ICAME NEWS

Newsletter of the International Computer
Archive of Modern English (ICAME)

Published by: The Norwegian Computing Centre for the Humanities, Bergen
The Norwegian Research Council for Science and the Humanities

Machine-readable
texts in
English language
research

No. 4

Sept. 1980

CONTENTS

Editor: Dr. Stig Johansson, Department of English,
University of Oslo, Norway.

THE DUTCH COMPUTER CORPUS PILOT PROJECT

Jan Aarts and *Theo van den Heuvel*
University of Nijmegen, Holland

During the last few years members of the English Departments of five Dutch universities (Nijmegen, Amsterdam, Utrecht, Groningen and Leiden) have been engaged on a project called the *C*omputer *C*orpus *P*ilot *P*roject. The project is now in its final stage, viz. the computational analysis of the tagged corpus. The major aspects of the project are briefly discussed below.

1. CORPUS

The corpus is a rather small one, which was already available at the University of Nijmegen in computerized form. It numbers 126,000 words. Apart from 6000 words of spoken English (T.V. sports commentary), it contains six 20,000 word samples. Varieties included are: scientific prose, popular scientific prose, discursive prose (literary criticism), drama, and fiction. The category of fiction is represented by two 20,000 word samples of detective writing from two different authors, one of whom is also the writer of the fragment of discursive prose, so that analysis of these three samples may give some insight into the relation between register and idiolect.

2. SYNTACTIC ANALYSIS

The type of syntactic analysis that has been used to describe the syntactic structure of the sentences of the corpus is basically that of the *Grammar of Contemporary English*. It proved to be necessary, however, to adapt the GCE system in some ways and to render it more explicit. We cannot go into details here; we can only say that the CCPP system of analysis has proved sufficiently explicit to serve as the basis of a context-free grammar (see 4.3).

3. TAGGING SYSTEM

A four digit code was assigned to each word. The first two digits contain information about word classes, while the second pair of digits marks constituent boundaries. The first digit of the word class code indicates the major word class, the second carries information about subclasses or about morphological characteristics. For verbs there is, instead of a digit, a letter in the first position of the word

class code. The fact that it contains an alphabetical character signals that the word in question is a verb, the choice of letter indicates the subclass; in this way the second position of the word class code can be used to provide morphological (inflectional) information.

In the constituent boundary code the end of a sentence is marked by assigning 00 to the last word. The last word of each of the immediate constituents of the sentence receives code 01. Next, the last word of each of the constituents making up the sentence constituents is given code 02, and this process is continued until each word of the sentence has received its constituent boundary code. Some words of the sentence are, of course, at the boundary of more than one constituent on different syntactic levels and consequently would have to be assigned more than one boundary code. This problem is solved by giving to such a word only the lowest relevant constituency number, so that only the largest constituent is marked explicitly. The boundaries that are not explicitly marked by the constituency code are easily recovered in the following way. If a word x has a constituency code that is lower in number than that of word $x - 1$, word x is assumed to have also the constituency code numbers starting from its own code up to and including that of the preceding word.

The example below illustrates both the word class code and the constituent boundary code; after the word class codes the syntactic information that they stand for is given in parentheses, and after the constituency codes of words that are at the boundary of more than one constituent, the codes that are not assigned but implied are added.

The	21	(determiner, def. art)	02
roof	31	(common noun, sg, common case)	02
of	91	(preposition)	03
the	21		04
house	31		01 (02, 03, 04)
collapsed	A3	(verb, intransitive, pa. tense)	01
when	63	(conjunction, subordinator)	02
a	25	(determiner, indef. art)	04
bomb	31		03 (04)
fell	A3		00 (01, 02, 03)

Tagging was done manually, mainly for reasons of practicability. It was carried out by a number of graduate students at different universities. Before the actual tagging of the whole corpus was undertaken the consistency of their analyses was checked twice, by having all analysts tag the same text. This resulted in a number of improvements, revisions and additions, both in the system of analysis and in the tagging system. When this had been done, it was found that after three or four days of practice the average analyst could tag, approximately, 1000 words a day. All the tagged material was checked by one analyst - a job that could be done at the rate of 2000 words a day.

