

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 suggest there to be three complete corpora: ICE



[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

Notwithstanding to the Commission...

[REDACTED]

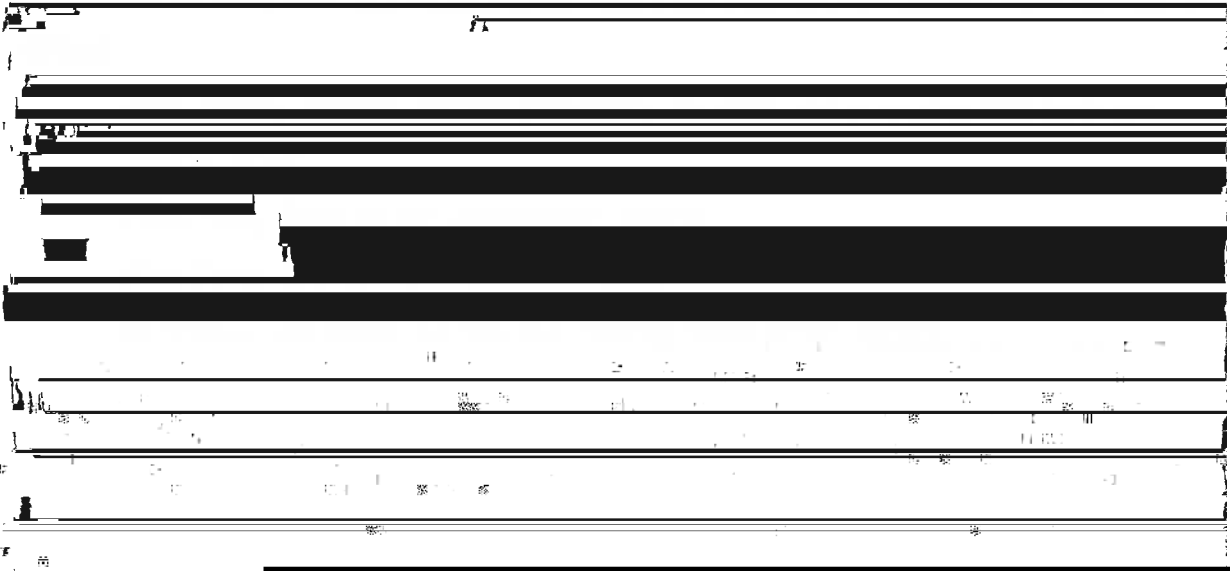
[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]



Nick Ponte, 'Manual for text processing for ICE (for generating full tracks & legends)' (3pp)

Alex Fang, 'Manual for AUTAS' (3pp)

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

Ni Niom, '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 interrogative verbs.

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 information (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

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
Ms S. Murison-Powie	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

married. And Rosta was appointed a lecturer at the Roehampton Institute of Higher

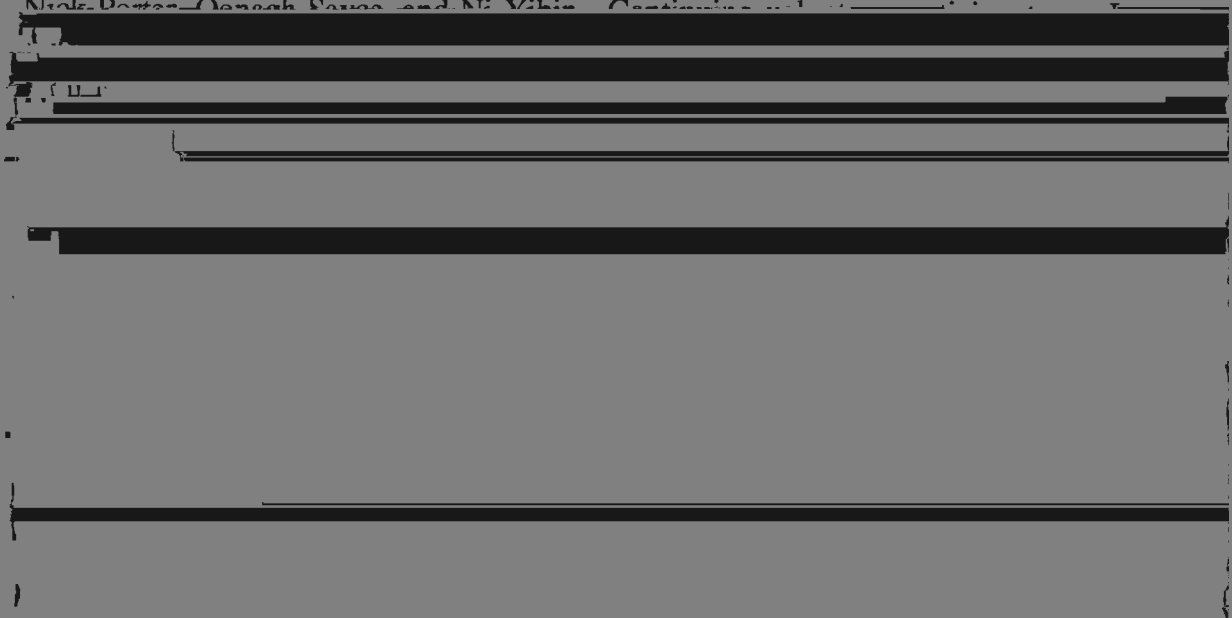
Education. Vlad Zaganj joined a research project at the University of Sussex.



