ICAME NEWS

Newsletter of the International Computer Archive of Modern English (ICAME). Published by NAVF's EDB-senter for humanistisk forskning, P.O. Box 53, University of Bergen, N-5014 Bergen, Norway. Editor: Dr. Stig Johansson, Department of English, University of Oslo, Norway

Number 1 March 1978

BACKGROUND

In February 1977, at a meeting of a small group of linguistic scholars in Oslo, a proposal was made to form an International Computer Archive of Modern English. The aims of the organization were specified in the following way in a letter sent out to scholars with an interest in computerized linguistic research:

"(1) collecting and distributing information on English language material available for computer processing;

(2) collecting and distributing information on linguistic research completed or in progress on the material;

(3) compiling an archive of corpuses to be located at the University of Bergen, from where copies of the material could be obtained at cost."

The participants at the meeting in Oslo were: W. Nelson Francis (Emeritus Professor of Linguistics, Brown University), Stig Johansson (Reader in English, University of Oslo), Geoffrey Leech (Professor of Linguistics and Modern English Language, University of Lancaster, U.K.), Arthur O. Sandved (Professor of English Language, University of Oslo), Jan Svartvik (Professor of English, University of Lund, Sweden). Jostein H. Hauge

(Director, NAVF's EDB-senter for humanistisk forskning, Bergen) took part in preliminary discussions.

One of the main aims in establishing the organization is to make possible and encourage the coordination of research effort and avoid duplication of research.

Initially, the archive will include three corpora: (1) the Brown University Corpus of American English ('The Brown Corpus'), (2) the complementary corpus of British English texts which has been under preparation for a number of years at the University of Lancaster, U.K., and which is now being completed in Norway ('The Lancaster-Oslo/Bergen Corpus'), (3) the spoken texts of the Survey of English Usage (University College London), which are now being transferred to magnetic tape for computer processing at the University of Lund, Sweden ('The London-Lund Corpus'). As needs and opportunities arise, the archive will be extended to include other corpora.

During the year work has continued on the three corpora mentioned. The present newsletter describes the status of the three projects and specifies the services offered or planned by ICAME in connection with the projects.

THE BROWN CORPUS

The well-known Brown Corpus, which was prepared in 1963-64 at Brown University under the direction of Professor W. Nelson Francis, has been available at cost since its completion. 150 copies of the computer tape containing it are in use at universities and other research centres all over the world. The work in assembling the Corpus and preparing it in machine-readable form is described in a forthcoming article:

> W. Nelson Francis, "Problems of Assembling and Computerizing Large Corpora", to appear in the volume on *Text-Corpora* edited by H. Bergenholtz & B. Schaeder (see below).

Over the last few years researchers at Brown University have been working on a grammatically tagged version of the Corpus. This phase will be described in:

W. Nelson Francis, "A Tagged Corpus-Problem and Prospects" (to be published).

The Brown Corpus is available in two formats:

A: with full graphic coding

B: with all external punctuation and other coding stripped off.

Both of these are available in either a 7-channel version (556 or 800 characters per inch) or a 9-channel version (800 or 1600 characters per inch). These are at present available only from the Department of Linguistics, Brown University. Those interested in acquiring a tape should write for a special order form.

Also available is a complete index to the Corpus, on a single tape. All tapes cost $75, post free, or $40 if you supply your own tape.

The tagged tape is not yet available for distribution, and probably will not be for at least a year.

NAVF's EDB-senter for humanistisk forskning, Bergen, has produced a version of the Corpus (the untagged version) with upper- and lower-case letters and other features which reduce the need for special codes and make the material more easily readable. A microfiche version of the text and a complete KWIC concordance are now available. This is described in greater detail at the end of the newsletter.

Inquiries on the Brown Corpus can be directed to: Professor W. Nelson Francis or Professor Henry Kučera, both of whom may be addressed at Box E, Brown University, Providence RI 02912, U.S.A.

THE LANCASTER-OSLO/BERGEN CORPUS

The aim of the project is to produce a British English equivalent to the Brown Corpus of American English. The project was originated and carried almost to completion at the University of Lancaster, U.K., under the direction of Professor Geoffrey

Leech. It has been described in the following article:

> Geoffrey Leech and Rosemary Leonard, "A Computer Corpus of British English", *Hamburger Phonetische Beiträge* 13 (1974), pp. 41-57.

In February 1977 it was agreed that the project would be completed at the Department of English, University of Oslo, in collaboration with NAVF's EDB-senter for humanistisk forskning at Bergen. A grant for this purpose was provided by the Norwegian Research Council for Science and the Humanities (NAVF).

The project is expected to be completed during the course of 1978. After this time the same services will be offered by ICAME as in connection with the Brown Corpus. For detailed information, see later newsletters.

Inquiries on the project can be directed to: Dr. Stig Johansson, Department of English, University of Oslo, P.O. Box 1003, Blindern, Oslo 3, Norway.

