1. The International Project

Work on the compilation of corpora within the ICE (International Corpus of English) framework is in progress in a number of countries. Most of the research teams have received research grants from their own institutions and/or from a national grant-giving council.

An ICE Workshop, convened by Professor Jan Aarts (University of Nijmegen), was held on June 2-3 at a hotel near Nijmegen, the Netherlands, before the start of the annual ICAME (International Archive of Modern English) conference. Twenty participants from 12 locations attended, including researchers from as far away as Australia, India, New Zealand, South Africa, and the USA.

2. ICE Documents

Three ICE Newsletters were distributed by the Survey this past year, reporting on progress in the international project: Newsletter 12 (October 1991), 13 (March 1992), and 14 (June 1992).

The Survey continues to supply guidance to ICE teams, to ensure consistency in the compilation and processing of the corpora. Five ICE documents were sent out to ICE teams in November 1991:

Sidney Greenbaum. 'The Compilation of the International Corpus of English and its Components' (20pp)

Gerald Nelson. 'File Header Information' (22pp)

Gerald Nelson. 'Markup Manual for Written Texts' (25pp)

Gerald Nelson. 'Markup Manual for Spoken Texts' (17pp)

Akiva Quinn. 'ICE Technology - Software & Hardware for the International Corpus of English' (15pp)
A considerable amount of time was spent during the past year testing the ICE word tagset on texts. Several versions of the tagset manual were produced, with assistance from colleagues at the University of Nijmegen and the University of Louvain. The manual is in use at the Survey and will shortly be used at Louvain in the ICLE (International Corpus of Learner English) project directed by Professor Sylviane Granger. The latest version is in two parts:

Sidney Greenbaum, ‘The ICE Tagset Manual’ (101pp)

Ni Yibin, ‘Appendix to the ICE Tagset Manual: A List of Closed-Class Items and a Quick Reference to the Manual’ (50pp)

The manual will be sent to ICE teams when they are ready to start on the word-tagging stage.

Two software manuals were produced for use with the ICECUP system and the tag selection system (see 3 below):

Akiva Quinn and Nicholas Porter ‘The ICECUP User Manual’, version 0.5, April 1992 (30pp)


In addition, technical documentation is incorporated in six internal papers:


Akiva Quinn and Nicholas Porter, ‘String Description Language’, May 1992 (3 pp)


Akiva Quinn, ‘ICECUP - TACT Data Files’, August 1992 (3 pp)

3. **ICE Software**

TAGSELECT, the ICE Tag Selection System, has been developed by the Survey to automate selection from the alternative word class tags generated by the Nijmegen Tagger. All actions are selected from menus or by clicking on the appropriate button. Where the first tag is correct (the majority of cases), the user simply advances to the next word. If the first tag is not the correct one, the user highlights an alternative tag. Where the correct alternative is not shown, a new tag can be added from the list of possible tags. A sentence window provides context for tagging, tags can be queried, and both queries and word/tag combinations can be searched for in the text. The system keeps track of who is selecting or checking each text, only permitting access by this individual. Progress reports are available at any time. Version 1.2 of TAGSELECT can be obtained from the Survey, and requires at least a 286 PC with 2MB of RAM plus Microsoft Windows version 3.0 or 3.1.

ICECUP, the ICE Corpus Utility Program, is under development at the Survey to provide a corpus searching and analysis tool for use on the ICE corpora and beyond. An easy-to-use Windows graphical user interface will provide access to all ICECUP functions. Principal functions that have been completed include searching for String Description Language expressions, concordancing, subcorpus selection, markup display options, and frequency analysis. Utilities are provided to check the consistency of markup, number text units, prepare texts for tagging by Nijmegen, and selectively strip markup from a corpus. ICECUP version 1 is expected to be released before the end of 1992. Like TAGSELECT, it requires at least a 286 PC with 2MB of RAM plus Microsoft Windows version 3.0 or 3.1.

4. **ICE-GB**

The Survey is responsible for the compilation and processing of ICE-GB, the British corpus within the ICE framework. The million-word corpus is now complete in machine-readable form, provided with ICE markup. It comprises 500 texts (samplings of language), each with about 2000 words.
in text categories containing a minimum of 10 texts. The composition of
ICE-GB is approximately the same as that of the other ICE national
corpora. The categories are listed below, the numbers in parentheses being
the number of texts in each category.

