UNIVERSITY COLLEGE LONDON

Survey of English Usage

ANNUAL REPORT 1992-93

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)

Ni Yibin, 'Appendix to the ICE Tagset Manual', 2nd edition (80pp)

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:

adjective
adverb
anticipatory it
article
auxiliary
cleft it
conjunction
connective

existential there
formulaic expression
genitive marker
interjection
nominal adjective
noun
numeral

particle
preposition
proform
pronoun
prop it
reaction signal
verb

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 extraposed part of a noun phrase
beginning and end of a noun phrase with untypical functions (premodifier or postmodifier of an adverb, adverbial, 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 withіna 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 Sigmund 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	collocations
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	EFL Gazette, UK
Dr G. James	Hong Kong University of Science & Technology
Dr E. Johnson	Cambridge University, UK
Mr D. Jowitt	University of Bayero, Kano, Nijeria
Professor Y. Kachru	University of Illinois at Champagne-Urbana, USA
Mr A.G. Kingscott	Language International, UK
Professor I. Lancashire	University of Toronto, Canada
Mr Dr Lewis	London, UK
Miss C.C. Luz	University of Alicante, Spain
Mrs F. McArthur	Cambridge, UK
Dr T. McArthur	English Today, UK
Dr A. Le Meur	Université de Haute Bretagne, France
Professor D. Mindt	Freie Universitat Berlin, Germany
Ms F. Morphy	Oxford University Press, UK
Mr H. Norbrook	BBC English, UK
Dr N. Ostler	Department of Trade and Industry, UK
Miss B. Picot	The Language Centre, UCL
Mr. A. Rosenheim	Oxford University Press, UK
Professor R. Rosner	Computer Centre, UCL
Mr D. Rowan	The Guardian newspaper, UK
Mr P. Schneider	University of Zurich, Switzerland
Mr R. Scriven	Oxford University Press, UK
Proftessor A.-B. Stenström	University of Bergen, Norway
Mr D. Tiomaju	University of Yaounde, Cameroon

Professor G. Tottie	University of Zurich, Switzerland
Ms L. Updahl	University of Bergen, Norway
Professor S. Wilbur	Department of Computer Science, UCL

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

Aijmer, K. (1990) 'Teaching spoken English', Proceedings from the Fourth Nordic Conference for English Studies, eds. G. Caie et al, 383-395. Copenhagen: Department of English, University of Copenhagen.

Altenberg, B. (1990) 'Speech as linear composition', Proceedings from the Fourth Nordic Conference for English Studies, eds. G. Caie et al, 133-145. Copenhagen: Department of English, University of Copenhagen.

Altenberg, B. (1993) 'Recurrent verb-complement constructions in the London-Lund Corpus', English Language Corpora, eds. J. Aarts et al, 227-245. Amsterdam: Rodopi.

Biber, D. (1990) 'Methodological issues regarding corpus-based analyses of linguistic variation', Literary and Linguistic Computing 5, 257-269.

Biber, D. (1992) 'Using computer-based text corpora to analyze the referential strategies of spoken and written texts' Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991, eds. J. Svartvik, 215-352. Berlin: Mouton de Gruyter.

Biber, D. (1992) 'On the complexity of discourse complexity: A multidimensional analysis', Discourse Processes 15, 133-163.

Biber, D. and E. Finegan (1992) 'The linguistic evolution of five written and speech-based English genres from the 17th to the 20th centuries', Enter the Helsinki Corpus: Computer-assisted approaches to English historical linguistics, ed. M. Rissanen et al, 688-704. Berlin: Mouton de Gruyter.

Clark, C.H. and R.J. Gerrig (1990) 'Quotations as demonstrations', Language 66, 764-805.

Collins, P. (1991) Cleft and pseudo-cleft constructions in English. London: Routledge.

Depraetere, I. (1992) 'Aspects of expressing anteriority in past domain relative clauses', Belgium Essays in Language and Literature, eds. P. Michel et al, 27-35. Liège: Liège Language and Literature.

Edwards, J.A. 'Design principles in the transcription of spoken discourse', Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991, ed. J. Svartvik, 129-144. Berlin: Mouton de Gruyter.