4. COMPUTATIONAL ANALYSIS OF THE TAGGED CORPUS

4.1 The initial data set

After the tagging and keypunching stages we have $\pm$ 140,000 punched cards containing a location code referring to the source text, the actual word (or punctuation mark, since some of the punctuation marks receive a tag), its word class code (WCC) and its constituent boundary code (CBC). Punctuation marks and other symbols that are not tagged are put immediately after the preceding word. To convert this data set to a more manageable form, we first make a lexicon of all words and punctuation marks (types). After that we replace every word and punctuation mark in the corpus by its pointer in the lexicon. During the analysis the lexicon does not play a role; analysis is only concerned with WCC and CBC. Instead of retaining the location code for every record we use a less direct and less space-consuming way of indicating location: after a word that is the last word of a line of the original text, we insert a special record representing a carriage return. Book and page numbers are only indicated at the start of a new page in the original text.

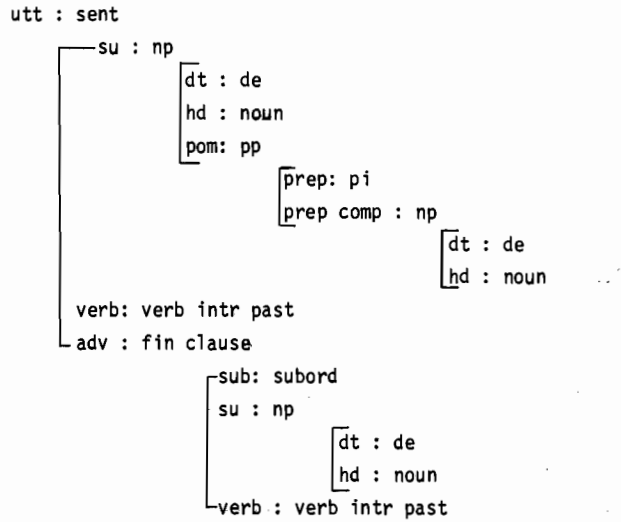
The data set that is the result of these conversions is called the reduced corpus. The reduction is one in size only; all the conversions are reversible and no information is lost.

For the work described in the rest of this section we use an IBM 370/158 under OS and some PDP 11/45's under UNIX.

4.2 Aim of the analysis

We want our software to yield two kinds of information: first, the category of a constituent (e.g. 'adjective phrase') and secondly its function in the constituent it is an immediate constituent of (e.g. 'premodifier'). Our example sentence

("The roof of the house collapsed when a bomb fell") would yield the following function-category tree:



The result of the analysis is thus a new data set relating, for every sentence of the corpus, a function-category tree to the lexicon and supplementary information.

4.3 A context-free basis

In order to automate the analysis which will yield results as exemplified above, the rather condensed and only partly explicit grammar of the *Manual for Coders* has been formalized in a context-free form. The limitations of a context-free grammar (cfg) for the description of natural languages are well-known. It should be borne in mind, however, that it is not our aim to describe English in full; given a sequence of codes, we only want a computer to tell us in what ways the grammar could have generated it. Moreover, most formal grammar systems are based, in one way or another, on a cfg. It is our purpose to extend our cfg so that it can also handle certain context-sensitive phenomena (see 4.5).

4.4 Parsing

It is not necessary for us to write a parsing program by hand. In the Nijmegen Department of Informatics a lot of software is available to manipulate formal grammars, including a program, written by Hans Meijer, which generates a parser

for possibly ambiguous cfg's.[1] This program accepts a cfg as input and outputs a parsing program. This program, in its turn, operates on character strings, yielding every possible analysis determined by the original cfg. More powerful generators are being developed at the moment, which can handle more powerful classes of grammars.

4.5 Extending the cfg

There are a number of reasons to extend our cfg. In the first place, certain context-sensitive aspects of the analysis of the corpus require such an extension. Secondly we want to reduce the bulkiness which is typical of cfg's. A third reason is that the parse trees yielded by the parsing program mentioned in 4.4 do not resemble the function-category trees of 4.2 at all, although they contain all the relevant information. The most efficient way to overcome this discrepancy is to extend our context-free grammar. This extension has not yet been fully accomplished. At the moment we can only say that it will be a restricted variety of a so-called 'affix grammar', a type of two-level grammar developed by C.H.A. Koster.[2]

5. THE CORPUS AND THE USER

For prospective users of the corpus it is not sufficient to make the analyzed corpus available in the form of a rather complex data set. Linguists should not be forced to become programming experts. To make the corpus more accessible for the average linguist, a 'query language' is being developed, i.e. an artificial language in which questions about the data can be formulated. Commands in such a query language are interpreted by a query system which has access to the data. In the CCPP work group discussion about the actual form of such a language took place at an early stage, in order to gain insight into the type of questions users of the corpus would tend to ask and to allow feedback to the grammar (and hence to the tagging system) so that linguistically interesting features would not remain concealed. To implement the CCPP query system (SYNTAXI) with the query language (TELL-ME) we will make use of the extended cfg's mentioned in 4.5. SYNTAXI will have access to the function-category trees, to the lexicon and to the reduced corpus.