**SPOKEN (300)**

**DIALOGUE (180)**

*Private (100)*
- direct conversations (90)
- distanced conversations (10)

*Public (80)*
- class lessons (20)
- broadcast discussions (20)
- broadcast interviews (10)
- parliamentary debates (10)
- legal cross-examination (10)
- business transactions (10)

**MONOLOGUE (120)**

*Unscripted (70)*
- spontaneous commentaries (20)
- unscripted speeches (30)
- demonstrations (10)
- legal presentations (10)

*Scripted (50)*
- broadcast news (20)
- broadcast talks (20)
- speeches (not broadcast) (10)
WRITTEN (200)

NON-PRINTED (50)

Non-professional writing (20):  
student untimed essays (10)  
student examination essays (10)

Correspondence (30):  
social letters (15)  
business letters (15)

PRINTED (150)

Informational (learned)  
humanities (10)  
social sciences (10)  
natural sciences (10)  
technology (10)

Informational (popular) (40):  
humanities (10)  
social sciences (10)  
natural sciences (10)  
technology (10)

Informational (reportage) (20):  
press news reports (20)

Instructional (20)  
administrative/regulatory  
skills/hobbies (10)

Persuasive (10)  
press editorials (10)

Creative (20)  
novels/stories (20)

ICE-GB is currently being annotatedgrammatically at the level of the word. Each word is assigned a word tag that indicates its word class and may also indicate one or more features that further characterize the word. The annotation is semi-automatic. A tagging program assigns one or more tags to each word, listed in order of probability. The human selectors at the
Survey check that the first tag is correct, and if not they choose another tag in the list or add a tag not on the list. The Survey's TAGSELECT program automates selection.

The grammatical annotation is performed in collaboration with the TOSCA Research Group at the University of Nijmegen, directed by Professor Jan Aarts. The TOSCA Group has developed the automatic tagging program and is applying it to ICE-GB using the ICE tagset. In the next stage, the automatic TOSCA parsing program will be applied to the tagged ICE-GB.

At the time of writing, 272 of the 500 texts in GB have been tag-selected and 172 of these texts have also been checked. Completion of the tagging annotation is expected by the end of 1992.

5. Funding
The support of the Economic and Social Research Council (ESRC) is gratefully acknowledged. Our work was funded in part by ESRC award R000-23-2077. We are also grateful for financial support from the Michael Marks Charitable Trust and the Sir Sigmund Sternberg Foundation.

6. Visitors
During the year we have been pleased to welcome the following scholars who have made use of our materials:

Professor L. Breivik
University of Bergen

cleft sentences

Mr Z. Eissiefy
Ain Shams University, Cairo

passives

Ms A. Fetzer
University of Stuttgart

pragmatics

Mr T. Fukaya
Sugiyama Jogakuen University, Japan

prepositions

Mr H. Higashi
Meijo University, Japan

modals
Professor Y. Ikegami  
Tokyo University

Ms C. Ilie  
Stockholm University

Professor M. Murata  
Chiba University, Japan

Mr T. Prcic  
University of Novi Sad, Yugoslavia

Professor I.M. Schlesinger  
Hebrew University of Jerusalem

Professor J. Svartvik  
University of Lund, Sweden

Mr A. Vogelmann  
University of Stuttgart

Ms K. Wales  
Royal Holloway & Bedford New College  
University of London

A number of scholars who are ICE participants came for discussion on
the ICE project:

Professor J. Aarts  
University of Nijmegen

Professor S. Granger  
University of Louvain, Belgium

Professor C. Mair  
University of Freiburg, Germany

Professor R. Morris  
University of Massachusetts-Boston

Dr A. Pakir  
National University of Singapore

Ms P. Peters  
Macquaries University, Australia

Dr J. Schmied  
University of Bayreuth, Germany

Professor S.V. Shastri  
Kolhapur University, India

John Bradley (Computing Services, University of Toronto) came for a week to
discuss possible collaboration between the Survey and the Centre for
Computing in the Humanities at the University of Toronto. The proposal, still under discussion, is for the development of software for text analysis.

Others who paid us brief visits were:

Dr K. Aijmer
Professor M. Akimoto
Professor R. Bailey
Professor J. Bately
Mr Simon Bell
Professor M. Benskin
Mr D. Campbell
Professor F. Cassidy
Dr M. Chayen
Mr D. Cooksey
Mr S. Crowdie
Professor A. Durant
Professor E. Finegan
Dr R.H. Flavell
Dr M. Fludernik
Mr E. Johnson
Dr D. Kalogjera
Mr G. Kaltenboeck
Professor Y. Nishimitsu
Mr Hamish Norbrook

