UNIVERSITY COLLEGE LONDON

Survey of English Usage

ANNUAL REPORT 1993-94

1. The International Corpus of English

The New Zealand Corpus has now been compiled, and the Singapore Corpus is very close to completion. We can therefore expect there to be three complete components of ICE, the International Corpus of English, by the end of 1994, the third being ICE-GB, the British Corpus compiled at the Survey.

There are now commitments for 15 components (national or regional corpora) of ICE:

1. Australia
2. Canada
3. Caribbean
4. Cameroon
5. East Africa:
   - Kenya
   - Tanzania
   - Zambia
6. Great Britain
7. Ireland
   - Eire
   - Northern Ireland
8. Hong Kong
9. India
10. New Zealand
11. Nigeria
12. Philippines
13. Singapore
14. South Africa
15. U.S.A.

In addition, academics in Ghana and Sierra Leone have expressed interest in joining the ICE project.

Associated with the ICE project is ICLE (International Corpus of Learner English) directed by Professor Sylviane Granger at the University of Louvain-La-Neuve, Belgium. The ICLE project draws on English essays by university students from nine language backgrounds: French, Spanish, German, Dutch, Swedish, Czech, Finnish, Japanese, Chinese. An automatic tagging system and a semi-automatic parsing system have been

contributed by the TOSCA Research Group at the University of Nijmegen, the Netherlands. The Survey has been supplying guidance and software on the compilation and markup of the ICE corpora, and is now developing automatic tagging and parsing programs.

The annual ICAME (International Computer Archive of Modern English) conference held in May in Denmark was attended by twelve ICE participants. Two of the conference papers were related to the ICE project. There was a special business meeting for ICE participants, which heard progress reports on all the ICE corpora.

2. The Leverhulme Project

The Leverhulme project is primarily an investigation into clause relationships. The data are drawn from a subset of ICE-GB that we have termed the Leverhulme Corpus. The Leverhulme Corpus comprises 42 texts, about 2,000 words each, distributed between several varieties of spoken and written English:

SPEECH (30)

- spontaneous conversations (20)
- unscripted monologues (5)
- broadcast discussions (5)

WRITING (12)

- personal letters (4)
- academic informative writing (4)
- non-academic informative writing (4)

The clause relationships have been annotated manually. Professor Greenbaum and Dr Nelson have written two papers on the Leverhulme Corpus and two further papers are in progress. Results from the research have also been incorporated in a chapter on sentences and clauses in The Oxford English Grammar, by Sidney Greenbaum, due to be published by Oxford University Press in 1995.

3. ICE Documents

Two ICE Newsletters were distributed by the Survey this year.

The following manuals were produced:

Judith Broadbent and Nick Porter, 'ICECUP User Manual for ICECUP version 2' (54pp)

Nick Porter, 'Manual for text processing for ICE', for preparing ICE texts for tagging (3pp)

Alex Fang, 'Manual for AUTASYS' (12pp)

Alex Fang, 'Manual for TQUERY' (9pp)

Ni Yibin, 'The Menu for the ICE Parsing Tree-Editor' (58pp)

Technical documentation is incorporated in five papers by Isaac Hallegua:

'Timed backup procedures' (2pp)

'Backup disk file data' (8pp)

'Retrieve data in dd format from tape' (2pp)

'Clearing up after the parser' (2pp)

'Do's and Don'ts before doing a Dump or Restore' (2pp)

4. ICE Software

ICECUP 2.0 has been modified and tested and is now ready for distribution. The new features incorporated in ICECUP 2.0 are (1) the subcorpus facility allows for biographical information to be included about listeners as well as speakers; (2) the user interface is much improved; (3) citations can be marked for outputting to file; (4) multiple queries can be specified, e.g. searching for all verbs in present tense irrespective of their transitivity or for all intransitive verbs irrespective of their other features; ICECUP has been converted to a WIN32s program, enabling it to run on Windows 3.1 (with WIN32s), Windows NT, and Windows 95 (the latest version, expected early next year). ICECUP (ICE Corpus Utility Program), a tool for searching and concordancing corpora, was primarily developed by Nick Porter and Akiva Quinn, but contributions also came from research projects undertaken by students from the Department of Computer Science at UCL. Neil Morgenstern is chiefly responsible for the subcorpus facility.

