what is it about concocted text that distinguishes it from the authentic equivalent. A more practical application would be to devise ways of bridging the gap between the language of the book and that of the real world – to establish a procedure, for example, for alerting the learner to qualities of 'naturalness' in text. Such awareness could also be put to use in the monitoring of teaching materials, and in a principled approach to the production of simplified readers.

5.3 SELECTION PROCEDURE

To cater for the various research interests, it was necessary to build a reasonably representative corpus, one which best reflected the language to which the greatest number of learners of English had been formally exposed over recent yea rs .

The British Council helped to obtain information on EFL book usage. In February 1982, we circulated a questionnaire around its offices abroad which sought to establish which EFL publications were most widely used in each country. The language officers (ELOs) were asked to identify practice within their own teaching operations (DTEOs) and in the host countries at large.

The answers to the questionnaire could have led to the creation of two different kinds of corpora. One would have been a collection of the EFL materials produced within each country to reflect the international language policy or to meet observed local needs. The other option was to collect a body of EFL course books which had been produced by major educational publishers for the international market. Both prospects had their attractions. In order to get a broad picture most economically, we decided to opt for the second type.

In terms of the actual titles listed, the questionnaire endorsed largely the experience of our EFL specialists, but it also gave valuable evidence of the relative popularity of the courses. From the information available, we made a selection of the major works. This was supplemented by a number of newer titles which are already showing promise in the UK and selectively abroad, and which could be expected to exert a similar degree of influence over current and future generations of learners.

The target size of a corpus is influenced by a variety of factors, the most important of which are the purposes for which the corpus is being assembled, and the cost. In this case, it was felt that the relatively small vocabulary and the assumed high rate of repetition suggested a fairly small corpus, and the initial
target was a little less than 1 million words. In the event, the corpus was slightly larger than anticipated, reaching approximately one million words, and made

up of some twenty-six books. These books together represent various stages in different course series, as can be seen in Appendix 6.

The books were to be processed in their entirety, with the exception of prefacing or interpolation addressed specifically to the teacher.

5.4 DATA PREPARATION

Before converting the texts to machine-readable form, we edited them, inserting extra- and intralinear codings. For reference purposes, each line of concordanced text was to be preceded by a code reference to the particular book in the corpus from which it came. Thus the line code 'e014' would identify an extract from *Starting Strategies.* Intralinear coding marked changes in language mode in the text. Mode was understood in the following terms and marked as indicated:

Mode	*Code*
instructional language	0
constructed spoken text	1
authentic spoken text in transcription constructed	2
written text	3
authentic written text	4

Of the text types, category '3' features most commonly in EFL books. This is not surprising in that it includes the core texts, dialogues, and exercises, all of which are artificially constructed to exhibit features of the language. Category '4' occurs least often in the corpus, since so few texts actually qualified as specimens of authentic writing. This was in spite of the large number of writtenextracts for which publishers' acknowledgements were provided. Acknowledged texts were discussed with the editors concerned wherever possible, and it was established that most texts which were presented as being 'authentic' had in fact been at least minimally abridged or altered.

To avoid confusion during the keyboarding process, each stretch of text within a given mode was colour-coded. The editing process was painstakingly carried out over a period of months by a research assistant, in collaboration with colleagues, and completed in December 1982. A full set of instructions regarding the treatment of aspects of the text, such as marking participant change in dialogue and gaps in exercises, was also prepared during this time.

5.5 PROCESSING OF DATA

5.5.1 Digitalisatlon of text

Experience with the main Corpus indicated that the paper and print quality of the EFL books would allow them to be KDEMed, but this proved impossible in view of the non-linear layout of the text. Course books published in recent years have been designed to achieve visual appeal, incorporating cursive script, a wide variety of type-faces, diagonally-set newspaper extracts, pictures, speech balloons, multi-coloured print and so on.

The bulk of the text was therefore sent out to a keyboarding agency in November 1982. Because of the coding insertions involved, it was work of a new order of complexity for the keyboarders, and progress was slow. A further delay was introduced by the method of verification, whereby the text was re-keyed by a second keyboarder to isolate discrepancies. Since the text was characteristically non-linear, the opportunities for deviation in the keying order were many, and time was wasted in this way.

