1. The International Corpus of English

Progress continues to be made by the many research teams working on the International Corpus of English. The annual ICAME (International Archive of Modern English) conference held in May in Switzerland was attended by sixteen ICE participants from thirteen locations, including researchers from as far away as Australia, Canada, Hong Kong, Singapore, and the USA. Six of the conference papers were related to the ICE project, in addition to a poster presentation and two software demonstrations. There was a special business meeting for ICE participants, which heard progress reports on all the ICE corpora.

2. ICE Documents

Three ICE Newsletters were distributed by the Survey this year.

The following manuals were produced on the grammatical annotation:

Gerald Nelson, ‘ICE Interactive Parsing Manual’ (4pp)
Gerald Nelson, ‘ICE Parse Selection Manual (3pp)
Ni Yibin, ‘The ICE Tagset’ (20pp)
Ni Yibin, ‘ICE Syntactic Marker Manual’ (21pp)
Ni Yibin, ‘ICE Parsing Manual’ (51pp)

In addition, Nelleke Oostdijk (TOSCA Group, University of Nijmegen) produced an ICE parser manual (35pp).

Three software manuals were produced:

Akiva Quinn & Nick Porter, ‘ICECUP 1.1 User Manual’ (30pp)
Akiva Quinn & Nick Porter, ‘ICECUP 2.0 User Manual’ (50pp)
Akiva Quinn, ‘ICEMARK User Manual’ (4pp)

In addition, technical documentation is incorporated in ten papers:

Isaac Hallegua, ‘Backup and restore procedures’ (3pp)
Isaac Hallegua, ‘Printers and print servers’ (2pp)
Nick Porter, ‘PC-NFS Licences and adding PCs to the network’ (3pp)
Nick Porter, ‘Adding a PC to the Backbone Network’ (1p)
Akiva Quinn, ‘Archiving parsed material’ (3pp)
Akiva Quinn, ‘Corpus-based teaching’ (2pp)
Akiva Quinn, ‘ICEMARK technical manual’ (2pp)
Akiva Quinn, ‘Indexing word-class tags, new indexing scheme’ (2p)
Akiva Quinn, ‘TAGSELECT technical manual’ (4pp)
Akiva Quinn, ‘Using FTP for transfer of files to/from timeshare’ (1p)
Akiva Quinn, ‘Using LEX and AWK under UNIX’ (4pp)

3. ICE Software
The Survey has developed a set of tools for annotating the ICE corpora:

The ICE Markup Assistant provides automation of, and simplified keypresses for, the
insertion of the standard set of markup symbols used throughout the ICE project.

TAGSELECT, the Ice Tag Selection system, automates selection from the alternative
word-class tags generated by the TOSCA tagger for each word in the corpus.

ICEMARK was developed this year. It is used to add syntactic markers to tagged
texts before the TOSCA parser does the parsing. One text unit is displayed at a time
and markers are added by selecting one from the list and pointing to the position
under the word where the marker should be added. Similarly, markers can be deleted
by pointing at the appropriate marker and clicking on the Delete Marker Button.
Other functions include changing a word-class tag, editing the sentence to correct
textual errors, and viewing the sentence with no tags.

ICECUP, the ICE Corpus Utility Program, has been under development for the last
three years. It is a tool for searching, concordancing, and analysing corpora. Version
2 was developed this year. New features include searching for markup; word partials,
tags and tag combinations, syntactic markers, and words with specified tags; file
output; display of subcorpus information; display of context for search query, in
accord with specified number of preceding and following text units; facility for viewing the whole of any text; context-sensitive help system.

4 Annotation of ICE-GB
During the past year we have been annotating grammatically ICE-GB, the million-word British corpus within the framework of the International Corpus of English. We have completed the tagging of each word in the corpus, with the assistance of the TOSCA tagger from the University of Nijmegen. We have also completed the syntactic marking; the computer-assisted manual insertion of syntactic markers, which are intended to assist the parsing by reducing ambiguities.

We are currently using the TOSCA parser to parse the corpus. The semi-automatic system of parsing consists of three stages. The first stage has been completed. It involves manual selection from the analyses offered by the parser. At the second stage, interactive parsing requires manipulation of the tags and markings to achieve a satisfactory analysis. The success rate at the first stage is about 50 per cent, and a further 25-30 per cent can be successfully parsed at the second stage. We have now covered about 70 texts in interactive parsing. At the final stage, parse trees will be drawn manually with computer assistance. We hope to complete the parsing by the end of 1993 or by early 1994.

