BAS AARTS
SEAN WALLIS
Department of English Language and Literature
University College
London, United Kingdom
b.aarts@ucl.ac.uk

Recent developments in the syntactic annotation of corpora: a demonstration of ICE-GB and DCPSE

Introduction

The field of Corpus Linguistics is rapidly developing and expanding. From the early 1990s a number of corpora of substantial scale, descriptive sophistication and diversity have been constructed.
The *British Component of the International Corpus of English* (ICE-GB, Nelson *et al* 2002)[1] is a one million word corpus of spoken and written English, based at the Survey of English Usage (SEU) at University College London. In this corpus each sentence is *POS-tagged*, which means that a part-of-speech label is assigned to every word. In addition ICE-GB is *parsed*, i.e. it is given a full grammatical analysis. This is done in the form of tree diagrams. An example of a tree from ICE-GB is shown in Figure 1 below.

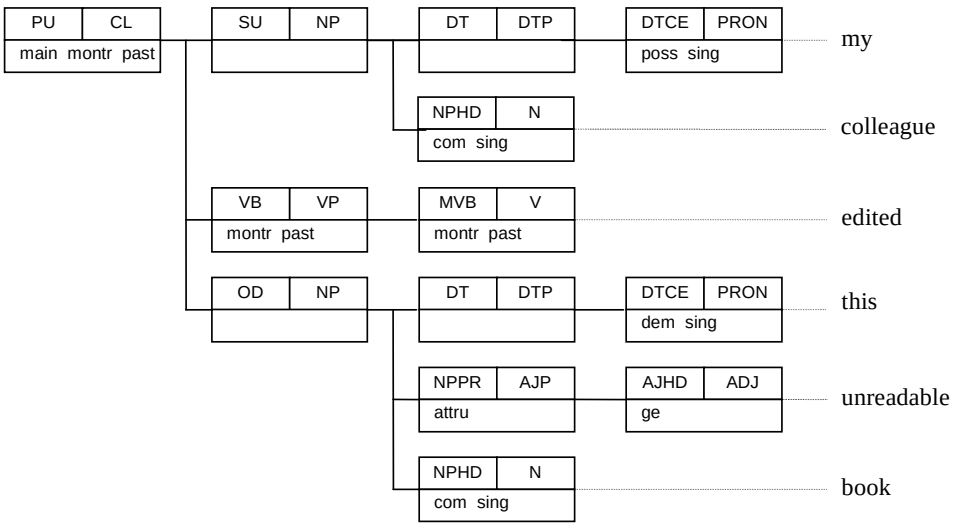

Figure 1: An ICE-GB tree for the sentence *My colleague edited this unreadable book*.[2]

In this representation each node (box) is assigned a *function* label (top left), a *category* label (top right), as well as *features* (lower part of the box) which may percolate upwards (e.g. the features 'montr' and 'past' have percolated up from the verb *edited* to the highest level, PU). ICE-GB contains 600,000 words of transcribed speech and 400,000 words of written English.
ICE-GB can be exploited with dedicated research tools, such as the innovative ICECUP (International Corpus of English Corpus Utility Program) software, developed at the SEU, with which linguists can search for grammatical constructions. This software is currently distributed with the ICE-GB corpus and over the web (together with a sample corpus). It runs over a network or on stand-alone PCs. At the heart of ICECUP is the Fuzzy Tree Fragments (FTFs) facility which enables users to construct approximate (hence 'fuzzy') models of tree structures, which the computer can search for in the corpus. Figure 2 shows an example of an FTF which matches all instances of a verb phrase (VP) followed by a direct object (OD).

[1] See www.ucl.ac.uk/english-usage/ice-gb, where information about the corpus and software downloads are available for review.
[2] Gloss (features are in italics): PU=parse unit, CL=clause, *main=main, montr=monotransitive, past=past tense,* SU=subject, NP=noun phrase, DT=determiner, DTP=determiner phrase, DTCE=central determiner, PRON=pronoun, *poss=possessive*, *sing=singular*, NPHD=NP head, N=noun, *com=common*, VB=verbal, VP=verb phrase, MVB=main verb, V=verb, OD=direct object, *dem=demonstrative*, NPPR=NP premodifier, AJP=adjective phrase, AJHD=adjective head, ADJ=adjective, *attru=attributive, ge=general*.