The majority of the text was eventually verified, however, and further work on error reduction has subsequently been done inhouse. Analysis of single and double occurrences indicates that the degree of accuracy is now acceptable.

5.5.2 Concordanclng and ficheing

The concordancing of a 1 million word corpus is a relatively modest task, in COBUILD terms. In view of the imminent dismantling of the University ICL mainframe machine, it was decided to concordance the text on the local PDP minicomputer. This process was completed over one weekend, during which time a series of word lists were also extracted. The concordance format was a modification of the KWIC format used for the main Corpus concordances. As with the main Corpus, tapes of the concordances were sent for ficheing to an outside agency.

By the end of January 1983, the 1 million word TEFL corpus was ready and accessible, both on microfiche and on magnetic tape via the PDP.

5.6 POSSIBLE ENHANCEMENTS TO THE TEFL CORPUS

The corpus as it currently stands contains only the language that is printed in the chosen course books. It would be useful to supplement this data ill a number of ways, two of which are particularly obvious;

5.6.1 Classroom interaction based around the contents of given units in the corpus books could be recorded. This could include writing on the board, verbal reinforcements of (or deviation from) the book content by teacher or learner, and so on. This information would be particularly useful in relation to elementary books which actually contain very few printed words and rely heavilyon classroom reinforcement.

5.6.2 Supplementary course material, in the form of work books, audio or video tapes and so on could be processed for analysis. Taped material increasingly plays an integral role in EFL courses.

5.7 DISCUSSION OF THE TEFL CORPUS

Some work of a preliminary nature has been done on the TEFL corpus, and a series of observations have been made which will need further investigation. They are offered here to show the kinds of information which the TEFL corpus can provide:

5.7.1 Table 1: A rank listing of the 50 most frequent word forms in the TEFL corpus, compared with their ranking in four other corpora

	COBUILD TEFL corpus	LOB (1978)	BROWN (1961)	COBUILD Spoken Corpus	Jones and Sinclair (1974)
the	1	1	1	1	1
to	2	4	4	5	6
a	3	5	5	3	5
you	4	32	33	9	4
and	5	3	3	2	3
in	6	6	6	8	9
I	7	17	20	6	2
of	8	2	2	4	8
is	9	8	8	11	12
he	10	12	10	44	21
it	11	10	12	10	7
what	12	-	-	26	32
for	13	11	11	20	26
was	14	9	9	14	14
are	15	27	24	24	36
at	16	19	18	29	35
do	17	-	-	34	34
have	18	26	28	17	25
on	19	16	16	19	23
that	20	7	7	7	11
this	21	22	21	13	20
she	22	30	37	-	-
x	23	-	-	-	-
about	24	-	-	39	42
not	25	23	23	38	30
your	26	-	-	-	-
they	27	33	30	15	15
with	28	14	13	33	47
be	29	15	17	21	33
or	30	31	27	36	27
but	31	24	25	16	17
like	32	-	-	50	-
we	33	40	41	18	50
there	34	36	38	37	37
his	35	18	15	-	-
when	36	44	45	-	-
yes	37	-	-	22	10
had	38	21	22	-	-
did	39	-	-	-	-
can	40	-	-	47	-
how	41	-	-	-	-
very	42	-	-	35	49
as	43	13	14	25	40
no	44	47	49	45	18
if	45	45	50	40	39
I'm	46	-	-	-	-
go	47	-	-	-	-
my	48	-	-	-	-
from	49	-	-	-	-
it's	50	-	-	31	16

(Table 1 continued).

5.7.2 A comparison of the number of word forms in the TEFL corpus and the LOB corpus, also expressed as a ratio of type to token

Total number of word forms or types in:

the 1 million word TEFL corpus	approx.	24,000
the 1 million word LOB corpus	approx.	46,800

Number of word forms down to and including a frequency of 10 in:

Type-token ratio in:

the TEFL corpus the LOB corpus = 1 :42(.024) 1:21(.047)

5.7.3 The relatively high element of repetition, (or in pedagogical terms 'reinforcement'), in the TEFL corpus which is apparent in the type-token ratio can also be seen in the first order of occurrence word listings for individual books within the corpus. In Longman's *First Things First*, for example, the number of word forms which are repeated in the first 100 words of text is 87. The pattern of reoccurrence is as follows:

Table 2

Word form	*instance of first occurrence*	*locations of reoccurrences*	*total no. of occurrences*
excuse	1st word	20th word	2
me	2	21	2
yes	3	13 26 72	4
is	4	9 15 22 28 40 49 56 62 69 74 79 83 86 90 94 98	17
this	5	10 23 55 63 70 80 84 87 91 95 99	12
your	6	11 24 50 53 64 81 85 88 92 96 100	12
handbag	7	12 25	3
it	14	27 67 71 73	5
thank	16	29 43 75	4
you	17	30 44 76	4
very	18	31 77	3
much	19	32 78	3
my	33	36 41 58	4
coat	34	54	2
and	35	52	2
umbrella	37	51 59 65	4
here	39	48	2
sir	45	61	2
—			—
18			87

(NB 12 tokens in the 100 word text are not repeated)

The type-token ratio is 1 :3.3(0.3) which is high for a small text sample and when it is compared with the corresponding ratio in the initial 100 words in other genres of text:

First Things First The 1 3.3(0.3) (TEFL course book) 1
Mighty Micro 1.4(0.72) (expository text)
Changing Places 1 1.5(0.67) (novel)

This high reinforcement factor in EFL teaching material is of course to be expected.

5.7.4 Table 3: A rank listing of the 50 most frequent word forms in the instructional language of the TEFL corpus, compared with their ranking in the non-instructional text

	instructional	*non-instructional*
the	1	1
and	2	6
in	3	7
to	4	3
you	5	4
a	6	2
of	7	8
about	8	43
ask	9	177
your	10	40
write	11	279
this	12	23
these	13	144
for	14	14
are	15	19
questions	16	311
or	17	48
with	18	37
what	19	13
words	20	364
is	21	10
sentences	22	507
unit	23	437
at	24	17
do	25	16
answer	26	306
have	27	15
on	28	18
example	29	450
like	30	38
make	31	153
as	32	58

	Instructional	*non-Instructional*
not	33	24
that say	34	20
each	35	121
following	36	305
look	37	438
exercise	38	120
use	39	494
it	40	262
how which	41	11
study can	42	45
then	43	8
listen	43	411
now	44	43
other	45	96
give	46	389
	47	62 135
	48	151
	49	

(Table 3 continued).

Instructional language is understood to include rubric, page heading, and any other language used in organizing the lesson content and activities. See 5.2.2 above for further discussion.

5.7.5 Table 4: Semantic categories represented in the TEFL corpus and in the COBUILD Corpus for the word form OBJECT, together with the number of instances in each category

Semantic Category ie. OBJECT in the sense of:	*No. of Instances In the TEFL corpus*	*No. of Instances in the COBUILD Corpus*
1 grammatical term	47(69%)	
2 thing	16(24%)	176(58%)
3 the verb 'to object'	5(7%)	41(14%)
4 aim, goal	-	42(14%)
5 focus of attention	-	39(13%)
and in the phrases:		
6 'object lesson'	-	2(0.7%)
7 'money is no object'	-	1(0.3%)
Total occurrences	68(100%)	301(100%)

6.0 NOTES

* The creation of computer corpora relies on the sympathetic cooperation of many publishers, authors and other copyright holders, whose help is gratefully acknowledged. For both the corpora described in this paper, we are also indebted to the British Council for advice and information in the selection of texts, which enhance the authority of the corpora.