5 The ICE tagset
The Survey has devised a tagset for the ICE project, which is based on the TOSCA tagset but differs from it substantially. It distinguishes 22 word classes:

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjective</td>
<td>existential there</td>
</tr>
<tr>
<td>adverb</td>
<td>formulaic expression</td>
</tr>
<tr>
<td>anticipatory it</td>
<td>genitive marker</td>
</tr>
<tr>
<td>article</td>
<td>interjection</td>
</tr>
<tr>
<td>auxiliary</td>
<td>nominal adjective</td>
</tr>
<tr>
<td>cleft it</td>
<td>noun</td>
</tr>
<tr>
<td>conjunction</td>
<td>numeral</td>
</tr>
<tr>
<td>connective</td>
<td>particle</td>
</tr>
<tr>
<td></td>
<td>preposition</td>
</tr>
<tr>
<td></td>
<td>proform</td>
</tr>
<tr>
<td></td>
<td>pronoun</td>
</tr>
<tr>
<td></td>
<td>prop it</td>
</tr>
<tr>
<td></td>
<td>reaction signal</td>
</tr>
<tr>
<td></td>
<td>verb</td>
</tr>
</tbody>
</table>

Some of these classes contain just one member or just a few members.
In addition, most of these word classes have one or more features attached to them. For example, verbs are marked for transitivity and verb form, and (where applicable) they are marked as enclitic or negative. In all, the ICE tagset contains 256 possible combinations of word class and features. These can be retrieved from the corpora through ICECUP.

6 The ICE syntactic marker set
To facilitate the parsing, a number of syntactic markers - adapted from the TOSCA marker set - have been established for the ICE corpora. The ICE marker set contains the following markers:

- beginning and end of a conjoin (unit in coordination)
- end of postmodifier of a noun phrase
- beginning and end of an adverb phrase functioning as premodifier of a noun phrase
- beginning and end of the extrapoosed part of a noun phrase
- beginning and end of a noun phrase with untypical functions (premodifier or postmodifier of an adverb, adverbal, postmodifier of a noun phrase, premodifier of a prepositional phrase, premodifier of a finite clause)
- beginning and end of an unanalysed noun phrase (used when the internal structure is not to be analysed in parsing because of ambiguity or uncertainty)
- beginning and end of a parenthetic clause
- beginning and end of a noun phrase as a vocative
- beginning of a sentence within a text unit

These can be retrieved from the corpora through ICECUP.

7. Annotation of the Survey Corpus
The original Survey Corpus has been tagged automatically with the ICE tags. The tagging has not been manually checked, but based on a sample that was manually checked the tagging program achieves an accuracy of 96-97 per cent for written texts and 93 per cent for spoken texts. We hope to improve the accuracy for the spoken texts on a later re-tagging. The Survey Corpus has been indexed for ICECUP, so that it will be possible to search for all the features provided in ICECUP, including subcorpus selection. We hope in due course to parse the Survey Corpus automatically with a program that will provide analyses that are less detailed than those offered by the Nijmegen parser used for ICE but that will be compatible
with the ICE parsing. It will then be possible to make diachronic comparisons within the Survey Corpus (which covers a period of about 30 years) and also between parts of the Survey Corpus and corresponding parts in the ICE Corpus.

At present the Survey Corpus can be consulted only at the Survey premises, but we expect to release the grammatically annotated corpus within the next year, possibly first in a tagged version and then in a version with both tagging and parsing. The 100 spoken texts of the Survey Corpus are already available with prosodic annotation as the London-Lund Corpus, which can be obtained from the Norwegian Computing Centre for Humanities at the University of Bergen.

8. 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, the Sir Sig mund Sternberg Foundation, and The Longman Pearson PLC. The Leverhulme Trust has awarded Professor Greenbaum a research grant of £100,720 over two years from 1st October 1993 for research into grammatical relationships between clauses in spoken English from the ICE-GB corpus.

9. Publicity
During the past year there were two broadcast interviews on the ICE project on the BBC World Service, one with Professor Granger and Professor Greenbaum and the other with Professor Greenbaum alone. Articles on the project appeared in the Guardian newspaper, Language International, and the EFL Gazette. An article on the project was distributed internationally by Gemini News Service.

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

Professor J. Algeo
University of Georgia, U.S.A. American and British grammar

Ms I. Bernard
University Aix Marseille I, France say and tell
Mrs I. Constaninescu
Institutul de Lingvistica
Bucharest, Romania

Dr I. Depraetere
Kulak University, Belgium
tense and aspect

Dr. H. Günther
Jena University, Germany
phraseological units

Ms H. Hasselgård
University of Oslo, Norway
time and space adverbials

Ms E. Hoffman
University of Aix-en-Provence I
France
as and so

Professor M. Ljung
University of Stockholm, Sweden
newspaper English

Professor A. Mittwoch
Hebrew University of Jerusalem
Israel
aspect

Ms L. Opdahl
University of Bergen, Norway
adverbs

Ms G. Papadopoulou
Department of Phonetics & Linguistics
University College London
pragmatic connectives