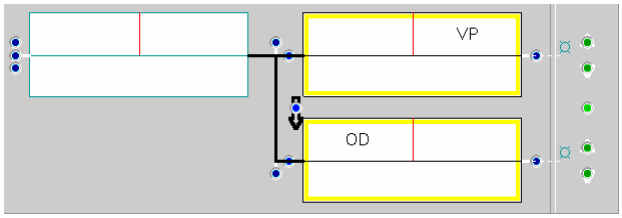

Figure 2: A simple FTF created with ICECUP

For more details on ICE-GB and ICECUP, see Greenbaum (1996), Aarts, Nelson and Wallis (1998) and Nelson, Wallis and Aarts (2002). The latter is a handbook for using ICE-GB and ICECUP.

In this paper I will present the latest version of the exploration software, applied to ICE-GB, as well as to the *Diachronic Corpus of Present-day Spoken English* (DCPSE), which will be described in the next section.

DCSPE: incorporating the LLC and ICE-GB

DCPSE, recently developed at the Survey of English Usage, contains spoken material from two corpora of Modern British English, both founded at the SEU: the *London-Lund Corpus (LLC)*, compiled in the 1960s, and the *British Component of the International Corpus of English*, compiled in the 1990s and described above. The London-Lund Corpus is the spoken part of the *Survey of English Usage Corpus*, founded by Randolph Quirk. It contains 510,576 words of 1960s spoken English. The corpus is divided into 'texts' of 5,000 words each which were transcribed and prosodically annotated (incorporating tone units, onsets, stresses etc.). Thirty-four texts were published in Quirk and Svartvik (1980). The corpus was computerised by Jan Svartvik (Svartvik 1990). Many scholars have used the LLC for their research, resulting in hundreds of publications, principal among them Quirk *et al.* (1972, 1985). It is still one of the largest and most widely used corpora of spoken English, not least because it is the only English corpus that is prosodically annotated. Kennedy (1998: 32) stresses the importance of the LLC in its own right for the study of spoken British English, but also as "a very important baseline record of data…by which other corpora of spoken English can be evaluated… [The texts] have been used by researchers in many countries for studies which go well beyond the study of phonology. The detailed annotation has also facilitated numerous studies of lexis, grammar and especially discourse structure and function". The SEU has enhanced the corpus by adding wordclass tags to it, using the ICE-GB scheme. In addition, the SEU has digitised the original sound recordings which will be supplied – for the first time in the LLC's history – with the new resource.

Diachronic and synchronic linguistics

Traditionally a distinction is made between diachronic and synchronic approaches to linguistics. The first considers language as it develops through time, whereas the latter takes a 'snapshot' look at languages viewed from the present. This old Saussurean dichotomy has recently been called into question, and some linguists have argued that the distinction is an artificial one. These linguists would argue that languages change all the time, even within the synchronic phases. As a result of these new attitudes to language development there is a new research impetus in linguistics which concerns itself with recent change (see Mair 1995, 1997; Mair and Hundt 1995, 1997, Denison 1998, Leech 2000, Smith and Leech 2001).

For linguists who are interested in recent change corpora are especially valuable for data-gathering. At present they will need two separate corpora from two different periods. Naturally, these corpora must be comparable as regards their internal composition (i.e. sampling criteria). An example of work done in this area is Aarts and Aarts (2002) which investigates the use of the English relative pronoun *whom*. In order to compare data from two periods of Present-Day English (PDE) the authors looked at material from the LLC and ICE-GB. They found that the overall use of *whom* as a Direct Object has become 90% less frequent over thirty years. Although ICE-GB is grammatically annotated and fully searchable, manual counts had to be carried out to find data in the older corpus. Thus, while the corpora were indispensable tools for this study, the research phase still required the careful pre-selection of comparable texts and manual searching of the LLC. A parsed LLC is essential to permit the systematic exploration of grammatical variation over time, and will greatly facilitate research of this type, especially if it involves complex grammatical patterns.