REFERENCES

ENGELS, L. K. (1981), *Leuven English Teaching Vocabulary-List, Based on Objective Frequency Combined with Subjective Word Selection;* Leuven: Dept. of Linguistics, Univ. of Leuven.
GEENS, D., L. K. ENGELS & W. MARTIN (1975), *Leuven Drama Corpus and Frequency List;* Leuven: PAL, I nstitute of Applied Linguistics, University of Leuven.
HOFLAND, K. & S. JOHANSSON (1982), *Word Frequencies In British and American English;* Bergen: The Norwegian Computing Centre for the Humanities.
JOHANSSON, S., ed. (1982), *Computer Corpora in English Language Re search;* Bergen: The Norwegian Computing Centre for the Humanities.
JOHANSSON, S., G. LEECH & H. GOOD LUCK (1978), *Manual of Information* to *Accompany the Lancaster/Oslo-Bergen Corpus of British English, for Use with Digital Computers;* Department of English, University of Oslo.
JONES, S. & J. MCH. SINCLAIR (1974), "English Lexical Collocations". In: *Cahlers de Lexlcologie,* 24, 15-61.
PHILLIPS, M. (1983), *Lexical Macrostructure In Science Text;* Birmingham: Dept. of English (PhD Thesis).
SINCLAIR, J. MCH. (1982), "Reflections on Computer Corpora in English Language Research". In: Johansson, ed. (1982).
SINCLAIR, J. MCH., S. JONES & R. DALEY (1970), *English Lexical Studies.* Report to OSTI on Project C/LP/08; Birmingham: Dept. of English, Univ. of Birmingham.
SVARTVIK, J. & R. QUIRK, eds. (1980), *A Corpus of English Conversation;* Lund: CWK GLeerup. (Lund Studies in English, 63).
SVARTVIK, J.', M. EEG-OLOFSSON, O. FORSHEDEN, B. tSRESTRtjM & C. THAVENIU!? (1982), *Survey of Spoken English: Report on Research* 1975-81; Lund: CWK GLeerup. (Lund Studies in English, 63).

Appendix 1

Content of the Written Component of the 7.3 Mill/on Word Corpus

Corpus Ref. No.	*Tltel*	*Author*
01	The Americans	A Cooke
02	The Third World War	Sir John Hackett
03	Superwoman	S Con ran
04	Life on Earth	D Attenborough
05	An Actor and his Time	Sir John Gielgud
06	Baby and Child Care	Dr B Spock
07	Manwatching	D Morris
08	The Fi re Next Time	J Baldwin
09	Lord of the Flies	W Golding
10	The Day of the Jackal	F Forsyth
12	The Companion Guide to London	D Piper
13	The Bedside Guardian 29	S Williams
15	Small is Beautiful	E F Schumacher
19	Tracks	R Davidson
21	Cosmopolitan (May 1981)	Various
22	Punch (May 1981)	Various
23	The Economist (May 1981)	Various
24	A Postillion Struck by Lightning	D Bogarde
25	Dispatches	M Herr
26	Bear Island	A MacLean
27	Elephants Can Remember	A Christie
28	Benefits	Z Fairbairns
29	Simple Steps to Public Life	P Anderson
30	What's Wrong with the Modern World	M Shanks
31	Future Shock	A Toffler
32	Zen and the Art of Motorcycle Maintenance	R M Pirsig
33	I n the Name of Love	J Tweedie
34	The Use of Lateral Thinking	E de Bono
35	Trout Fishing in America	R Brautigan
36	The Pendulum Years	B Levin
37	The Boys from Brazil	I Levin
38	The Next Horizon	C Bonnington
39	Changing Places	D Lodge
40	Summerhill: a Radical Approach to Education	AS Neill
43	The Mighty Micro	C Evans

Antoinette Renouf

Corpus Development at Birmingham University

44	Akenfield: Portrait of an English Village	R Blythe
50	Daniel Martin	J Fowles
51	Tell me a Riddle	T Olsen
52	The Human Factor	G G reene
53	The Glittering Prizes	F Raphael
54	Asking for Trouble	D Woods
55	Roots	A Haley
56	An Autobiography - Angela Davis	A Davis
57	Cosmopolitan (July 1981)	Various
58	The Illustrated London News	Various
59	Newsweek (May 11 1981)	Various
60	Kiss Kiss	R Dahl
61	How to. be an Alien	G Mikes
62	Inside the Third World	P Harrison
63	Jaws	P Benchley
64	Newsweek (27 July 1981)	Various
65	I'm Okay-You're Okay	T A Harris
66	You can get there from here	S MacLaine
67	Punch (29 July 1981)	Various
68	The Needle's Eye	M Drabble
69	Success without Tears	R Nelson
71	Rich Man, Poor Man	I Shaw
72	Lolita	V Nabokov
73	The Third World	P Worsley
74	It's an odd thing but...	P Jennings
75	Working with Words	J Mace
76	The Alienated: Growing Old Today	G Elder
77	Beyond the Crisis in Art	P Fuller
78	Kings of the Castle	G Eaton
79	Revolutionaries in Modern Britain	P Shipley
80	The Prerogative of the Harlot	H Cudlipp
81	But what about the Children?	J Hann
82	The War and Peace Book	D Noble
83	Tony Benn	R Jenkins
84	Love Story	E Segal
85	Punch (August 12 1981)	Various
87	Portrait of a Marriage	N Nicolson
117	A Backward Place	R Prawer Jhabvala
119	The History Man	M Bradbury
126	Homes and Gardens (October 1981)	Various
143	Punch Book of Short Stories	
152	National Geographic (January 1980)	A Coren
	D. H. S. S. Leaflets	Various
154	The Sunday Times Magazine	Various
155	(20 Jan. 1980) The Herald Tribune (25 July 1980)	Various
	Personal Letters	
156	The Guardian	Various
	(7 September 1981) The Guardian	Various
161		
200	The Times	Various
		Various
201		Various
202		