Ms P. Typadi
Department of Phonetics and Linguistics
University College London
parentheticals

Ms J. Van Dyke
Newport News
VA, U.S.A.
semantics

Mrs Katie Wales
Royal Holloway and Bedford
New College, UK
personal pronouns

Ms B. Wörner
Stuttgart University, Germany
articles
A number of scholars who are ICE participants came for discussion on the ICE project:

Professor J. Aarts  
University of Nijmegen, The Netherlands

Mr P. Bolt  
Hong Kong Polytechnic

Dr K. Shields-Brodber  
University of the West Indies, Jamaica

Professor S. Granger  
University of Louvain, Belgium

Dr J. Kirk  
Queen’s University, Northern Ireland

Professor C.F. Meyer  
University of Massachusetts-Boston, USA

Professor J. Schmied  
University of Bayreuth, Germany

Other visitors were:

Ms J. Ayling  
Cambridge University Press, UK

Professor Janet Bately  
King’s College London, UK

Dr R. Baumgardner  
The Asian Foundation, Pakistan

Professor J. Campbell  
Department of Computer Science, UCL

Professor W.Q. Chen  
Tianjin Medical College, China

Mr S. Crowdie  
Longman, UK

Mrs D. Dolores  
The Language Centre, UCL

Dr B. Erman  
Stockholm University, Sweden

Mr A. Fang  
Guangzhou Foreign Language Institute, China

Professor Y. Fu  
The State Language Commission, China

Mr T. Fukaya  
Sugiyama Jogakuen University, Japan

Professor N. Givishiani  
Moscow State University, Russia

Professor M. Görlach  
University of Cologne, Germany

Mr P. Hanks  
Oxford University Press, UK
Mr J. Hughes
Dr G. James
Dr E. Johnson
Mr D. Jowitt
Professor Y. Kachru
Mr A.G. Kingscott
Professor I. Lancashire
Mr Dr Lewis
Miss C.C. Luz
Mrs F. McArthur
Dr T. McArthur
Dr A. Le Meur
Professor D. Mindt
Ms F. Morphy
Mr H. Norbrook
Dr N. Ostler
Miss B. Picot
Mr. A. Rosenheim
Professor R. Rosner
Mr D. Rowan
Mr P. Schneider
Mr R. Scriven
Professor A.-B. Stenström
Mr D. Tiomaju

EFL Gazette, UK
Hong Kong University of Science & Technology
Cambridge University, UK
University of Bayero, Kano, Nijeria
University of Illinois at Champagne-Urbana, USA
Language International, UK
University of Toronto, Canada
London, UK
University of Alicante, Spain
Cambridge, UK
English Today, UK
Université de Haute Bretagne, France
Freie Universitat Berlin, Germany
Oxford University Press, UK
BBC English, UK
Department of Trade and Industry, UK
The Language Centre, UCL
Oxford University Press, UK
Computer Centre, UCL
The Guardian newspaper, UK
University of Zurich, Switzerland
Oxford University Press, UK
University of Bergen, Norway
University of Yaounde, Cameroon
11. Staff

And Rosta re-joined us this year. Continuing staff from last year are Judith Broadbent, Justin Buckley, Yanka Gavin, Gerry Nelson, Nick Porter, Akiva Quinn, Oonagh Sayce, Ni Yibin, and Vlad Žegarac. Voluntary participants are Isaac Hallegua, Neil Morgenstern, René Quinault, Richard Wilson, and Tariq (‘Zeb’) Zaidi. Richard and Zeb had worked on ICECUP for their MSc research projects, for which both were awarded a distinction.

Akiva has been in charge of computing work at the Survey, and has continued his role in the design and development of ICECUP. In addition, he has produced ICEMARK. Nick has taken an active role in the design and development of ICECUP. Isaac Hallegua has been engaged on system and data management. Richard, Zeb and Neil have worked on various aspects of ICECUP.

Two MSc students from the UCL Department of Computer Science worked on ICECUP for their degree research project: Christine Papadakis on English word morphology and Dina Quaraishi on the indexing and playback of digitized speech.

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

The other members of staff have worked on the language side. Gerry Nelson has had general responsibility on this side of the project.

At the ICAME conference in Switzerland, Gerry Nelson and Akiva Quinn gave a
poster presentation and Akiva also gave a demonstration of ICECUP. Gerry spent two weeks in Nijmegen learning the parsing system.

Justin Buckley published (in collaboration with two other musicians, as Osmium), a CD and vinyl record of rock music.

Professor Greenbaum gave a talk in a symposium at the University of Bergen. He contributed a paper to the ICAME conference and chaired a session. He has joined the Editorial Advisory Board of Language Sciences (Pergamon Press).

12. Publications based on Survey material


Sidney Greenbaum
Director, Survey of English Usage

September 1993