UNIVERSITY COLLEGE LONDON

Survey of English Usage

ANNUAL REPORT 1991-92

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)

Akiva Quinn, 'The ICE Tag Selection System: User Manual', August 1992 (5 pp)

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

Akiva Quinn, 'A Command Language for the ICE Corpus Utility', October 1991 (3 pp)

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

Akiva Quinn and Nicholas Porter, 'The ICECUP Technical Manual', June 1992 (23 pp)

Akiva Quinn, 'The ICECUP Text and Reference Indices', August 1992 (5 pp)

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

Nicholas Porter, 'ICECUP Retrieve Module', August 1992 (6 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 annotated grammatically 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	causation
Ms C. Ilie Stockholm University	pragmatics and rhetorical questions
Professor M. Murata Chiba University, Japan	collocations
Mr T. Prcic University of Novi Sad, Yugoslavia	agentive suffixes
Professor I.M. Schlesinger Hebrew University of Jerusalem	cognitive and semantic categories
Professor J. Svartvik University of Lund, Sweden	revision of Communicative Grammar
Mr A. Vogelmann University of Stuttgart	progressive forms in speech
Ms K. Wales Royal Holloway & Bedford New College University of London	personal pronouns

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	University of Lund, Sweden
Professor M. Akimoto	Aoyama Gakuin University, Japan
Professor R. Bailey	University of Michigan at Ann Arbor
Professor J. Bately	King's College London
Mr Simon Bell	Routledge
Professor M. Benskin	University of Oslo
Mr D. Campbell	BBC
Professor F. Cassidy	University of Wisconsin-Madison
Dr M. Chayen	Hebrew University of Jerusalem
Mr D. Cooksey	BBC
Mr S. Crowdie	Longman
Professor A. Durant	Goldsmiths' College, University of London
Professor E. Finegan	University of Southern California
Dr R.H. Flavell	Institute of Education, University of London
Dr M. Fludernik	University of Vienna
Mr E. Johnson	University of Cambridge
Dr D. Kalogjera	University of Zagreb, Yugoslavia
Mr G. Kaltenboeck	University of Vienna
Professor Y. Nishimitsu	University of Kobe, Japan
Mr Hamish Norbrook	BBC

Professor H. Nyyssonen	University of Oulu, Finland
Mrs A.G. Obermer	Michael Marks Charitable Trust
Mr T. Onuma	Kenkyusha Press, Tokyo
Dr N. Ostler	Department of Trade & Industry
Mr J. Price	Routledge
Mr Paul Proctor	Cambridge University Press
Mr K. Ricketts	BBC
Mr L. Song	Institute of Education, University of London
Professor A.-B. Stenström	University of Bergen, Norway
Professor N. Takahashi	Tokyo University of Foreign Studies
Dr W. Teubert	University of Mannheim, Germany
Professor J. Thiesmeyer	Hobart & William Smith College, USA
Mr D. Tiomajou	University of Yaounde, Cameroon
Dr L. Urbanova	University of Prešov, Czechoslovakia
Dr E. User	Michael Marks Charitable Trust
Dr T. Varadi	Hungarian Academy of Sciences, Budapest
Professor G. Veikhman	Moscow Linguistics University
Professor E. Yakovleva	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

Aarts, B. (1992) Small Clauses in English: The Nonverbal Types. Berlin: Mouton de Gruyter.

Altenberg, B. (1991) 'The London-Lund Corpus of Spoken English: Research and Applications', Using Corpora. Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research, 71-83. Ontario: UW Centre for the New OED and Text Research, University of Waterloo, and Oxford: Oxford University Press.

Antonopoulou, E. (1991) Agent-defocusing mechanisms in Spoken English: A cognitive explanation of impersonalization. Athens.

Bald, W.-D. (1991) 'Haupt- und Nebenfunktion in der englischen Grammatik', Fremdsprachen Lehren und Lernen 20, 132-143.

Bald, W.-D. (1991) 'Modal auxiliaries: Form and function in texts', Anglistentag 1990 Marlburg: Proceedings, eds. C. Uhlig and R. Zimmermann, 348-361. Tubingen: Niemeyer.

Biber, D. and E. Finegan (1991) 'On the exploration of computerized corpora in variation studies', English Corpus Linguistics. Studies in Honour of Jan Svartvik, eds. K. Aijmer and B. Altenberg, 204-220. London: Longman.