Appendix 2

An analysis of the written component of the 7.3 *million word Corpus*

2.1. The Book Component

2.1.1 N umber of texts: 66

single authorship 63 joint authorship 1 anthologies 2

2.1.2 Text size: an average of 70,000 words

2.1.3 *Genres and Topics*

2.1.3.1 Non-fiction

Genre	*Topics*	*No. of Books*
Exposition		
- essays	child care left-wing politics modern art and criticism Third World economics natural history social anthropology behavioural psychology computer technology C19th rural Britain old age	10
- articles	American people and culture various (Guardian anthology)	2
Procedure		
Handbooks	household management child care British institutions career advancement	4
Guidebooks	London topography and architecture	
Argument		
Polemic - essays	Third World politics nuclear weapons behavioural psychology micro-economics progressive education adult literacy Black America	7
- letters		
Discussion - essays	love and human relations contemporary society future society recent social history religion and modern society	5
Narration		
Travelogue	desert mountain	2
Biography	press barons British politicians British upper classes	3

Autobiography	theatre	
	cinema	
	rural life; cinema Vietnam war	
	modern technology and man	
	apartheid in S Africa Black America	
		7

2.1.3.2 Fiction

Genre	*Topics*	*No. of books*
General novels	human behaviour and relations	5
	slavery	1
	post- Raj Indian society	
		7
Mystery - short stories - human weaknesses		
Thriller novels	assassination murder at sea neo-nazism; cloning	
	espionage	
	American society; sharks	
		5
Detective novels	- murder	
Adventure novels	civilization; island castaways	
Academic novels	life-swapping Oxbridge	
	life in SO's provincial university life in 60's	
		3
Science fiction	abortion; future society	
	Third World War	
		2
Soap opera novel	- American society	2
		2
Humour collections of articles short stories	British society and institutions	2
	various (Punch)	1
		3

2.2 The newspaper component

2.2.1 Genres

circulation	*frequency*	*English lang. variety*	*amount of text*	*no.*
national	daily	British	all	3
international	daily	American	all	1
				4

2.3 The magazine component

2.3.1 Genres and topics

circulation	*frequency*	*Eng. lang. variety*	*topic*	*amount no. of text*	
national	wkly	British	review of current affai rs	all	4
national	wkly	British	economics	all	1
inter-national	wkly	American	current affairs	all	2
national	mthly	British	modern woman	all	2
national	mthly	British	home and garden	all	1
national	mthly	British	cu rrent affai rs	all	1
i nter-national	mthly	American	ethnic & cultural affairs	all	1
				-	12

2.4 Government documents

Genre	*Topic*	*No. of texts*
Department of Health and Social Security leaflets	retirement pension unemployment benefit death social security for school leavers and students cash help	5

2.5 Letters

Genre	*Topic*	*No. of texts*
Personal correspondence	- general family news	6

Appendix 3

Additional statistics relating to the book component of the Corpus - (single authorship only)

3.1 Date of publication

1980-81 6 1975-79 31 1970-74 13 1960-69 10 1950-59 3 pre-1950 1

64

3.2 Age of author at time of publication

16-25	1
26-35	19
36-45	20
46-55	11
56-70	10
70 +	3

64

3.3 Sex of author

male 49 female 15

64

3.4 Ethnic group

white 61 black 3

64

3.5 Language variety of author

British English	45	
American English	16	
Other	3 - consisting of	- 1 Australian -1 South African -1 Anglo-Russian

An Analysis of the Spoken Content of the 7.3 million Word Corpus

Appendix 4.