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,

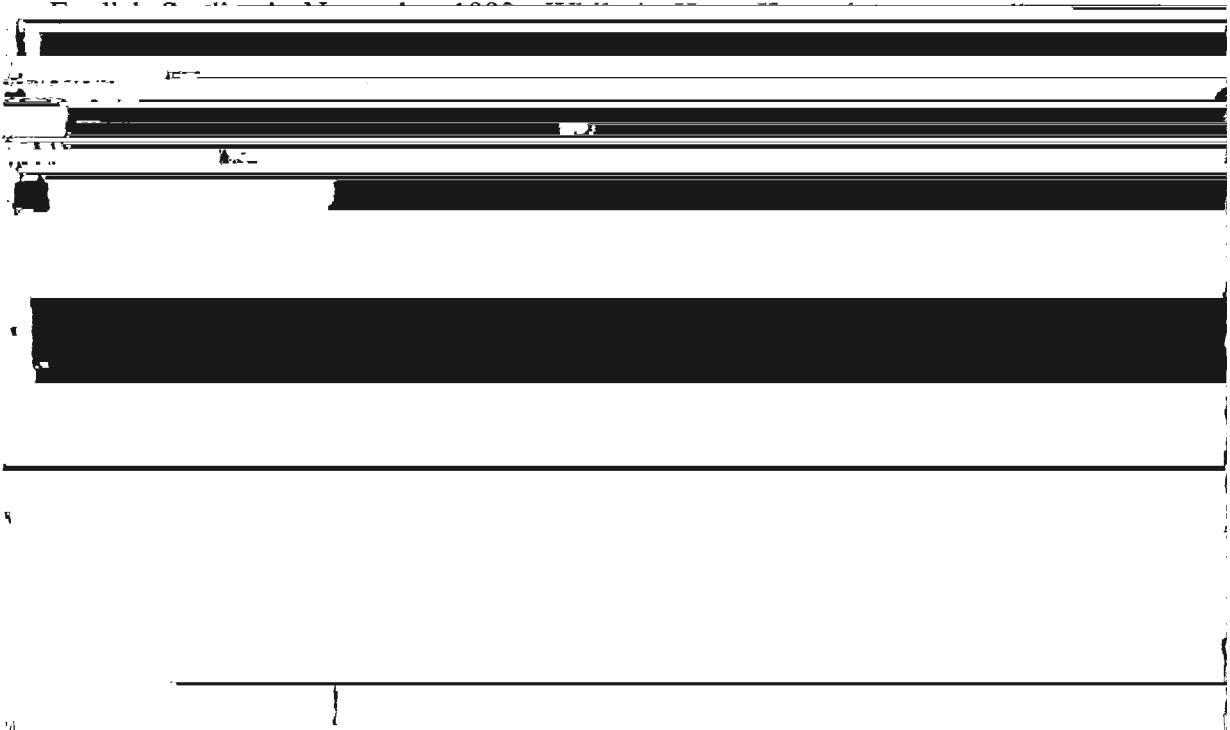
Nick Porter, Oonagh Savage, and Ni Vihia. Continuing staff from last year are



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



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;

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 \emptyset in two corpora of spoken English', English

- 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: 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', 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

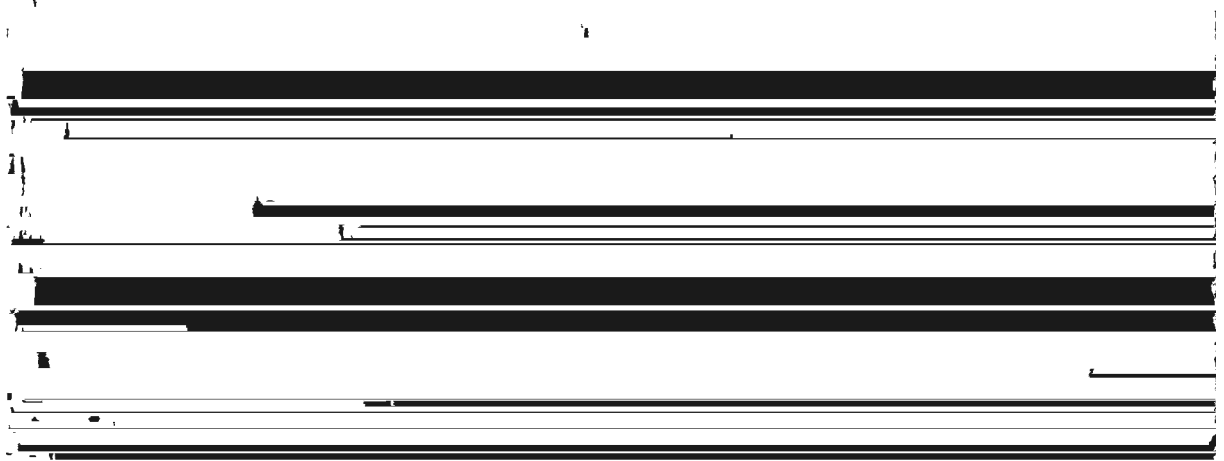
Today 10, 19-24.

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 I. Svartvik (1993) 'Impassable speech: Pauses and other



nonfluencies in spoken English', Corpus-based Research into Language: In Honour