University of Lund, Sweden
Aoyama Gakuin University, Japan
University of Michigan at Ann Arbor
King's College London
Routledge
University of Oslo
BBC
University of Wisconsin-Madison
Hebrew University of Jerusalem
BBC
Longman
Goldsmiths' College, University of London
University of Southern California
Institute of Education, University of London
University of Vienna
University of Cambridge
University of Zagreb, Yugoslavia
University of Vienna
University of Kobe, Japan
BBC
Professor H. Nyssonen  
Mrs A.G. Obermer  
Mr T. Onuma  
Dr N. Ostler  
Mr J. Price  
Mr Paul Proctor  
Mr K. Ricketts  
Mr L. Song  
Professor A.-B. Stenström  
Professor N. Takahashi  
Dr W. Teubert  
Professor J. Thiesmeyer  
Mr D. Tiomajou  
Dr L. Urbanova  
Dr E. User  
Dr T. Varadi  
Professor G. Veikhman  
Professor E. Yakovleva  
University of Oulu, Finland  
Michael Marks Charitable Trust  
Kenkyusha Press, Tokyo  
Department of Trade & Industry  
Routledge  
Cambridge University Press  
BBC  
Institute of Education, University of London  
University of Bergen, Norway  
Tokyo University of Foreign Studies  
University of Mannheim, Germany  
Hobart & William Smith College, USA  
University of Yaounde, Cameroon  
University of Prešov, Czechoslovakia  
Michael Marks Charitable Trust  
Hungarian Academy of Sciences, Budapest  
Moscow Linguistics University  
Moscow University

7. Staff

Several new members of staff joined us this year. Judith Broadbent (a PhD student in the UCL Department of Phonetics and Linguistics, who started work in July), Justin Buckley (a graduate from the UCL Department of English), Nicholas Porter (who has an MSc in Cognitive Science from the University of Sussex), and Ian Warner (a graduate from the UCL
Department of English, who left in July because of other commitments).

During the summer of 1992 David Elkan was employed to complete the sub-corpus selection module of ICECUP; he is now beginning an MSc course in Cognitive Science at the University of Edinburgh. During the same period three MSC students from the UCL Department of Computer Science - Riaz Hussan, Richard Wilson, and Tariq Zaidi - worked on aspects of ICECUP for their degree research projects.

Isaac Hallegua (recently retired from General Electric) joined us in July in a voluntary capacity, generously donating his time and expertise to the computational side of our work. He has been engaged on improving and documenting the back-up system, and is now in charge of backing-up our data.

Continuing staff from last year are Yanka Gavin, Gerald Nelson, Ni Yibin, René Quinault, Akiva Quinn, Oonagh Sayce, and Vladimir Žegarac.

Besides carrying on with his usual archival duties, René Quinault has embarked on the task of transferring the recordings of the original Survey corpus from reel to cassette tape to make them more readily accessible.

Akiva Quinn has been in charge of computing work at the Survey. He has developed TAGSELECT to version 1.2, and has been responsible for the design of ICECUP, the technical specification for ICECUP, and implementation of some of the modules, including those for indexing, processing 'markup', and extracting text-header information. Nicholas Porter has been working mainly on ICECUP: modifying the code, developing the retrieval module, providing a Windows Interface, and developing and implementing the String Description Language.

The other members of staff - Judith Broadbent, Justin Buckley, Yanka Gavin, Gerald Nelson, Ni Yibin, Oonagh Sayce, Ian Warner, and Vladimir Žegarac - worked on the language side. Except for Judith, they were engaged on transcribing and keypunching spoken material. They have all since worked on tag selection. Gerald Nelson has general responsibility for the tagging, including checking for possible textual errors and transmitting texts to the TOSCA Research Group in Nijmegen. Gerald and Yibin provide
the final check on tagging queries, and Yibin has responsibility for consistency of tag selections.

Gerald Nelson and Akiva Quinn gave presentations at the ICE Workshop in Nijmegen. Akiva also gave papers at the ALLC/ACH Conference at Oxford and the ICAME Conference in Nijmegen. Vladimir Žegarac gave a paper at the LAGB Conference in Brighton. Gerald taught at King's College London, Vladimir at the Central School of Speech and Drama, and Ni Yibin at Middlesex University. Vladimer was awarded the Ph.D. degree in Linguistics from the University of London.

Professor Greenbaum gave a lecture at the University of Nijmegen, chaired the ICE Workshop and gave presentations at it, and contributed a paper to the ICAME Conference in Nijmegen. He was interviewed about the ICE project on the BBC World Service. He has been elected to the Management Committee of the Society of Authors.

8. Publications


Sidney Greenbaum
Director, Survey of English Usage

September 1992