COLT: a progress report

Gisle Andersen & Anna-Brita Stenström
University of Bergen

This is an update on COLT (The Bergen Corpus of London Teenage Language), which has been financially supported by The Norwegian Research Council, The Norwegian Academy of Science and the Faculty of Humanities at the University of Bergen.

Phase 1

The COLT material was collected in London by a research team at the University of Bergen in 1993. It consists of roughly half-a-million words of spontaneous conversations between 13- to 17-year old boys and girls from socially different school districts. During the period 1994–95, the conversations were transcribed orthographically (including indication of pauses and overlapping speech) by transcribers engaged by the Longman Group, and tagged for word-classes by a team at Lancaster university. In this form, COLT has become part of the British National Corpus (BNC). A demo of this version of COLT has been made available on Internet.

At this point, the entire corpus has been checked and edited by the team in Bergen. The frequent occurrence of <unclear> labels and the numerous instances of a question mark for speaker identity in the original transcripts indicate that the transcribers were faced with considerable problems. During our checking process, a great many instances of <unclear> have disappeared, most of the speakers have been identified (with a substitution of the original names by fictitious ones), and mistakes in the original transcription have been straightened out. As a result, we have not only ended up with a transcription that is more faithful to the tape-recordings but also with a larger corpus; the number of words has increased by at least 15 per cent. This, in turn, has had the effect that the original word class tagging has become partly inadequate and that the edited corpus will have to be retagged.

The retagging, which will be done by means of the Xanthippe software with assistance from Lancaster university, will be completed in the early

autumn. A first, orthographically transcribed, word class-tagged version of COLT, with a search program, will then be produced on CD-ROM with help from the Norwegian Computer Centre for the Humanities at the University of Bergen, and will be launched in the autumn of 1996.[1]

Phase 2

Phase 2 involves the production of a second, more sophisticated version of COLT on CD-ROM, including sound files and a prosodically rather than orthographically transcribed text. The prosodic analysis into tone units will be carried out by research assistants at the University of Bergen, and the mark-up will be in line with the conventions outlined in Haslerud and Stenström (1994).

From the outset of the COLT-project, it has been a stated aim to launch a final version of the corpus including sound files comprising the actual recordings of the conversations that have been transcribed. The advantage of this is obvious: it enables the researcher to make judgements as to the phonemic and phonological properties of the recorded speech, and to analyze prosodic aspects such as pitch, loudness, tempo and rhythm in a much more subtle way than our simplified prosodic marking of the texts will allow.

So far, nothing has been done to improve the sound quality of the audio tapes containing the COLT-conversations. The sound quality varies a lot from one recording to the next. The fact that quite a few of the conversations took place on the bus, near a road with heavy traffic, in the school playground or in very noisy classrooms (!) underlines the need for sound editing of the recordings. However, the removal of disturbing noise takes place at a certain risk and must be performed with caution to avoid the sound of the actual conversations being impaired. To assist us in this large-scale audio-editing, we are fortunate enough to have the collaboration of the Bergen University Media Centre.

The next step in this connection is to digitize the edited analogous recordings. Each conversation will be divided into sound files of standard length, each sound file corresponding to an unspecified number of words in the text. We are not certain which time-span will be optimal, but it seems clear that each sound file will not exceed 20 seconds. To ensure cohesion, the sound files will overlap by approximately two seconds.

Finally, the concordance of sound and text requires a re-indexation of the texts, including the insertion of a tag which indicates where one sound file ends and the next begins. The links from text to sound files

will allow the researcher to access the audio version of the corpus by a mere mouse-click on the extract s/he wants to listen to.

COLT-based research

COLT material has provided the basis for both PhD and MA theses at the University of Bergen. One MA thesis has already been completed (Andersen, unpublished); six are under way, on the following topics: 'The use of tags', 'The pragmatic particle *innit*', 'Backchannelling', 'Vague language', 'Conflict talk' and 'Metaphors'. One PhD dissertation, on 'Age-specific discourse strategies', is also under way. Most of these studies are sociolinguistic in nature. In some, relevance theory is adopted for interpreting utterances. Finally, two PhD students, who are researching learner language, use COLT conversations for comparison.

Small samples of COLT are also used for research outside Bergen, eg at Stockholm University and Åbo Academy, and guest students from abroad have spent time at the COLT project to be able to study the entire corpus as well as listen to the recordings.

Note

To obtain a copy of COLT I, please contact knut.hofland@hd.uib.no or stenstroem@eng.uib.no

References

Andersen, Gisle. Omission of the primary verbs BE and HAVE in London teenage speech – a sociolinguistic study. Unpublished MA thesis. Department of English, University of Bergen.

Andersen, Gisle & Anna-Brita Stenström. 1996. More trends in teenage talk: A corpus-based investigation of the discourse items *cos* and *innit*. In Ian Lancaster *et al*. (eds.) *Synchronic corpus linguistics*. Amsterdam: Rodopi.

Haslerud, Vibecke & Anna-Brita Stenström. 1994. COLT: Mark-up and trends. *Hermes Journal of Linguistics*, 55–70.

Haslerud, Vibecke & Anna-Brita Stenström. 1995. The Bergen Corpus of London Teenager Language (COLT). In G. Leech, G. Myers & J. Thomas (eds.). *Spoken English on computer.* London: Longman, 235–242.

Stenström, Anna-Brita. 1995. Taboos in teenage talk. In G. Melchers & B. Warren (eds.) *Studies in anglistics.* Stockholm Studies in English 85. Stockholm: Almqvist & Wiksell International.

Stenström, Anna-Brita. Forthcoming. Chatting about boys and discussing taboo habits. In R. Horowitz (ed.) *Talking about text: Developing understanding of the world through talk and text.* The University of Texas-San Antonio.