PROSPECTS

So far, only an account of the manner of syntactic analysis and the tagging system has been given in a *Manual for Coders* (U. of Nijmegen, 1978). As soon as the first data from the tagged corpus are available it is intended to present these together with a full description of the system of analysis, the tagging

system and the computer programs.

As another direct result of the CCPP, a new project will start at the U. of Nijmegen in January 1981. Its aim is to develop a complete package of software which will enable users to retrieve syntactic information from a computer corpus (not necessarily an English one), allowing them to input grammars or parts of grammars of their own choosing. The project will issue its own bi-annual newsletter. Those interested in receiving the newsletter are requested to write to Jan Aarts, English Dept., U. of Nijmegen.

NOTES

1. Cf. Hans Meijer: *A Parser Generator for Possibly Ambiguous Context-Free Grammars on the IBM 370/158*, U. of Nijmegen, 1979
2. Cf. C.H.A. Koster: *Two-Level Grammars*, Mathematisch Centrum, Amsterdam, 1970

THE GILL CORPUS

J.M. Gill
Warwick Research Unit for the Blind
University of Warwick, Coventry, England

The corpus of text was created as part of a study on braille contractions. The braille system, in England, utilises 190 abbreviations and contractions. The use of those contractions is governed by a complex set of rules, many of which depend on pronunciation or meaning of the word. The project involved studying the effect of these contractions on:

(i) ease of learning
(ii) reading speed
(iii) writing speed
(iv) space saving
(v) ease of production

Space saving is a significant factor since braille is very bulky. For instance one copy of the Bible takes up 72 volumes (about a quarter of a cubic metre). Ease of production is also a factor because of the increasing use of computers to translate the ink print to contracted braille. A reduction in the number of rules dependent on pronunciation or meaning would reduce the cost of producing braille by computer-assisted methods.

One part of this project involved measuring the frequency of use of text strings. The Brown Corpus was used but it had the disadvantage that it was not typical of the material read in braille by blind people in the U.K.

Therefore a new corpus was created in an attempt to be more representative. Some of the material was in the form of short documents which had been requested by blind people for transcription into braille. These could be divided into:

Agendas and minutes	15%
Instruction booklets	6%
Employment	14%
Students' handouts	11%
Leisure (e.g. record sleeves)	7%
Songs and poems	1%

Recipes	2%
General information (e.g. government leaflets)	9%
Religious	1%
Timetables	1%
Accounts	4%
Correspondence	8%
Miscellaneous	21%

Added to this were samples of short stories and books, both fiction and non-fiction, which had been transcribed into braille by the Royal National Institute for the Blind.

The corpus contains 1030 short pieces in the English language giving a total of 2,561,308 words (a word being defined as an alphanumeric character string delimited by spaces or punctuation).

The corpus will be made available, for research purposes, on digital tape. For information, contact Louis Burnard, The Archive, Oxford University Computing Service, 13 Banbury Road, Oxford OX2 6NN, England.

A COMPUTER CORPUS OF PRESENT-DAY INDIAN ENGLISH

S.V. Shastri
Department of English, Shivaji University
Kolhapur, India

Under my direction, a team of workers at the Department of English, Shivaji University, Kolhapur (India) has been engaged for over a year now in assembling a computer corpus of present-day Indian English to match the American (Brown) and the British (LOB) corpora[1]. The corpus is primarily expected to serve as a data base which could be used as a source for purposes of comparative studies of American, British and Indian English, and secondly to produce, later, a description of present-day Indian English. If it is to serve the first purpose the Indian corpus has to be as similar to the other two corpora as possible. The Lancaster-Oslo/Bergen (LOB) Corpus approximates this mainly by sticking to the same sampling year as the Brown Corpus, and conforming very closely to the distribution and weighting of the fifteen different categories from which the texts are drawn. It departs marginally in respect of the distribution of some subcategories.

