*Vibecke Haslerud & Anna-Brita Stenström**

COLT: Mark-up and Trends

Abstract

The following is a progress report on **COLT** (**A Corpus of London Teenager Language)**[1], the first large English corpus focussing on the language of teenagers. The paper consists of two parts: a description of the mark-up system that will be used for the prosodic transcription of the material is followed by a report on the most obvious linguistic and interactional trends that immediately struck us when we looked at the data.

1. Introduction

COLT is a young corpus, both as regards year of collection and age of the speakers. It was collected in 1993 and consists of the spoken language of 13 to 17-year-old teenagers. The boys and girls involved are from five different school districts in London and have various backgrounds, representing the entire social scale. This means that the corpus is open to linguistic as well as sociolinguistic research.

The complete corpus, half a million words, has already been orthographically transcribed by trained Longman transcribers, and in this form, orthographic transcription, it has been incorporated in the British National Corpus (BNC). However, this is not the final COLT version. The next step at the Bergen end will be to transform the sentence-bound and written-like orthographic transcription into a prosodic, tone-unit-bound and spoken-like transcription. Eventually, COLT will be obtainable on CD-ROM, with tagged text and sound and provided with search programs, from the Norwegian Computing Centre for the Humanities, Bergen.

[1] COLT was funded in 1993/94 by the Norwegian Research Centre for the Humanities (NRF) and by the Faculty of Arts at Bergen University.

* *Anna-Brita Stenström*
* *Vibecke Haslerud*
English Department
University of Bergen
5007-Bergen (N)

2. Mark-up

The transcription level and mark-up model to be selected for any text should be determined by the requirements of the users. Different users will have very different needs. In the case of COLT we know **who** the users will be: researchers around the world who are interested in spoken language and researchers and graduate students at the University of Bergen. What we do not know, however, is the exact **use** to which they will put the corpus. Therefore, we choose to use a model for transcription and encoding which can cater for many different uses: discourse analysis, lexicography, grammatical studies, intonation and phonology studies, etc.

Even though the transcription and mark-up should be as true as possible to the speech and as informative as possible, one should aim at readable texts in which important information is not concealed in over-abstract representations (Milroy 1987). We feel that the COLT system is well balanced in this respect. It is likely that even with a system designed for multiple uses, most researchers will have to add more information to the texts, according to their specific needs. Having the sound available as well as the tagged transcripts makes this reasonably easy.

2.1. Implementing the COLT system

The Longman transcription conventions are an excellent starting point, but since we feel that the Longman level of transcription and mark-up is somewhat restricted, the COLT texts will be converted to a more detailed set of conventions, namely the COLT system for transcription and encoding (cf Haslerud & Stenström forthcoming). In the following, the two mark-up systems, shown in Table 1, will be compared.

As will become clear in what follows, there is agreement between most of the features marked in the two systems. That means that most of the Longman tags can be easily converted to COLT tags by a simple conversion program. Some features which we want to mark in the COLT system are not tagged in the Longman transcripts, however. They will have to be added to the COLT version on the basis of a new auditory analysis. In this process, the orthographic transcription will be refined as well, to provide a more accurate printed representation of the conversations.

Table 1. Corresponding Longman- and COLT-tags

2.1.2. Basic units of analysis

In the Longman transcripts, sentence-like units are marked by a capital letter at the beginning and a punctuation mark (, . ? !) at the end, and the length of the units is determined on the basis of writing conventions and the transcriber's intuition (Crowdy 1991). The COLT system, on the other hand, focuses on tone units, as we are aiming at a representation more true to the spoken language. By 'tone unit' we mean a tonic syllable expounded by a nuclear tone and optionally preceded and/or followed by other syllables (cf Crystal 1975).