Erman, B. (1990) 'From lexical to pragmatic meaning: Three levels of functions of pragmatic expressions in relation to their lexical meaning', Proceedings from the Fourth Nordic Conference for English Studies, eds. G. Caie et al, 303-317. Copenhagen: Department of English, University of Copenhagen.

Erman, B. (1990) You know in face-to-face and telephone conversation', English today: Papers read at the English Studies Conference in Umeå, June 2-3, 1988, eds. I Henrysson and G. Persson, 14-21. Umeå: Department of English, University of Umeå.

Fludernik, M. (1993) The Fictions of Language and the Languages of Fiction. London: Routledge.

Fukaya, T. (1993) 'Corpus Linguistics in English Language Research', Journal of Sugiyama Jogakuen University 24, 139-150.

Fukaya, T. (1993) 'Motivations for Deferred Prepositions', Sophia Linguistica 33, 101-129.

Geisler, C. (1992) 'Relative infinitives in spoken and written English', New directions in English Language corpora: Methodology, results, software developments, ed. G. Leitner, 213-230. Berlin: Mouton de Gruyter.

Geluykens, R. (1992) From discourse process to grammatical construction: On left-dislocation in English. Amsterdam: John Benjamins.

Greenbaum, S. (1993) 'The tagset for the International Corpus of English', Corpus-based computational linguistics, eds. C. Souter and E. Atwell, 11-24. Amsterdam: Rodopi.

Hudson, J. (1990) 'A computerized study of multi-word fixed-phrase adverbials', Proceedings from the Fourth Nordic Conference for English Studies, eds. G. Caie et al, 335-341. Copenhagen: Department of English, University of Copenhagen.

Johansson, C. (1993) 'Whose and of which with nonpersonal antecedents in written and spoken English', Corpus-based computation linguistics, eds. C. Souter and E. Atwell, 97-116, Amsterdam: Rodopi.

Knowles, G. (1990) 'The use of spoken and written corpora in the teaching of language and linguistics', Literary and Linguistic Computing 5, 45-48.

Lavelle, T. (1990) 'Rules, tendencies and predictions: English nominalizations in S-V-C clauses', English today: Papers read at the English Studies Conference in Umeå, June 2-3, 1988, eds. I. Henrysson and G. Persson, 40-48. Umeå: Department of English, University of Umeå.

Leitner, G. (1992) 'International Corpus of English: Corpus design - problems and suggested solutions', New directions in English language corpora: Methodology, results, software developments, ed. G. Leitner, 33-64. Berlin: Mouton de Gruyter.

Ljung, M. (1990) A Study of TEFL Vocabulary. Stockholm: Almqvist and Wiksell.

Nevalainen, T. (1992) 'Intonation and discourse type', Text 12, 397-427.

Opdahl, L. (1991) 'Close or closely as verb modifier? In search of explanatory parameters', Proceedings from the Fourth Nordic Conference for English Studies, eds. G. Caie et al, 201-212. Copenhagen: Department of English.

Papp, N. (1989) 'Future time and future tense in English and Hungarian', Annales Universitatis Scientiarum Budapestenensis 20, 5-32.

Powell, M. J. (1992) 'Semantic/pragmatic regularities in informally marked lexis: British speakers in spontaneous conversational settings', The Eighteenth LACUS Forum 1991, ed. R.M. Brend, 379-391. Lake Bluff, Illinois: LACUS.

Quinn, A. (1993) ‘An object-oriented design for a Corpus Utility Program’, English Language Corpora, eds. J. Aarts et al., 215-225. Amsterdam: Rodopi.

Tesch, F. (1990) Die Indefinitpronomina ‘some’ und ‘any’ im autentischen englischen Sprachgebrauch und in Lehrwerken. Tubingen: Gunter Narr.

Tottie, G. (1991) ‘Lexical diffusion in syntactic change: Frequency as a determinant of linguistic conservatism in the development of negation in English’, Historical English syntax, ed. D. Kastovsky, 439-467. Berlin; Mouton de Gruyter.

Virtanen, T. (1992) ‘Temporal adverbials in text structuring: On temporal text strategy’, Nordic research on text and discourse: NORDTEXT Symposium 1990, eds. A.C. Lindeberg et al, 185-197. Åbo: Åbo Academy Press.

Sidney Greenbaum
Director, Survey of English Usage

September 1993