Annual Report
1996

1. General

It is with sadness that we announce the death in May 1996 of Professor Sidney Greenbaum. He died in Moscow while on a lecture tour. Sid had been Director of the Survey since 1983 when he came to UCL from the University of Wisconsin at Milwaukee to become the Quain Professor of English Language and Literature. He retired from the Quain chair in 1990 to devote all his energies to the Survey. Greenbaum initiated the compilation of the International Corpus of English (ICE), the British component of which we hope to complete by the autumn of 1997. For further details, see the ICE section below.

Our completely redesigned and updated Home Page has been relaunched at the address shown below.

The Survey’s Annual Report will appear around February of each year from now on, covering the preceding calendar year.

2. Research

New project: The Internet Grammar of English (IGE) (Project No. JTAP-49)

This project, funded by the Joint Information Systems Committee (JISC), is a World Wide Web based grammar intended for people who want to learn about English grammar (or improve their existing knowledge) through self-study. It is also useful for teachers and researchers. IGE consists of two modules: The Grammar Module and The Exercise Module. The first of these consists of descriptions of the English language. The Exercise Module offers exercises which are linked to sections in the Grammar Module. There will also be a Foundation Course, designed especially for absolute beginners. The descriptive outlook of IGE is predominantly traditional, but it will incorporate some insights from theoretical grammar. The Web
address for *IGE* is the same as the Survey's:  http://www.ucl.ac.uk/english-usage/. Simply follow the *IGE* link. This project started in August 1996, and we anticipate completing it in July 1998. In due course *IGE* will also be available on CD-ROM for institutions of higher education and schools. Gerald Nelson is the senior researcher working on the project. Justin Buckley works on the computational side, his chief task being the design of the Web site. Judith Broadbent assists on a part-time basis.

**Continuing projects:**

**The International Corpus of English**

The parsing of the British ICE corpus (ICE-GB) has been completed, and work has begun on aligning the tagged and parsed versions. This will yield a single, unified version of the corpus. The ICETree program has been modified to handle the alignment process.

We have begun the speech digitization project, computerizing the recordings of spoken texts in ICE-GB and aligning them to the orthographic transcriptions. The digitized sound files are being archived on CD-ROM. As of October 1997 ICE-GB will be on CD-ROM.

ICECUP III, the ICE Corpus Utility Program, has been developed for searching, browsing and analysing corpora. It is the corpus equivalent of a database browsing program. ICECUP III is a new version of the program which is based around the parsed version of the ICE-GB corpus, but with compatibility for lexical and tagged corpora. The current version of ICECUP III contains a retrieval system for simple searches and a browser for sentences. Another module is used to construct hash tables and indices in advance. The main parts of the work outstanding are integrating the visualisation of trees and the specification of complex queries into the tool.


Sidney Greenbaum and Gerry Nelson co-edited a special issue of *World Englishes* devoted to the ICE project. The volume (1,15) was published in March and contains eight research papers based on component corpora of ICE.
The Survey Parser Project (EPSRC Grant No GR/K75033)

The Survey Parser Project is for the development of an automatic syntactic parser of English and a lexical database that provides useful information for parsing. Alex Chengyu Fang was employed for the parsing component and commenced work in September 1995. Sean Wallis was employed for the lexical component and joined us in October 1995. The Survey Parser is now functioning. It is fully automatic, fast, and produces only one analysis for the input string. The current design of the parser comprises two modules: wordclass tagging and syntactic parsing. Raw texts can be fed into the parser without having to be pre-edited. The output is an indented tree structure with appropriate labels. The parser is able to process 30 words per second, a speed that is unaffected by the length of the input string. The single analysis produced can be complete or partial. A complete analysis is one where all the constituents have been successfully labelled. A partial parse has one or more missing syntactic functions. The parser has been used to parse the residue of the British component of the International Corpus of English (ICE-GB), that could not be dealt with by the TOSCA parser (Nijmegen, the Netherlands). This residue represents about 320,000 words or 28,000 sentences. The Survey Parser successfully produced either a global or a partial analysis for all the input strings. The next major development will be the interface between the parser and the lexical database. It is expected that information from the lexical database will contribute to better parsing outputs. The lexical database (LDB) currently consists of 162,989 entries. Each entry consists of a unique headword and a series of sub-entries, distinguished by the part of speech, and containing cross-references to other appropriate word forms, including inflected, derived, compounded and alternative forms. Each entry may additionally include atomic semantic information, and a list of possible collocational frames. This information has been aggregated from a number of sources: the Collins English Dictionary, the Oxford Advanced Learner’s Dictionary, and the Longman Dictionary of Contemporary English.

3. Staff

Bas Aarts was appointed Director of the Survey at the beginning of 1997.

Continuing staff from last year are Judith Broadbent, Justin Buckley, Celine Bijleveld, Marie Gibney, Isaac Hallegua, Gerry Nelson, René Quinault, Sean Wallis, and Jonathan White. Other staff who joined us this year to work on
the alignment of computerised texts are: Brian Davies, and Gunther Kaltenboeck.

We congratulate Judith Broadbent on obtaining her PhD.

We also congratulate Yibin Ni on obtaining his PhD, and on his appointment as lecturer at the University of Singapore, a post he will take up later in the year.

Alex Fang has left the Survey and now works on the EPSRC project in the Department of Phonetics and Linguistics at UCL, where he can be reached by e-mail: alex@phonetics.ucl.ac.uk.

Ken Fletcher has left the Survey for Central China, where he will be working for VSO.

Oonagh Sayce retired from the Survey in December, but still works on an occasional basis.

4. Publications based on Survey material


Bas Aarts
Director
February 1997