ICEParse 1.2, by Nick Porter, generates statistical information on parsing progress and control.

ICETreee 1.0, by Nick Porter and Dennis Tech-Yong, is a Windows-based application for graphically building or editing parse trees.

Prepare1.1, by Nick Porter, converts texts to ICE format in preparation for ICE word-class

tagging and use with ICECUP.

HyperGram, by Nick Porter, is a prototype Hypertext grammar system.

AUTASYS, by Alex Fang, is an automatic tagger for tagging words with ICE tags.

TQUERY, by Alex Fang, is a system for retrieving syntactic informaiton (word-class tags and parses).

5. Annotation of ICE-GB

We have been using the TOSCA parser to parse ICE-GB, the million-word British ICE corpus. The TOSCA parser has been developed by the TOSCA Research Group under the direction of Professor Jan Aarts at the University of Nijmegen. During the past year the TOSCA parser has been produced in several versions, each improving on the previous version, and we have been applying these to increase the success rate of the parsing. We estimate that about 70 per cent of the ICE-GB has been parsed. We are now about to apply to the remaining 30 per cent of the corpus a parser being developed at the Survey by Alex Fang. Whatever remains after that application will be manually parsed with the aid of ICETree during the academic year 1994-95.

6. Funding

The major financial contribution for 1993-94 has come from the Leverhulme Trust. We are also grateful for financial support from the Michael Marks Charitable Trust and the Sir Sigmund Sternberg Foundation.

7. Visitors

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

Professor John Algeo University of Georgia, U.S.A.	American and British grammar
Mr Francisco Gonzalvez Garcia University of Granada, Spain	object-plus-infinitive constructions; ICE systems
Dr Joseph Hung Chinese University of Hong Kong	modal auxiliaries; conversational analysis, ICE systems

Professor Takuro Ikeda Aoyama Gakuin University, Tokyo, Japan	idioms
Mr Gunther Kaltenboeck University of Vienna, Austria	anticipatory it
Professor Magnus Ljung University of Stockholm, Sweden	nonfinite adverbial clauses
Professor Izchak Schlesinger Hebrew University of Jerusalem, Israel	semantic roles

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

Professor Jan Aarts	University of Nijmegen, The Netherlands
Nancy Belmore	Concordia University, Montreal, Canada
Professor Sylviane Granger	University of Louvain, Belgium
Professor Graeme Kennedy	Victoria University, New Zealand
Dr John Kirk	Queen's University, Belfast, Northern Ireland
Dr Anne Pakir	National University of Singapore
Professor Josef Schmied	Technische Universitat, Chemnitz, Germany
Pam Peters	Macquarie University, Sidney, Australia

Other visitors were:

Mr J. Amey	Department of Trade and Industry, UK
Professor. P. Anusas	Vilnius University, Lithuania
Dr E. Atwell	University of Leeds, U.K.
Professor W.-.D. Bald	University of Cologne, Germany
Dr A.S. Bobda	University of Yaounde, Cameroon
Dr S. Cmerjkova	Inst. of Czech Languages, Prague, The Czech Republic
Mr M. Cutts	Plain Language Commission, UK

Professor R. de Beaugrande	University of Vienna, Austria
Professor J. Firbas	Brno University, The Czech Republic
Professor M. Fludernik	University of Freiburg, Germany
Ms. S.Y.C. Fung	Law Drafting Division, Hong Kong
Mr P. Gibbins	Sharp, UK
Mr G. Hill	London, UK
Mr J. Hughes	University of Leeds, UK
Mr I. Johnson	Sharp, UK
Mr R. Kilgariff	BBC
Mr M. Le Fanu	Society of Authors, UK
Mr T. McArthur	English Today
Lord and Lady Marks of Broughton	Michael Marks Charitable Trust, UK
Mr J. Milton	Hong Kong University of Science & Technology
Mr S. Murisson-Bowie	Oxford University Press
Professor Y. Murata	Rikyo University, Japan
Mr H. Norbrook	BBC
Mr R. Scriven	Oxford University Press
Mr A.N. Watson-Brown	Law Drafting Division, Hong Kong
Professor I. Yasui	University of Tsukuba, Japan