No capitals or punctuation marks will appear as we wish to avoid imposing writing conventions on the spoken language. Not only would punctuation and capitals divide the sound material into units which do not always exist in speech, they would also influence the researcher's perception of the sound when listening to it. For instance, often there is no pause between what we think of as sentences but if we see the text divided into sentences on the page, we might think that we hear pauses. This is illustrated below[2], in transcript sample 1: the utterance *Teacher! Teacher! Teacher!* has in fact no audible pauses at all, as we can see in the COLT transcript. Instead of punctuation and capitals, we use tone unit boundaries, indicated by cross-hatches (#). Incomplete tone units are marked by two equal symbols (==).

Punctuation may also impose the wrong function on an utterance. An utterance may have the interrogative form and thus be given a question mark by the Longman transcribers. It may, however, actually function as an exclamation or a statement. We wish to leave as much as possible of this interpretation to the researcher.

[2] Transcript samples 1-5 are drawn from COLT conversation number 132601, which is between three 14 and 15 year-olds, supposedly doing homework under supervision in the school library. All names have been changed to preserve the anonymity of the speakers. See appendix A for a larger excerpt.

Longman transcript

```
<u who=2 id=12>        [I] think you’re really
<u who=1 id=13>        Teacher! Teacher! Teacher!
<u who=2 id=14>        I think you’re really sexy Billy!
```

COLT transcript

```
<1>      sh [quiet +sh +sh +sh] teacher + teacher + teacher==
<2>          [I think you’re R\EAlly s=]#
<2>      *I think you’re R\EAlly# S\/Exy Billy#
```

Transcript sample 1

The different wording and order of utterances in the COLT transcript are the result of a more detailed transcription, which is part of the conversion scheme. The Longman transcribers do not mark repetition and restart. Tags for these features, plus (+) for repetition and asterisk (*) for restart, are added manually to the texts in the conversion process[3]. In the COLT system, the left brackets of overlapping strings are aligned to preserve the readability of the texts (cf Edwards 1993).

2.1.3. Tonic marks

Rhythm is effectively depicted by capitalized nuclear syllables, in addition to the aforementioned tone unit boundaries. We will also add tonic marks (/, \, /\, \/ and _) before the vowel of the nucleus, to represent the function of the utterance and the attitude of the speaker, which are very difficult to interpret from ordinary punctuation. An example of this can be seen in the second line of transcript sample 2: *Alright, too late, he’s yours*. The commas and full stop are not very informative as to the attitude of speaker 2, whereas the rising intonation which is given in the COLT version represents a rather surprised attitude.

[3] See Table 2 for coding conventions.

Longman transcript
<u who=1 id=15> Too late, he’s mine.
<u who=2 id=16> Alright, too late, he’s yours. [<nv>laugh</nv>]

COLT transcript
<1> too late he’s M\INE#
<2> alright too late he’s Y/OURS#
<1> [@@@]

Transcript sample 2.

2.1.4. Pauses

The COLT system is very similar to the SGML tags used by Longman to encode pauses (, ...(5 secs)), except on one point: the brief and medium pauses. In the mark-up Longman uses, a comma/full stop is used if there is a brief/medium pause OR simply if the punctuation is syntactically appropriate (Crowdy 1991) - regardless of whether there is a pause at all. If *Too late, he’s mine* in transcript sample 2 is compared with the COLT version, we see that there is, in fact, no pause in this utterance. For a more detailed discussion of this, see Haslerud & Stenström (forthcoming).

Commas are not used in the COLT system. We do, however, for want of a better symbol, use full stops (.) to indicate brief pauses, but they are given a space on each side to distinguish them from punctuation marks. This is exemplified in the first line of the COLT transcript sample 3: *cor look at that . B\UM [now]#*, where the nucleus is highlighted by the short, preceding pause.

2.1.5. Laughter

A small, but significant difference between the two mark-up systems is the way in which laughter is illustrated. The Longman <nv>laugh</nv> is replaced by @ symbols in the COLT-version, one for each laughter syllable.

