Annual Report

1997

1. General

There has been steady progress on the Internet Grammar of English (IGE), which we hope to complete in the course of 1998. IGE will be freely available over the Web, but we will also having to write an expression in logic. Preliminary results show that this method will be easy to use, and impressively fast to process, due to the fact that each individual element of a tree is separately indexed. (Simple single-element searches are very fast: under 2 seconds on a 100MHz Pentium from the hard disk), while compound element (FTF) searches are better than proportional to the number of candidates. Obviously, actual performance will vary according to the computer being used and the speed of the drive on which the corpus resides. We will provide a variety of installation options to allow users to move the entire corpus, or portions, onto their hard disk. Other facilities will include a capacity to perform concordances on lexical items and entire linguistic structures, and simple statistical facilities. Updates to the software will be possible via the Survey web page. The ESRC funding was awarded to extend, improve and evaluate the software.

New Project: Landmarks in English Grammar

Landmarks in English Grammar: The Eighteenth Century is a collection of five classic eighteenth-century grammars of English. They are bundled on a single CD-ROM together with a copy of Acrobat Reader, the software used to view the texts. The texts in this collection have

Landmarks in English Grammar will be available from the Survey of English Usage early in 1998.

Continuing Projects:

The International Corpus of English

At the ICAME conference in Chester it was decided that Professor Chuck Meyer (meyer@cs.umb.edu) would take over the coordination of the International ICE project. ICE-GB will continue to be led from the Survey, where advice will be given to international teams by Gerry Nelson.

A CD-ROM containing the ICE-GB corpus as well as The International Corpus of English Corpus Utility Programme (ICECUP) is now almost ready, and will be released in May. See http://www.ucl.ac.uk/english-usage for the most up-to-date information.

The Survey's software for searching, browsing and analysing the corpus, ICECUP III, contains facilities to perform fast searches for individual lexical items, tags, and individual nodes of a tree. It will also include a facility to perform a search for "fuzzy tree fragments", described above.

A paper on the human-computer aspect of the Survey's work in checking the ICE-GB corpus (Wallis & Nelson, 1997) was presented to the European workshop on Knowledge Acquisition in Catalonia, Spain. The ICE-GB corpus checking procedure was treated as a case study to compare the Survey tree editor ICE Tree II (written by Sean Wallis) and the earlier editor, ICE Tree. The large scale of the task is unprecedented in the KA community, a perhaps unexpected spin-off of the work on ICE-GB.

ICE Tree II is available from the Survey, and is free for ICE Teams upon application from the team coordinator. A time-limited demonstration version can be downloaded from the Survey Web Site.

The Internet Grammar of English (JISC/JTAP Project No. 49)

We have made good progress on the Internet Grammar. Sections on all the word classes have now been written, as well as sections on phrases and clauses. Most of these are currently available online to reviewers, to whom we give access via a login and password. We have received useful feedback from several of these reviewers. Please let us know if you would like to have access to the site. We welcome all comments.

The Internet Grammar will include a comprehensive Glossary of grammatical terms, and an Index. The Glossary is currently being compiled, as the grammar is being written.

We are working on various ways to make the Grammar more interactive. We have incorporated interactive exercises into the text pages, and we are experimenting with animated graphics, which we hope to use to illustrate various features of phrase and clause structure. We have been testing the Grammar on a range of platforms, to ensure that is it accessible to as many people as possible.

Completed project:

The Survey Parser Project (EPSRC Grant No. GR/K75003)

This UK EPSRC-funded project finished successfully at the end of October 1997, and the final report has been submitted. The parser exploits ICE-GB to inductively generate rules governing the

phrase structure of ‘canonical grammatical phrases’, i.e. adverb, adjective, noun, prepositional, and verb phrases. These rules are qualitative, rather than probabilistic, which means that it is far easier to modify them by hand, and they can be more linguistically meaningful. The parser is robust, fast, and fairly accurate, achieving rates of 70% coverage of noun phrases extracted from dictionary definitions, with 90% having an exact match with an annotation performed independently by hand. Since this is a parser that has been trained on a corpus, we expect the coverage to improve as the range and quality of the original corpus annotation is improved. Four samples of about 68,000 words each were analysed for verb phrase boundaries, internal structures, and labels. Over 95% accuracy was achieved for all of the samples (Fang, 1997). To support the parser, a large-scale lexical database was constructed, with 160,000 individually indexed entries, each of a distinct word form. The database includes cross-reference information (word forms, derivations, compounds etc.), which is supported by semantic information specifying the classes of nouns, verbs and adjectives. In the project, this information was used to support the parser when deciding between functional types of prepositional phrases.

3. Staff

Celine Bijleveld has left to take up a full-time appointment. We thank her for the work she did for us and wish her luck in her new position as sub-editor at EL Gazette.

Judith Broadbent was appointed lecturer at the Roehampton Institute. We wish her luck in her new job, and thank her for her work in the Survey over the past five years.

Justin Buckley continues to work as a Web designer on the IGE project. He has also taken care of the technical implementation of the *Landmarks* CD-ROM, and has an advisory role on the writing of the ICE-GB manual.

Marie Gibney continues as the Survey's administrator. Her knowledge of the Survey and such matters as funding application procedures has been invaluable. We hope that she will remain with us in the Survey well into the next century.

Isaac Hallegua continues to work on systems and data management. His assistance is highly valued.

Gerry Nelson is principally employed on the IGE project, for which he writes the content. He is also working on the ICE project and initiated the *Landmarks* project described above. Needless to say, his knowledge of the ICE project is invaluable to us, and to ICE teams worldwide.

René Quinault has been working on re-checking some of the more difficult transcriptions in ICE- GB, as well as on the splitting of the text soundfiles in line with the sentence numbering. He has also looked after the needs of visitors requiring to study the recordings of the first Survey corpus. We are grateful for his continued support.

Sean Wallis has worked on the EPSRC parser project, and full-time on ICECUP since the parser project came to an end. He designed, and will be working on, the new ESRC funded FTF project described above.

Jonathan White works on IGE, as well as on the ICE-GB project, specifically the ICE-GB manual.

4. Publications, conference presentations, and studies using Survey material

Aarts, Bas, (1997) *English syntax and argumentation*. Macmillan Modern Linguistics Series. Basingstoke and London: Macmillan.

Aarts, Bas (1997) The role of argumentation in the description of English. In: Jan Aarts, Inge de Mönninck and Herman Wekker (eds.)

Fang, Alex Chengyu (1997) Prepositional Phrases: Towards the Automatic Determination of their Syntactic Functions. Paper presented at the18th ICAME Conference, Chester, UK. May 21-25.

Hoye, Leo (1997) *Adverbs and Modality in English*. London: Longman.

Ljung, Magnus, (1997) A genre-based study of English subordinator-headed non-finite and verbless adverbial clauses. In: *To explain the present: Studies in the changing English Language in honour of Matti Rissanen*. Terttu Nevalainen and Leena Kahlas-Tarkka (eds.) Helsinki: Societé Néophilologique. Mémoires de la Société Néophilologique de Helsinki, LII, 375-394

Nelson, G. (1997) A study of the top 100 wordforms in the ICE-GB text categories. *International Journal of Lexicography*, 10.2, 112-134.