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

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

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.

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
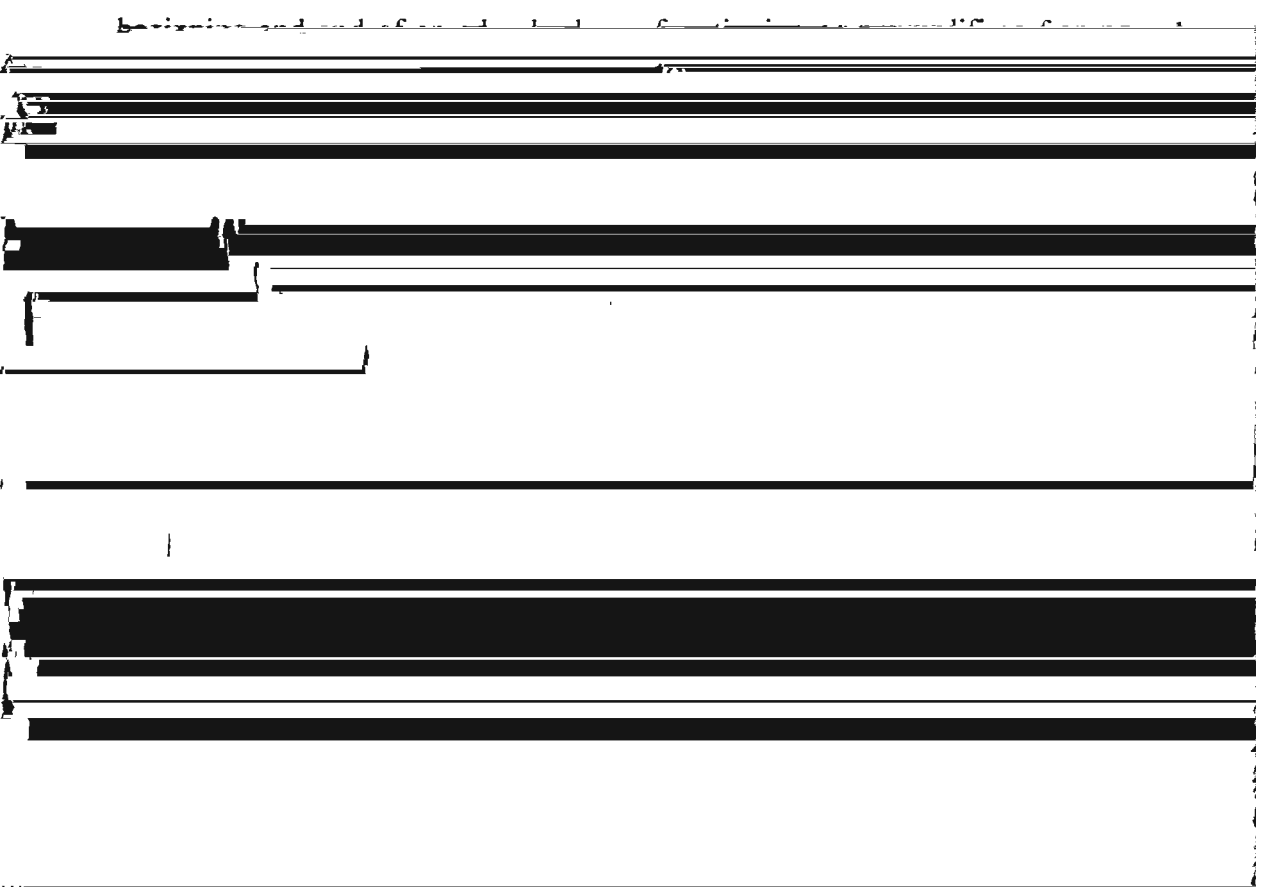
with the ICE parsing. It will then be possible to make diachronic comparisons within the Survey Corpus (which covers a period of about 30 years) and also between parts of the Survey Corpus and corresponding parts in the ICE Corpus.

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

Mrs I. Constaninescu
Institutul de Lingvistica
Ruch̄or̄est̄ R̄om̄on̄io   colloc̄atiō

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

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

Mr S. Crowdie	Longman, UK
Mrs D. Dolores	The Language Centre, UCL
Dr B. Erman	Stockholm University, Sweden