The Indian Corpus departs from the Brown Corpus in two ways. It does not stick to the sampling year of 1961 but draws its texts from writings published in 1978, i.e. nearly two decades later. Yet the compilers feel that this should not affect comparability, because the emergence of Indian English as a distinct variety of English is a recent phenomenon; and again the English language situation in India is more volatile than in England or America, while language change in England and America would be trivial over these years; hence comparability would not suffer much. It is also felt that a 1961 corpus (assembled after 1980) would be unhelpful for the second projected purpose, viz. for the description of present-day Indian English.

Another respect in which the Indian corpus is likely to deviate more from the Brown Corpus than does the LOB Corpus is in respect of the distribution and weighting of the different text categories and sub-categories from which the text samples are drawn. Of the two major categories - I. Informative Prose consisting of 374 texts and II. Imaginative Prose consisting of 126 texts, it is obvious that the overall production of the second category of writing is bound to be proportionately lower in a second language situation (as in India) than in a first language situa-

tion (as in America or Britain). Moreover, the genres of Indian writing are unlikely to match those of native text production. Nevertheless, it has been decided to maintain the native corpus weighting as regards the two major categories of writing. However, in respect of the second major category, i.e. Imaginative Prose, it may be necessary to alter the distribution and weighting of subcategories. This is because in the Indian situation certain categories of writing are not produced at all. For example, there is hardly any full-length novel representative of science fiction and hardly any full-length book representative of humour. Therefore there are likely to be more samples taken from short stories than from full-length novels or books in these categories.

At the time of reporting, a majority of the texts from "books" representing almost all the categories have been collected, as well as a number of texts from Press materials. It is expected that the collection of texts will be completed by the end of 1980. Coding of data and card-punching is already in progress. The corpus is expected to be available for use some time in 1982.

The initial phase of the project is being funded by Shivaji University.

NOTE

1. The other members of the team at the moment are Dr. V.V. Badve and Dr. P.R. Kher.

THE AUGUSTAN PROSE SAMPLE AND THE CENTURY OF PROSE CORPUS

Louis T. Milic
Cleveland State University, U.S.A.

The Augustan Prose Sample and the Century of Prose Corpus are period corpora of British prose covering parts of the periods called Restoration and Eighteenth Century. APS was begun in 1972 and consists of 52 selections by 51 authors nominally spanning the period 1675-1725 on the basis of the publication date. It began as an *ad hoc* compilation of minor writers designed to serve as a norm for a particular study, a comparison of the prose styles of Steele and Addison. The design originally required 50 selections of 2000 words each, one for each year of the span. Many difficulties (the lack of appropriate bibliographies of minor writers, obtaining the books, identifying time of writing as distinct from time of publication) intervened, making strict adherence to the original design impossible. Although it is a compromise, APS nonetheless has served its original purpose and has become, as far as I am aware, the only general-purpose corpus for the period it covers. The final product tape, which is available gratis to academic users upon request, consists of 79,208 words comprising 1650 sentences. The distribution of samples over the half-century is quite uneven, with twenty of the years not represented by any selections and three of the years covered by three, four, and six selections. Moreover, the selections are quite varied in length, ranging from 352 to 4298 words from 5 to 99 sentences. The two chronological halves of APS however (pre-1700 and post-1700) are nearly equivalent: 26 selections each.

The basic form of APS consists of sentence-length records in chronological order by selection, with original spelling (APS.CHROS). Other formats are in preparation: chronological order with regularized spelling (APS.REG), random order (APS.RAND), and single-word records (MONOAPS), chronological, old-spelling. A manual giving a detailed description of the contents is also in preparation and should be available by the end of 1980.

Certain shortcomings of APS led to plans for a larger and more thoroughly designed corpus. COPC (1680-1780) consists of two parts: Part A has 15,000-word samples by the twenty major prose writers of the time (Addison, Berkeley, Boswell, Burke, Chesterfield, Defoe, Dryden, Fielding, Gibbon, Goldsmith, Hume, Johnson, Junius, Locke, Richardson, Smollett, Steele, Sterne, Swift, Walpole); and Part B contains

100 selections of 2000 words each. The samples in Part A include several selections examining each writer's productive life. The Part B samples cover ten categories (travel, science, education, periodicals, biography, learning, essays, polemics, fiction and letters), one selection for each decade. The Part B samples constitute the background or norm for the writers whose work is collected in Part A. Any decade or category will be accessible vertically or horizontally. The whole will be available only in regularized spelling (American variants), when it is completed. As of this date, 95% of Part B has been encoded, and 35% of Part A. It is doubtful whether the complete tape (500,000 words) will be available before July 1981, though Part B may be.