In order to support research into current change Christian Mair at the University of Freiburg has constructed two corpora of 1990s English: FLOB (Freiburg-Lancaster-Oslo-Bergen) and FROWN (Freiburg-Brown). These corpora are intended to match the LOB (Lancaster-Oslo-Bergen) and Brown corpora containing written English from the 1960s. These are excellent resources enabling linguists to research changes in written English over 30 years. Manual searches are still unavoidable, however, because these corpora have not been parsed. We have taken Mair's initiative further. We have constructed DCPSE to provide linguists interested in recent changes in English with a new and innovative database containing spoken English covering a period of 25-30 years. We opted for a corpus of spoken English because it is generally recognised that spoken language is primary and the first locus of changes in lexis and grammar. DCPSE differs from FLOB and FROWN in a number of important ways. Firstly, the corpus is unique in containing exclusively spontaneous spoken English. We provide a playback facility enabling linguists to listen to the original recordings. Secondly, the corpus is parsed which permits research into synchronic and diachronic grammatical variation. Thirdly, the corpus is fully searchable using the ICECUP software that we developed for ICE-GB. This software has been modified to operate on the new data. We envisage that DCPSE

will be a major new resource complementing the Freiburg corpora, allowing access for the first time to recordings that could hitherto only be listened to at the SEU premises.

An overview of DCSPE:

- Contains 800,000 words of parsed spoken English from comparable categories in the LLC and in ICE-GB (400,000 words from each corpus). The design of these corpora is similar, and it will thus be possible to select identical categories of spoken English. These include face-to-face conversations, telephone conversations, radio discussions, class discussions, parliamentary debates, legal cross examinations, business transactions, spontaneous speeches and interviews.
- The LLC portion of the corpus contains prosodic markup.
- Digitised sound recordings of both corpora can be listened to.
- The new resource also features a lexicon (a database of word-tag combinations in the corpus) and a grammaticon (a database of node combinations). These will enable users to contrast lexical and grammatical distributions in the LLC and ICE-GB.

References

1. Aarts, B., Nelson, G., and Wallis, S.A. (1998) Using Fuzzy Tree Fragments to Explore English Grammar. *English Today* 14, 52-56.
2. Aarts, F. and B. Aarts (2002) Relative *Whom*: a 'Mischief Maker'. In: A Fischer and G. Tottie (eds.) *Text Types and Corpora*.
3. Denison, D. (1998) Syntax. In: S. Romaine (ed.). *The Cambridge History of the English Language*. IV: 1776-1997. Cambridge. 92-329.
4. Kennedy, G. (1998) *An Introduction to Corpus Linguistics*. London.
5. Leech, G. (2000) Diachronic linguistics across a generation gap: from the 1960s to the 1990s. Paper read at the symposium Grammar and Lexis. University College London Institute of English Studies.
6. Ljung, M. (1997)(ed.) *Corpus-Based Studies in English*. Amsterdam.
7. Mair, C. (1995) Changing Patterns of Complementation and Concomitant Grammaticalisation of the Verb help in Present-Day English. In: B. Aarts, and C.F Meyer (eds.). *The Verb in Contemporary English*, Cambridge. 258- 272.
8. Mair, C. (1997) Parallel Corpora: a Real-Time Approach to the Study of Language Change in Progress. In: M. Ljung, M. (ed.). 195-209.
9. Mair, C. and Hundt, M. (1995) Why is the Progressive Becoming More Frequent in English? A Corpus-Based Investigation of Language Change in Progress. *Zeitschrift für Anglistik und Amerikanistik* 43.2. 111-122.
10. Mair, C. and M. Hundt (1997) The Corpus-Based Approach to Language Change in Progress. In: U. Böker and H. Sauer, H. (eds.). *Anglistentag* 1996. Dresden. 71-82.
11. Nelson, G., Wallis, S.A., and Aarts, B. (2002). *Exploring Natural Language*. Amsterdam.
12. Quirk, R., Greenbaum, S., Leech G., and Svartvik, J. 1972. *A Grammar of Contemporary English*. London.
13. ——— 1985. *A Comprehensive Grammar of the English Language*. London.
14. Smith, N. AND G. Leech (2001) Grammatical change in recent written English, based on the FLOB and LOB corpora. Paper read at the ICAME conference. Louvain-la-Neuve, Belgium.
15. Svartvik, J. 1990 (ed.). *The London-Lund Corpus of Spoken English: Description and Research*. Lund Studies in English 82. Lund.
16. Svartvik, J., and QUIRK, R. 1980. *A Corpus of English Conversation*. Lund.
17. Wallis, S. (1999) Completing parsed corpora: from correction to evolution. In: A. Abeillé (ed.). *Journées ATALA sur les Corpus Annotés pour la Syntaxe* – Treebanks Workshop. 7-12.