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 Universität Berlin, Germany
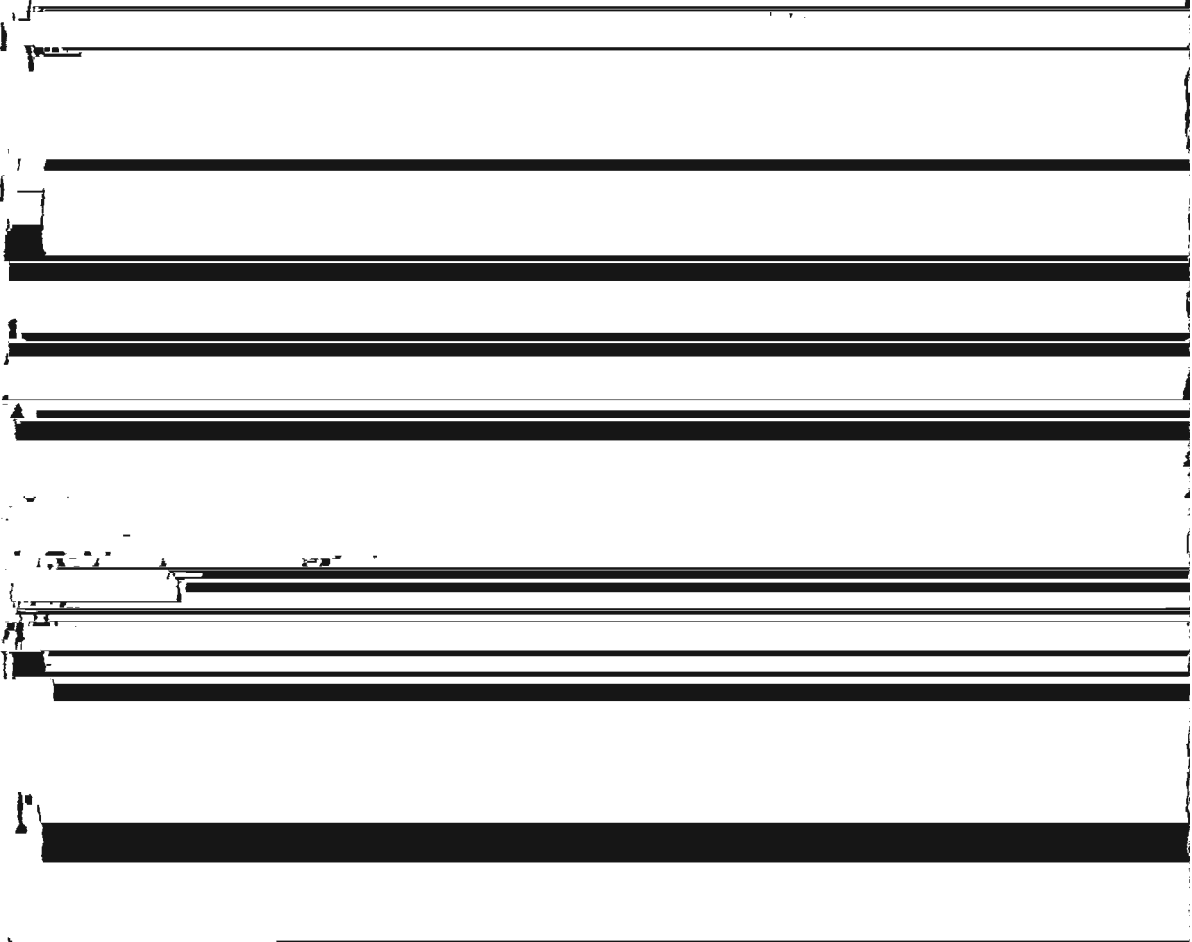
Professor G. Tottie   University of Zurich, Switzerland