Encoding practice has required extensive improvisation, as no standard encoding conventions exist. The Brown Corpus method is unsuited to the texts under consideration. Apart from the fact that the text is all upper-case (with $ for proper names) the main deviation from normal text is the limitation on the use of sentence-final punctuation (only period and question mark), neither of which may be used medially. A complete description will be available later.

COMPUTATIONAL TEXT ANALYSIS AT THE UNIVERSITY OF BIRMINGHAM

John McH. Sinclair
University of Birmingham, England

The University of Birmingham is pursuing a strong interest in computational text management and analysis, as a number of different interests converge on this area.

(a) Linguistic computing began in the middle sixties with a research project to study lexical patterns in transcribed conversations.[1] This study demonstrated that word-collocation was an important feature of texts, and that it was distinct from syntactic structure, and applied to all words, and not just "vocabulary" words. It also indicated that the available texts of up to a million words[2] were much too short to allow any comprehensive lexical description of a language or even a variety of it.[3] But the computers of the day had difficulties with long texts, and the software was quite inadequate, so it was decided that effort should be concentrated on encouraging the provision of the right conditions for research of a different order.

(b) The Department of Computer Science developed an interest in the teaching of computing, and made a series of video films on the subject.[4] One part of this was the problem of teaching Arts students, who lack the mathematics background that used to be taken for granted - unnecessarily - in the teaching of computing. Courses have been developed, usually in conjunction with other Departments, and the current picture is as follows:

Undergraduate courses:
A one year subsidiary course for students in the Faculty of Arts.
A seminar course for English students - 'Computer Aided Study of English Texts'.
A course for students of Linguistics.

Postgraduate courses: There is nothing specific for postgraduates but in some cases it may be possible for students with a first degree in Arts to do an M.Sc. in Computer Science. The English Department runs a regular seminar for staff and postgraduates.

Future: The Computer Centre is at present drafting regulations for a Combined Honours course, of which half will consist of Computer Science. It will be a three year course for the degree of B.A.

(c) It was understood some years ago that much valuable information could be retrieved from long texts by researchers without the need for strenuous training in computer science. Accordingly, a package was designed to give easy access to word-counts, concordances, co-occurrences and collocations. This is called CLOC[5], written by Alan Reed of the Computer Centre. It is used in batch processing, and has been in use and frequently upgraded for about 5 years. The researcher has to write about half-a-dozen simple instructions, to identify the text he wants and specify the CLOC operation, and the results appear in a convenient format. Word counts may be produced in alphabetic, reversed alphabetic (starting with the last letter of a word), frequency or text order; concordances are in KWIC format and selections may be made according to a variety of conditions; co-occurrence outputs show the patterns of phrases (not necessarily continuous) and collocation outputs place one word in the centre of a span of n words and show the tendency of other words to collocate with that word. CLOC is easy to use and reliable, and operates both as an introduction to beginning Arts students and a primary research tool.

(d) Linguistic needs overlap with other areas of information retrieval where numerical operations are not central, and the need grew in several quarters for a programming language which put a priority on easy handling of files, and which did not require great skill in computing. Members of the Department of Computer Science devised such a language and called it ATOL - a text-oriented language.[6] To the user, it is similar to Basic in the way in which the programmes are built up, but whereas Basic is very clumsy with text files, ATOL is elegant and, after a little practice, easy to think in. It is designed for interactive use, and its feedback gives the programmer helpful guidance as he builds up a program. As well as its immediate utility, ATOL is valuable as a means of communication between the researcher in Arts subjects and the computer scientist - as a sort of inter-language where natural language is not precise enough. Programs can be written in ATOL to the satisfaction of the user, in that he gets the results that he wants in a trial run, and understood as a logical structure by the professional programmer, who can then advise on implementation, since ATOL may not be the most efficient vehicle, for complex work on long texts, for example.

ATOL has been running on several machines for up to two years, and has proved flexible and labour-saving. One program, written by an English undergraduate, guides a complete beginner through setting up a CLOC job, and outputs the CLOC control information.

