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...

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

A considerable amount of time was spent during the

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 C

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 report

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 options are selected f

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 f

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)

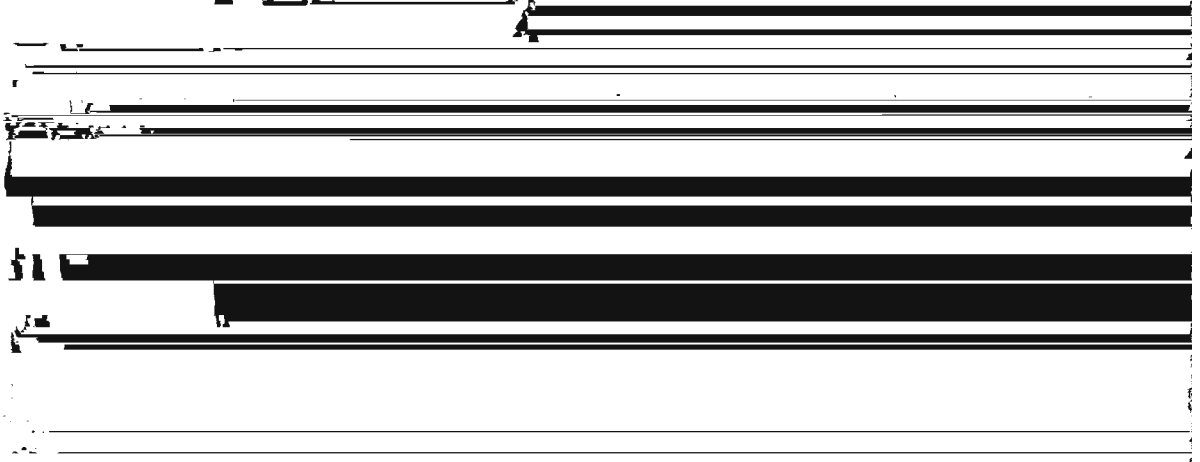**Scripted (50)**
broadcast news (20)
broadcast talks (20)
speeches (not broadcast) (10)

WRITTEN (200)

NON-PRINTED (50)

Non-professional writing (20):

student 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)
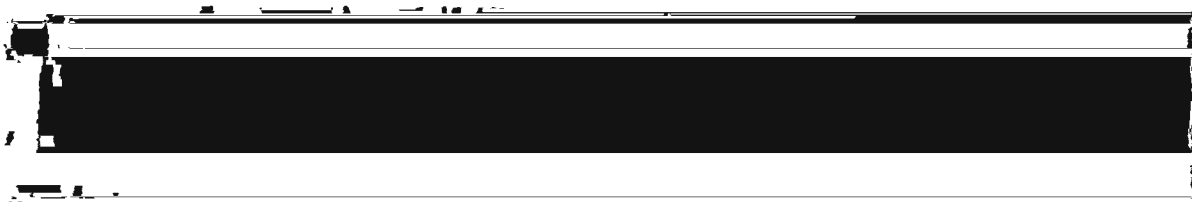
informational (popular) (40):
humanities (10)
social sciences (10)
natural sciences (10)
technology (10)

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

administrative/regulatory
skills/hobbies (10)

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

program and is applying it to ICE-GB using the ICE tagset. In the next stage the automatic TOSCA

Professor Y. Ikegami
Tokyo University

causation

Ms C. Ilie

pragmatics and

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

revision of

Mr A. Vogelmann

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

	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

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 J. Urbanom	University of Paris, France

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.

René Quinault, Akiva Quinn, Oonagh Sayce, and Vladimir Žegarac.

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

Oxford and the ICAME Conference in Nijmegen. Vladimir Žegarac gave a paper at the LAGB Conference in Brighton. Gerald taught at King's College

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 N...

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

Kennedy, G. (1992) 'Preferred ways of putting things with implications for language teaching', Directions in Corpus Linguistics. Proceedings of

, eds. K. Aijmer and
gman.

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

ndon: 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 Universit