8. Staff

Several members of staff left us this year, with our good wishes. Akiva Quinn emigrated to Australia, where he has been working for a computing company; he has recently

married. And Rosta was appointed a lecturer at the Roehampton Institute of Higher Education. Vlad Zegarač joined a research project at the University of Middlesex and teaches part-time at the University of Sussex. Yanka Gavin completed an MA in Anglo-American Cultural Relations at UCL and joined a publishing company.

Continuing staff from last year are Judith Broadbent, Justin Buckley, Gerry Nelson, Nick Porter, Oonagh Sayce, and Ni Yibin. Continuing voluntary participants are Isaac Hallegua and René Quinault. We welcome a new member of staff: Alex Fang, formerly at the Guangzhou Institute of Foreign Languages.

Nick Porter is now in charge of computing work at the Survey. He has completed ICECUP version 2.0, ICEParse version 1.2, and Prepare version 1.0, and he has been working on ICETree version 1.0 in collaboration with Dennis Tech-Yong. Dennis and Kate Millard were MSc students from the Department of Computer Science who undertook MSc research projects for ICE during the summer. Dennis, who gained a distinction for his project, has continued to contribute to ICEtree voluntarily while engaged on a full-time computing job for the police. Kate's project was a module for displaying overlapping speech. Domenico Pirlo, an undergraduate student in Computer Science, is developing a morphological analysis module for ICECUP. Neil Morgenstern, who is now employed in the Department of Computer Science, has developed the subcorpus facility for ICECUP and has been working on a program to align the tagged corpus and the parsed corpus for ICECUP. Isaac Hallegua has continued his activities on system and data management.

In addition to his usual archive duties, René Quinault is working on reducing the ICE recordings to those used in the ICE corpus, having finished analogous work on the original Survey corpus.

The other members of staff have worked on the language side. Gerry has been engaged chiefly on the Leverhulme project, but has continued to provide guidance to ICE teams internationally. Contributions to the Leverhulme project have also come from Justin and Oonagh. Judith, Justin, Oonagh, and Yibin have been engaged on interactive parsing of the ICE corpus using the TOSCA parser. Justin is currently working on the Help system for the tree-builder/editor, Judith on the ICECUP Help system, and Yibin on the ICE parsing manual.

Alex has been working on tagging and parsing programs. He developed AUTASYS for use with the ICE tagset, which involved cross-tagset mapping of LOB to ICE, and created TQUERY for retrieval of tags and parses. He tagged the Survey Corpus (one million words) and the 1988 and 1989 issues of the Wall Street Journal (twenty-two million words) with the ICE tagset. He is currently developing an automatic parsing program.

Professor Greenbaum was the 1993/94 Distinguished Visiting Scholar at United College, the Chinese University of Hong Kong, where he delivered three public lectures on English Studies in November 1993. While in Hong Kong, he gave a talk at a seminar for teachers organized by Longman Hong Kong. He chaired a session at the ICAME conference in Denmark in May 1994 and an ICE business meeting.

A joint paper was presented at the ICAME conference by Professor Greenbaum and Gerry Nelson and another paper by Nick Porter. Judith Broadbent presented a paper at the summer meeting of the LAGB (University of Middlesex) and at the International Workshop on Phonological Structure (University of Durham). Alex Fang presented a paper at the International Workshop on Directions of Lexical Research in Beijing and at a seminar in the UCL Department of Phonetics and Linguistics.

Apart from papers based on Survey material, several articles were published by members of staff: 'War and the OED', Verbatim XX (1994) 17-19, by Gerry Nelson; 'Punning Rebuses in the Chinese Decorative Arts' by Ni Yibin, Daily Telegraphy 25 July 1994, p.16; 'Notes from London' (translation) by Ni Yibin, New Chinese Writing from London, eds. J. Chang, L. Pan, and H. Zhao, pp 79-84, London: Lambeth Chinese Community Association 1994.