(e) The bottleneck of keyboarding long texts is being eased by modern technology in two ways. One is computer typesetting, where material to be printed is held in machine readable form. Because of the welcome co-operation of the Birmingham Post & Mail Ltd., the University computers now have access to virtually unlimited quantities of this output, and programs have been written which make the text capable of being processed in the established routines.

The other new developments are the document readers, which examine typed or printed text optically, identify the characters, and record the information on magnetic tape. A project soon to begin in Birmingham, financed by the British Library, is to study the feasibility of "electronic journals" and a document reader is on order for this work. Experience gained in its use will open up access to unpublished material which will not be available through computer typesetting. One important area for language study here is that of transcribed spoken text, which is at present very laborious to prepare for the computer. Until the next technological breakthrough in the automatic transcription of conversation, which seems a long way off, the document readers will supply an important link.

The last few years, then, have prepared the ground for text study in several different ways. A lot of effort has gone into attracting Arts scholars into computing, de-mystifying it for non-mathematicians and refining the tools so that new work can be done which was not practicable before. The accent has been on access to language text, preserved as close to its natural occurrence as possible. Any organisation that leaves traces in text itself is in principle retrievable over texts of any length.

Length is felt to be important. Even a glance at the statistics of word occurrence suggests that to gain access to the characteristics of language, one requires texts of a length that puts them well out of scale of direct human observation. About half of any running text is made up of occurrences of very frequent words; vocabulary items that occur frequently usually do so because they have several distinct meanings; in texts of a million words, getting on for half the words occur once only, so little can be discovered about them. Fixed phrases, variable phrases and idioms further muddy the waters.

Linguists do not take readily to indirect observation of language; they paused at the sentence for some 2,500 years, and are not yet properly equipped for going much beyond it. One danger that can be foreseen is that they will assume that the

descriptive categories appropriate to sentences (deriving from the patterning that Saussure called *langue*) will be imposed, with some ingenuity and much labour, on text (Saussure's *parole*). For this kind of reason research at Birmingham has not explored automatic category labelling, even lemmatisation, though facilities are available for this when it is required. We intend to continue the work of making text features accessible to researchers, and are currently investigating more efficient ways of processing extremely long texts.

One principle of the work is that we should design systems that will answer questions that have not as yet been asked; that will be so flexible that any conceivable text pattern can be studied immediately, instead of incurring the cost and waiting time of writing special programs for each new investigation. The economics of computing change continuously, and we are now at a stage where storage is getting cheap, while processing time remains costly, however fast, and high-level languages use a lot of machine time although they save enormously in human effort. It may be possible to exploit this situation.

NOTES:

1. OSTI Project C/LP/08, *English Lexical Studies*. Final Report in the British Library.
2. E.g. the Brown Corpus.
3. See OSTI report and Jones and Sinclair, "English Lexical Collocations", *Cahiers de Lexicologie*, no. 24, 1974, 15-61.
4. Films on computing made by the TV & Film Unit, University of Birmingham, between 1970 and 1976.
5. CLOC USER GUIDE, written by Al Reed, Computer Centre, University of Birmingham, 1978.
6. *ATOL – A Simple Language with Powerful Data Structuring Facilities*, Axford, Burkhardt, Dodd, Laflin, Parkyn and Ramsay (1978), available from the Computer Centre, University of Birmingham.

ICAME PROJECTS

Though this issue of our newsletter has concentrated on other projects, work has also advanced on the three main ICAME projects. As reported before, a grammatically tagged version of the Brown Corpus has been produced at Brown University, where lemmatized frequency lists are currently being prepared. Grammatical tagging of the London-Lund Corpus is in progress at the University of Lund. Similar work has started on the LOB Corpus through cooperation between the universities of Lancaster, Oslo, Bergen, and the Norwegian Computing Centre for the Humanities. A number of grammatical and other studies are in progress using the three corpora.

The printed version of part of the London-Lund Corpus announced in *ICAME NEWS* 3 has now appeared: Jan Svartvik and Randolph Quirk, *A Corpus of English Conversation*, Lund Studies in English 56, Lund: CWK Gleerup. Copies can be ordered from: Liber, Box 1205, S-221 05 Lund, Sweden.

Word frequency lists for the LOB Corpus (rank list, alphabetical list, a comparison of frequencies in different text categories) as well as lists comparing word frequencies in the Brown Corpus and the LOB Corpus have been produced but are not yet available in printed form.