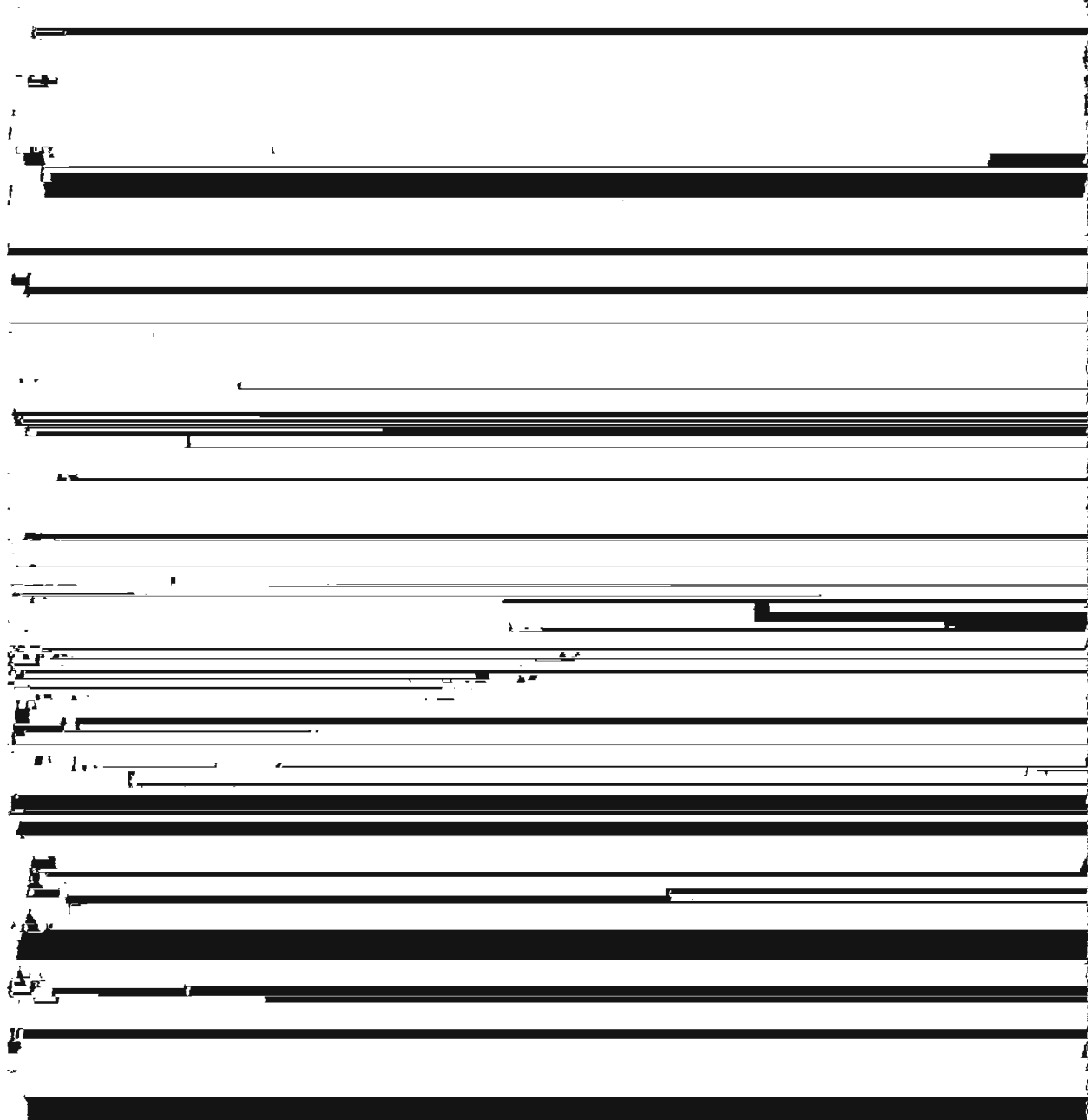Ms L. Updahl   University of Bergen, Norway

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,

participants are Isaac Hallegua, Neil Morgenstern, René

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

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 <u>Language Sciences</u> (Pergamon Press).

Biber, D. and
E. Finegan (1992)

Mouton de Gruyter.

Clark, C.H. and   'Quotations as demonstrations', Language 66, 764-805.
R.J. Gerrig (1990)
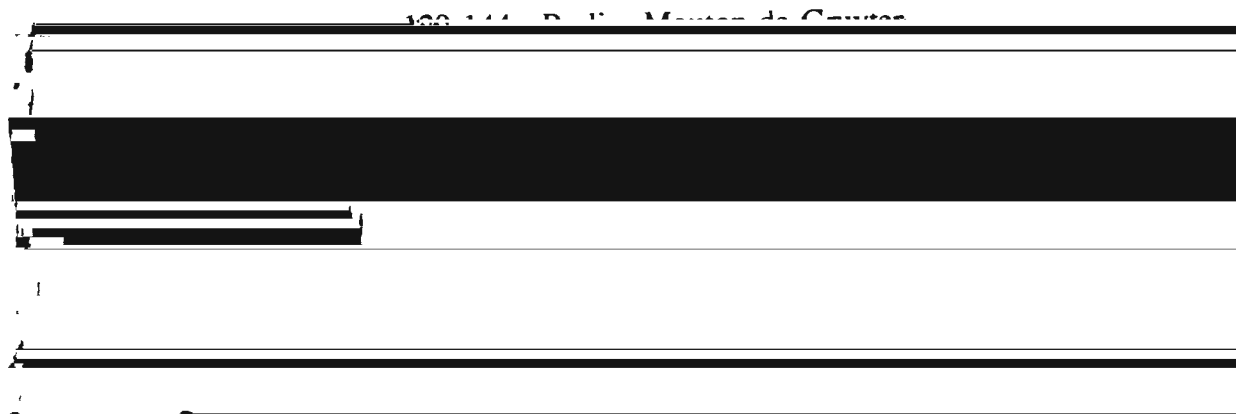
Collins, P. (1991)   Cleft and pseudo-cleft constructions in English.   London:
                       Routledge.

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.

Erman, B. (1990)   'From lexical to pragmatic meaning: Three levels of functions

of English, University of Copenhagen.

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

Fludernik, M. (1993)

London: Routledge.

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

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.

O... (199?)   'An object-oriented design for a Corpus Utility Program'.

Die Indefinitpronomina 'some' und 'any'...

Amsterdam: Rodopi.

Tesch, F. (1990)

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. ... 439-467. Berlin: Mouton de Gruyter.