9. Publications based on Survey material

Aarts, B. (1994) 'The syntax of binominal Noun Phrases in English', Dutch Working Papers in English Language and Linguistics 30, 1-28.

Aarts, F. (1993) 'Who, whom, that and ∅ in two corpora of spoken English', English Today 9, 19-21.

Altenberg, B. (1994) 'On the functions of such in spoken and written English', Corpus-based Research into Language: In Honour of Jan Aarts, eds. N. Oostidjk and P. de Haan, 223-240. Amsterdam: Rodopi.

Bäcklund, I. (1992) 'Macrostructure in conversation', Nordic Research on Text and Discourse: NORDTEXT Symposium 1990, eds. A.-C. Lindeberg, N.E. Enkvist, and K. Wikberg, 61-71. Åbo: Åbo Academy Press.

Biber, D. and E. Finegan (1994) 'Intra-textual variation within medical research articles', Corpus-based Research into Language: In Honour of Jan Aarts, eds. N. Oostdijk and P. de Haan, 201-221. Amsterdam: Rodopi.

De Rycker, T. (1993) 'A corpus-based analysis of "elliptical" imperatives in conversational discourse', Linguistica Antverpiensia 27, 109-130.

Eeg-Olofsson, M. and B. Altenberg (1994) 'Discontinuous recurrent word combinations in

the London-Lund Corpus', Creating and Using English Language Corpora, eds. U. Fries, G. Tottie, and P. Schneider, 63-77. Amsterdam: Rodopi.

Fang, A.C. (1994) 'ICE: Applications and possibilities in NLP', Proceedings of the Post-COLING94 International Workshop on Directions of Lexical Research, eds. N. Calzolari and C. Guo, 23-46. Beijing: Tsinghua University.

Fang, A.C. and G. Nelson (1994) 'Tagging the Survey Corpus: a LOB to ICE experiment using AUTASYS', Literary and Linguistic Computing 9, 189-194.

Geluykens, R. (1991) 'Information flow in English conversation: A new approach to the given-new distinction', Functional and Systemic Linguistics: Approaches and Uses, ed. E. Ventola, 141-167. Berlin: Mouton de Gruyter.

Greenbaum, S. and Y. Ni (1994) 'Tagging the British ICE Corpus: English Word Classes', Corpus-based Research into Language: In Honour of Jan Aarts, eds. N. Oostdijk and P. de Haan, 33-45. Amsterdam: Rodopi.

Greenbaum, S. and R. Quirk (1994) A Student's Grammar of the English Language (Korean translation). Seoul: Hansin.

Quinn, A. and D. Quinn (1993) 'CORTEX: A corpus-based teaching expert', AI '93: Proceedings of the 6th Australian Joint Conference on Artificial Intelligence, Melbourne, Australia, 16-19 November 1993, eds. C. Rowles, H. Liu, and N. Foo, 377-382. Singapore: World Scientific.

Quinn, D. and A. Quinn (1994) 'Linguistic Modelling for a corpus-based CALL system', Corpora in Language Education and Research: A Selection of Papers from Talc 94, eds. W. Wilson and T. McEnery, 87-98. Lancaster: Unit for Computer Research of the English Language, Lancaster University.

Quinn, A. and N. Porter (1994) 'Investigating English usage with ICECUP', English

Today 10, 19-24.

Schmied, J. (1994) 'Analysing style variation in the East African Corpus of English', Creating and Using English Language Corpora, eds. U. Fries, G. Tottie, and P. Schneider, 169-174. Amsterdam: Rodopi.

Shimizu, M.(1990) 'A DRS approach to reflexives', The Bulletin of the Kyushu Institute of Technology 38, 35-57.

Stenström, A.-B. (1994) An Introduction to Spoken Interaction. London: Longman.

Stenström, A.-B. and J. Svartvik (1993) 'Imparsable speech: Repeats and other nonfluencies in spoken English', Corpus-based Research into Language: In Honour of Jan Aarts, eds. N. Oostdijk and P. de Haan, 241-254. Amsterdam: Rodopi.

Sidney Greenbaum
Director, Survey of English Usage

November 1994