THE LONDON-LUND CORPUS

The Survey of Spoken English was started at Lund University in 1975 by Professor Jan Svartvik, in close co-operation with the Survey of English Usage, University College London. The primary aim of the project is to make available in machine-readable form the unique spoken English material collected and transcribed at University College London under the direction of Professor Randolph Quirk. The project is described in:

> Jan Svartvik, "Projektet engelskt talspråk". Report from the Survey of Spoken English. Lund 1976.
>
> Randolph Quirk and Jan Svartvik, "A Corpus of Modern English", to appear in the volume on *Text-Corpora* edited by H. Bergenholtz & B. Schaeder (see below).

The Spoken English material, which is in orthographic form with prosodic analysis, will be made available on computer tape (as well as in printed form). This part of the project is expected to be completed by 1979, by which time copies of the magnetic tape can be distributed via ICAME.

The second phase of the project involves grammatical tagging of the material. Later newsletters will contain further information on the project and on the services offered by ICAME in connection with the project.

Inquiries on the project can be directed to: Professor Jan Svartvik, Survey of Spoken English, Porthuset, Allhelgona Kyrkogata 14, S-22362 Lund, Sweden.

OTHER CORPORA OF MODERN ENGLISH TEXTS

During the year information has been collected on other computerized corpora of Modern English, some of which may later be included in the archive. Future newsletters will contain information on such material and specify possible services offered by ICAME in connection with the material.

We are grateful for further information on Modern English material prepared for computer processing. Do you have a corpus available which you would be willing to share with others? If so, under what conditions? Do you know of other researchers who have compiled corpora of Modern English which might be included in the archive? Write to: Dr. Stig Johansson, Department of English, University of Oslo, P.O. Box 1003, Blindern, Oslo 3, Norway.

OTHER TEXTUAL ARCHIVES

In connection with the distribution of the ICAME announcement we have received information that there are already, or are being planned, other computer archives of modern English texts (or including modern English texts). The Oxford University Computing Laboratory is establishing an archive of English literature. At the University of California, San Diego, there are plans to form an archive of computer readable texts in modern languages. Both of these enterprises seem to have a distinct literary bias, whereas ICAME will mainly concentrate on material especially prepared for, or especially suitable for, linguistic analysis.

Very similar to ICAME, though wider in scope, is the Stanford Computer Archive of Language Materials (CALM). The aim of CALM is "to collect and to archive (in a computerized data bank) a large body of primary observational evidence about language systems and language usage, in order to meet such research needs of humanistic disciplines as grammatical analysis, language universals, lexicology and lexicography, historical phonetics, language pedagogy, and child language development" (D. Sherman, "A Computer Archive of Language Materials", in S. Lusignan & J.S. North, eds., *Computing in the Humanities: Proceedings of the Third International Conference on Computing in the Humanities*, Waterloo, Ontario: The University of Waterloo Press, 1977, p. 283). CALM contains typological and lexicographic as well as textual data files. Access to the material will be provided by "printed reference works such as handbooks, indexes and concordances, and on-demand information retrieval services" (loc.cit.). Detailed information on CALM is contained in the article referred to and in: C.A. Ferguson, "Constructing a Textual and Lexicographic Archive for Language Research", Linguistics Department, Stanford University, 1977.

A FORTHCOMING BOOK ON CORPUS LINGUISTICS

The current interest in assembling corpora and establishing textual archives is reflected in a forthcoming book: H. Bergenholtz & B. Schaeder, eds., *Text-Corpora: Materialien für eine empirische Sprach- und Literaturwissenschaft*, Kronberg/ Ts.: Scriptor Verlag (to appear in 1978). Among the contributions could be mentioned:

S. Allén, "Lexical Morphology: A Method and an Application"

K.-H. Bausch, "Probleme der Intuition als empirische Basis in der Linguistik"

H. Bergenholtz & B. Schaeder, "Zur Methodik der Auswertung von Text-Corpora"

G. Engelien, "Die maschinelle Aufnahme von Texten für ein Corpus"

W. Nelson Francis, "Problems of Assembling and Computerizing Large Corpora"

R. Quirk & J. Svartvik, "A Corpus of Modern English"

The book will also contain a survey of available corpora of modern-language texts in machine-readable form.

CONDITIONS ON THE USE OF ICAME CORPUS MATERIAL

Users of ICAME material and researchers who have put material at the disposal of ICAME should note the following rules for the use of the material:

Material is distributed at cost and to bona fide researchers only.

The person/s/ who originally prepared the material in computerized form will be notified every time a new copy of the material in question is distributed through ICAME.

All recipients of ICAME corpus material will be required to sign a form where they agree to the following conditions:

1. No copies of corpora, or parts of corpora, are to be distributed under any circumstances without the written permission of ICAME.

2. Print-outs of corpora, or parts thereof, are to be used for bona fide research of a non-profit nature. Holders of copies of corpora may not reproduce any texts, or parts of texts, for any purpose other than scholarly research without getting written permission of the individual copyright holders, as listed in the manual or record sheet accompanying the corpus in question. (For material where there is no known copyright holder, the person/s/ who originally prepared the material in computerized form will be regarded as the copyright holder/s/.)

3. Commercial publishers and other non-academic organizations wishing to make use of part or all of a corpus or a print-out thereof must obtain permission from all the individual copyright holders involved.

4. The person/s/ who originally prepared the material in computerized form must be acknowledged in every subsequent use of it.

Researchers who have put material at the disposal of ICAME may add further conditions of use.

Editorial note

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