Longman transcript

<u who=1 id=9>   Cor! Look at that bum [now!]
<u who=2 id=10>   [<nv>laugh</nv>]

COLT transcript

<1>   cor look at that . B\UM [now]#
<2>   [@@@@@]

Transcript sample 3.

This way of representing laughter was originally the idea of Du Bois (Du Bois 1991, Du Bois et al. 1993), who pointed out that the @ symbols resembled the ‘smiley face’ icon. We choose to consider laughter part of the language, not merely a paralinguistic feature. Laughter may be a substitute for words and can function as a backchannel, marker of termination or transition, etc. Different ways of laughing are associated with personality, gender, etc. It is our intention to enable the users of the corpus to extract this information from the text. The number of laughter syllables is recorded because it provides information about the degree to which the speaker feels entertained. The @ symbol is easily reiterated in a minimum of space and stands out in the text, even when printed inside a word (Du Bois et al. 1993).

As appears in Table 1, <@> will be used to mark syllables spoken in a laughing voice as well, to replace the Longman tag <laughing> used for paralinguistic features. (See example in appendix A.)

2.2. Application of TEI mark-up

To facilitate international interchange and make it easy for the CD-ROM user to convert the COLT-tags into his/her preferred mark-up system, the texts will also be available with mark-up following the recommendations of the TEI (Text Encoding Initiative). Converting the COLT tags to conform to the TEI recommendations seems to be a fairly straightforward matter. As is demonstrated in Table 2, there is a one to one relationship for most tags. This means that most COLT tags have an equivalent TEI element defined in more or less the same way, covering the same instances.

TABLE 2

Note that many of the TEI elements listed in Table 2 may receive attributes other than those mentioned there (e.g. the attribute *who*, used to identify the vocalist), see example in transcript sample 4 and further Sperberg-McQueen and Burnard 1994.

2.2.1. Overlap

There is no indication in Table 2 of how the COLT notation for **overlapping speech** will be converted to follow the TEI guidelines. According to Sperberg-McQueen and Burnard (1994), there are several ways of doing this. We will use <anchor> elements with *id* and *synch* attributes, but without reference to absolute time, as demonstrated in transcript sample 4 below.

COLT transcript

<1> cor look at that . B\UM [now]#
<2> [@**@@@**@]

TEI-tagged transcript

<u who=1> <seg type=tone.unit>**cor look at that**
<pause type=brief> **b&f;um**<anchor id=P1>**now**
<anchor id=P2></seg></u>
<anchor synch=P1><vocal who=2 desc=laugh iterated=y>
<anchor synch=P2>

Transcript sample 4.

The spatial arrangement of aligned overlapping strings will disappear in the conversion to TEI, but this is really of no consequence, since the TEI mark-up is not intended to appear on the user‘s screen, but is merely an underlying standard for interchange.

2.2.2. Cases of no corresponding TEI element

Although most of the tags are unproblematic to convert, some problems can be foreseen. For some TEI elements we will have to create new attributes, such as for instance *unit* to <pause> in Table 2. Thus, COLT will not be 100% TEI compatible. Certain features do not have a suggested TEI element, such as **incomplete word** and **restart** (see Table 2).

There are, however, tools in the TEI system for defining new entities. To compensate for the lack of suggested elements corresponding to the COLT tonic marks, a set of new entities will be defined: &r; for rising tone (/), &rf; for rise fall (/\), and so on. In spite of the fact that the capitalization of the nuclear syllable will be lost in the transition from COLT to TEI, the nucleus will be easy to identify due to the entities printed within them.

In the marking of **laughter**, some information will inevitably be lost, however. So far, there is no way in which the number of laughter syllables can be recorded in the TEI system, since the iteration attribute can take only ‘yes’, ‘no’ or ‘unmarked’ as values.

3. Trends

Judging by the conversations that we have had the opportunity to study so far, there is no doubt that teenagers’ spoken interaction differs a great deal from that of adults. It is not just a matter of vocabulary, pronunciation, grammar and choice of topics but, perhaps first and foremost, overall interactive behaviour.