The next issue of the newsletter will bring more detailed reports from our current work including lists of publications and work in progress. *Users of computer corpora, in particular the three corpora distributed from Bergen, are asked to submit information on publications and work in progress (use the form on the next page).* Relevant information will be reproduced in the newsletter. In this way we hope to achieve one of the main aims of ICAME, which is to make possible and encourage the coordination of research effort and avoid duplication of research.

PUBLICATIONS/WORK IN PROGRESS

Published books/articles using computer corpora:

Unpublished reports:

Work in progress:

Name and address:

(Return to: Stig Johansson, Department of English, University of Oslo,
P.O. Box 1003, Blindern, Oslo 3, Norway)

MATERIAL AVAILABLE FROM BERGEN

Two versions of the Brown Corpus are available:

Text I: Typographical information is preserved; the same line division is used as in the original version from Brown University except that words at the end of the line are never divided.

Text II: Typographical information is reduced; the line division is new.

The tagged version can only be ordered from: Henry Kučera, TEXT RESEARCH, 196 Bowen Street, Providence, R.I. 02906, U.S.A.

The LOB text can be obtained on tape from Bergen. KWIC concordances for the LOB Corpus and the Brown Corpus are available on tape or microfiche. The microfiche set for the Brown Corpus, but *not* for the LOB Corpus, includes the complete text of the corpus. A printed manual accompanies the tape of the LOB Corpus. Printed manuals for the Brown Corpus cannot be obtained from Bergen.

Part of the London-Lund Corpus, consisting of spontaneous conversation with prosodic coding, is available on tape from Bergen. The printed version, which is a necessary tool for the user of the magnetic tape, can only be ordered from the publishers (see above) or their representatives.

The material has been described in greater detail in previous issues of *ICAME NEWS*. Technical specifications are given on the order form at the end of the newsletter.

EDITORIAL NOTE

Further ICAME newsletters will appear irregularly and will, for the time being, be distributed free of charge. The Editor is grateful for any information or documentation which is relevant to the field of concern of ICAME.

ORDER FORM

To obtain material you must *enclose a cheque (bank draft, cashier's cheque) with your order made out in Norwegian currency (or the equivalent in English pounds or U.S. dollars) to:* The Norwegian Computing Centre for the Humanities, Bergen, Norway.

For *postage and handling*, add to the prices given below:				
for each 1200 ft. tape	25	Norwegian	kroner	
for each 2400 ft. tape	35	"	"	
for each microfiche set	10	"	"	(overseas air mail: 20)

(Use of some of the material for research by commercial institutions may be possible on the same conditions. *Commercial institutions are, however, asked to contact us before ordering any material.*)

(Return to: The Norwegian Computing Centre for the Humanities,
P.O. Box 53, University of Bergen, 5014 Bergen, Norway)

Indicate in the table below what you wish to order (prices are given in Norwegian kroner):

Magnetic tapes	9 track, 1600 FPI Number of tapes	Price	7-track, 800 FPI Number of tapes	Price	7-track, 556 FPI Number of tapes	Price
Brown: Text I	1 1200 ft.	175	1 2400 ft.	275	1 2400 ft.	275
Brown: Text II	1 1200 ft.	200	1 2400 ft.	300	1 2400 ft. 1 1200 ft.	300
Brown: Texts I+II	1 1200 ft.	250	1 2400 ft. 1 1200 ft.	525	2 2400 ft.	550
Brown: KWIC concordance	4 2400 ft.	1200	11 2400 ft.	3375	15 2400 ft.	3900
LOB: Text	1 1200 ft.	175	1 2400 ft.	250	1 2400 ft.	250
LOB: KWIC concordance	5 2400 ft.	1400	12 2400 ft.	3650	16 2400 ft.	4100
London-Lund: Text	1 1200 ft.	175	1 1200 ft.	200	1 1200 ft.	200

☐ Brown: KWIC concordance (microfiche) Price: 350

☐ LOB: KWIC concordance (microfiche) Price 350

Name/Institution: _____

Address: _____

_____

We understand that the material listed above is for research purposes only and agree not to distribute the material or reproduce any part of it for any other purpose than scholarly research. (To be signed by a responsible official of the institution making the order.)

Date _____ Signed _____

Print name _____

Official Position _____

ICAME NEWS is published by the Norwegian Computing Centre
for the Humanities.
Address: Harald Hårfagresgate 31, P. O. 53, 5014 Bergen - University, Norway