MEDIUM/MODE NO. OF SPEAKERS

	Radio	TV/ Video	face to face	private	1	2	3	4	5	6	7	8	several	total texts
face to face, informal conversation			+	+		4	8	1					2	15
telephone conversation				+		6								6
face to face discussion			+	+					1	1			1	3
lesson discussion incl. 1 teacher/lecturer			+			1	1						2	4
radio discussion, incl. chair/presenter	+		+			**11**	2	5	10	6	2	1	7	44
TV discussion, incl. chair/presenter		+	+			6								6
face to face formal interview			+			1	2							3
radio interview, incl. **interviewer**	+		+			48	10	2						60
videoed interview		+	+			1								1

	Radio	TV/ Video	face to face	private	1	2	3	4	5	6	7	8	several	total texts
personal narration			+	+	2									2
talk given in class			+		1									1
oral demonstration given in class			+		1									1
radio talk	+				2									2
university lecture			+		19									19
	106	7	159	26	25	78	23	8	11	7	2	1	12	167

Appendix 5

Text types and topics represented In the 1.3 *million word spoken component of the Corpus*

Text Type	*General Topic Area*	*No. of Texts*
face to face, informal conversation	various, domestic	13
	current affairs	2
		-
		15
telephone conversation	service encounters	6
		-
		6
face to face discussion	- education	3
		-
lesson discussion		3
	accountancy	2
	physics	1
	geography	1
		-
radio discussion		4
	current affairs the arts education	8
	finance energy	24
	issues technology	3
	law	3
		4
		1
		1
TV discussion	- politics	-
		44
		6
face to face interview		-
	interview for teaching appointment	6
	domestic matters personal preference	
		3
radio interview		
	government and law	3
	politics	3
	economics	2
	the arts	9
	religion	4
	education	6
	society and sociology	10
	language	1
	history	1
	biology	4
	technology	6
	physics	5
	other sciences maths	5
		1
		60
video interview	interview for undergraduate admission to university cou rse	
personal narration	- content of dream	2
		2
talk given to class	- British Education system	1
oral class demonstration	- math s	
radio talk	- physics	2
		2
university lectu re		
	perception	9
	artificial intelligence	1
	instrumentation biology	1
	technology	1
	various	3
		4
		19

Antoinette Renouf

Corpus Development at Birmingham University

Appendix 6

The Content of the TEFL Corpus

Code No.	*Title*	*Author(s)*	*Publisher*
EOO1	Encounters	J Garton-Sprenger et al	Heinemann Educational Books
EOO2	Kernel One	R O'Neill	Longman
EOO3	Kernel Lessons Intermediate	R O'Neill et al	Longman
EOO4	Kernel Lessons Plus	R O'Neill	Longman
EOO5	Access to English Starting Out	M Coles B Lord	Oxford University Press
EOO6	Access to English Getting On	M Coles B Lord	OUP
EOO7	Access to English Turning Point	M Coles B Lord	OUP
EOO8	Access to English Open Road	M Coles B Lord	OUP
EOO9	Streamline English Departures	B Hartley P Viney	OUP
E010	Streamline English Connections	B Hartley P Viney	OUP
E011	Streamline English Destinations	B Hartley P Viney	OUP
E012	Contact English 1	C Granger A Hicks	Heinemann
E013	Contact English 2	C Granger A Hicks	Heinemann
E014	Starting Strategies	B Abbs I Freebairn	Longman
E015	Building Strategies	B Abbs I Freebairn	Longman
E016	Developing Strategies	B Abbs I Freebairn	Longman
E017	Studying Strategies	B Abbs I Freebairn	Longman
E018	Follow Me 1	L G Alexander R Kingsbury	Longman
E019	Follow Me 2	L G Alexander R Kingsbury	Longman
E020	English Alive	S Nicholls et al	Edward Arnold
E021	English Alive 2	S Nicholls et al	Edward Arnold
E022	English Alive 3	S Nicholls C Wrangham	Edward Arnold
E023	First Things First	L G Alexander	Longman
E024	Practice and Progress	L G Alexander	Longman
E025	Developing Skills	L G Alexander	Longman
E029	Main Course English Exchanges (Pt A & B)	P Prowse et al	Heinemann