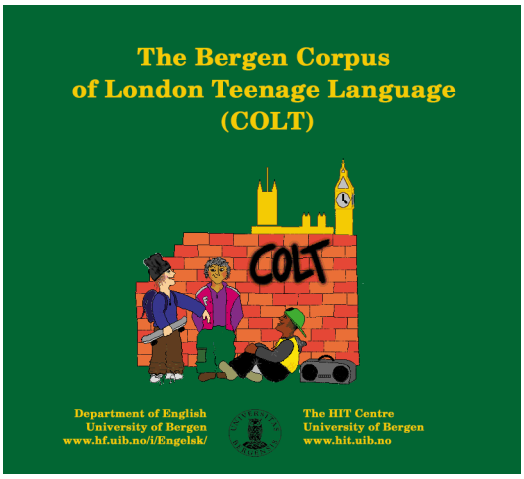
The HIT Centre
University of Bergen
Allégt. 27
N-5007 Bergen, Norway
Tel: +47 55 58 29 54
http://www.hit.uib.no/HIT/colt-pro.htm



The Bergen Corpus of London Teenage Language (COLT)

- COLT is the first, and so far the only, existing corpus of English teenage talk that is available for research worldwide;
- it was collected in five different London school boroughs in 1993 and consists of roughly half a million words of spontaneous conversation (55 hours);
- the conversations were recorded (surreptitiously) by student 'recruits', equipped with a Sony Walkman, a lapel microphone and a log book;
- the speakers, including the recruits, were 13 to 17-year-old girls and boys with different social backgrounds (with the occasional teacher, parent or sibling);
- the recordings have been orthographically transcribed by trained British transcribers;
- part of the material has been subjected to a simplified prosodic analysis;
- the entire corpus has been tagged for word-classes by means of the CLAWS 6 tagset developed at Lancaster University;
- the original sound tapes have been digitised and the digitised sound files have been improved with respect to sound quality by means of the CoolEdit software. The sound files will be available in the Summer of 2000 as a set of 3 CDs with MP3 files.
- a sound/text alignment procedure now enables the corpus user to browse the text with hyperlinks to sound files or to search the text corpus and retrieve a search result that includes hyperlinks to the relevant sound file, with adjustable parameters for sound file length. This has been implemented on a web interface.
- the COLT material is published jointly by the HIT Centre and the Department of English at University of Bergen.
- the material is only available for non-commercial purposes.

For further information contact colt@hit.uib.no

COLT CD-ROM

UNIVERSITAS BERGENSIS