Breivik, L.E. (1990) Existential 'There': A Synchronic and Diachronic Study, 2nd edition. Oslo: Novus Press.

Collins, P. (1991) 'Pseudocleft and cleft constructions: A thematic and informational interpretation', Linguistics 29, 481-519.

Collins, P. (1991) 'The modals of obligation and necessity in Australian English', English Corpus Linguistics. Studies in Honour of Jan Svartvik, eds. K. Aijmer and B. Altenberg, 145-165. London: Longman.

Firbas, J. (1992) Functional sentence perspective in written and spoken communication. Cambridge: Cambridge University Press.

Geluykens, R. (1991) 'Discourse Functions of It-Clefts in English Conversation', Communication and Cognition 24, 343-358.

Greenbaum, S. (with Janet Whitcut) (1991) The Longman Guide to English Usage. London: The Softback Preview [paperback edition].

Greenbaum, S. (1992) 'A new corpus of English: ICE', Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm 4-8 August 1991, ed. J. Svartvik, 171-179. Berlin: Mouton de Gruyter.

Greenbaum, S. (1992) associate editor [for grammar entries], The Oxford Companion to the English Language, ed. T. McArthur. Oxford: Oxford University Press [contributed over 130 entries].

Hasselgard, H. (1991) 'Sequences of temporal and spatial adverbials in spoken English: Some pragmatic considerations', ICAME Journal 15, 3-17.

Ilson, R.F. (1991) Assembling, Analysing and Using a Corpus of Authentic Language. Budapest: Institutum Linguisticum Academiae Scientiarum Hungaricae.

Ilson, R.F. (1991) 'Lexicography', The Linguistics Encyclopedia, ed. K. Malmkjaer, 291-298. London: Routledge.

Ilson, R.F. (with Simon Jenkins) (1992) The Times English Style and Usage Guide. London: Times Books.

Ilson, R.F. (1992) 'Looking for the Words', Disabling World, eds. P. Barker and D. Jones, 19-22. London: Channel 4 Television.

Ilson, R.F. (1992) contributor to The Oxford Companion to the English Language, ed. T. McArthur. Oxford: Oxford University Press.

Kennedy, G. (1992) 'Preferred ways of putting things with implications for language teaching', Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82. Stockholm 4-8 August 1991, ed. J. Svartvik, 335-373. Berlin: Mouton de Gruyter.

Meyer, C.F. (1992) 'A corpus-based study of apposition in English', English Corpus Linguistics. Studies in Honour of Jan Svartvik, eds. K. Aijmer and B. Altenberg, 166-181. London: Longman.

Meyer, C.F. (1991) Apposition in Contemporary English. Cambridge: Cambridge University Press.

Mindt, D. (1991) 'Syntactic evidence for semantic distinctions in English', English Corpus Linguistics. Studies in Honour of Jan Svartvik, eds. K. Aijmer and B. Altenberg. 182-196. London: Longman.

Opdahl, L. (1991) '-ly as adverbial suffix: Corpus and elicited material compared', ICAME Journal 15, 19-35.

Quinault, R. (1992) contributor to The Oxford Companion to the English Language, ed. T. McArthur. Oxford: Oxford University press.

Stenström, A.-B. (1991) 'Expletives in the London-Lund Corpus', English Corpus Linguistics. Studies in Honour of Jan Svartvik, eds. K. Aijmer and B. Altenberg, 239-253. London: Longman.

Svartvik, J. (1991) 'What can real spoken data teach teachers of English?' Georgetown University Round Table on languages and Linguistics 1991, ed. J.E. Alatis, 555-566. Washington, D.C.: Georgetown University Press.

Svartvik, J. (1992) 'Lexis in English language corpora', Euralex '92 Proceedings, eds. H. Tommola, K. Varantola, T. Salmi-Tolonen, and J. Schopp, 17-31. Tampere: Department of Translation Studies.

Tottie, G. (1991) 'Conversational style in British and American English: The case of backchannels', English Corpus Linguistics. Studies in Honour of Jan Svartvik, eds. K. Aijmer and B. Altenberg, 254-271, London: Longman.

Sidney Greenbaum

Director, Survey of English Usage

September 1992