The extract below recorded in a Hackney school library is a good illustration. The speakers are <1> Catherine 15, <2> Sue 14, <3> Billy 15, and <4> a teacher. The pupils are expected to be working under the teacher’s supervision, but there are obviously distractions (‘id’ = utterance number):

<u who=1 id=9>	Cor! Look at that bum [now!]
<u who=2 id=10>	[<nv>laugh</nv>]
<u who=1 id=11>	You should [relax]
<u who=2 id=12>	[I] think you’re really
<u who=1 id=13>	Teacher! Teacher! Teacher!
<u who=2 id=14>	I think you’re really sexy Billy!
<u who=1 id=15>	Too late, he’s mine.
<u who=2 id=16>	Alright, too late, he’s yours.[<nv>laugh</nv>]
<u who=1 id=17>	[<nv>laugh</nv>]
<u who=2 id=18>	What are you going out with Billy?
<u who=1 id=19>	Yes I am.
<u who=2 id=20>	Oh, is she going out with you?
<u who=1 id=21>	Yes I am.

<u who=? id=22> [Are you?]
<u who=2 id=23> [No, say.]
<u who=3 id=24> I’m only joking Cath. I’ve gotta like Catherine now.
<u who=1 id=25> <nv>laugh</nv>
<u who=2 id=26> Bloody idiots!
<u who=1 id=27> <nv>laugh</nv> ... You’re so thicky!
<u who=2 id=28> I see, it’s screw me and leave me. [Aha.]
<u who=1 id=29> [Oh like] your seeing Trevor
<u who=2 id=30> You know
<u who=1 id=31> again?
<u who=? id=32> [Yeah.]
<u who=2 id=33> [I quite] understand.

<u who=? id=34> Where’s this bloody book!
<u who=1 id=35> Oh. Eh? Oh.
<u who=3 id=36> Urgh! Urgh!
<u who=3 id=37> <nv>laugh</nv>
<u who=1 id=38> Oh my God!
<u who=2 id=39> Er, er er ... er er

<u who=1 id=40> It was so funny, I had this weird dream the other night, you know. [I mean]
<u who=2 id=41> [If it’s about] Take That I don’t wanna hear.
<u who=1 id=42> Oh yeah. It was.
<u who=2 id=43> No, I don’t wanna hear it.
<u who=1 id=44> I got off [with <unclear>]
<u who=2 id=45> [Oh shut up!]
<u who=1 id=46> [and it was]
<u who=4 id=47> [<unclear>]
<u who=1 id=48> nice though. It’s erm, [it was at the, and he was]

<u who=4 id=49> [Come on! You’re in here] to work don’t [<unclear>]
<u who=2 id=50> [We are.]
<u who=4 id=51> I said no ... no eating!
<u who=2 id=52> Sorry miss. I’ll spit it out straight away. I swear to God.

<u who=4 id=53>   Do I have to say it twice?
<u who=1 id=54>   Yes.
<u who=2 id=55>   Well

Transcript sample 5.

3.1. Interactive behaviour

3.1.1. Turntaking

Turntaking is a tricky business. The most obvious conversational rule says that speakers should wait their turn and not butt in just anywhere. The speakers in the library extract do not pay too much attention to this rule; there are several instances of overlapping speech, four of which are clear cases of interruption.

First, Sue takes over before Catherine has finished (id 12); second, Catherine is trying to go on telling the other two about a dream, when Sue interrupts her (id 41); third, she interrupts her again, this time more rudely *(Oh shut up)*, apparently with no effect (id 45). Finally, the teacher's attempt to take the turn is without success (id 49). The remaining overlaps are instances of simultaneous speech, not interruptions.

Clearly, adult speakers interrupt each other as well, but they tend to do so less frequently, and they are definitely not supposed to tell each other to shut up. This would be perceived as extremely rude. Teenagers, on the other hand, do not seem to mind (unless they are told by a grown-up).

3.1.2. Reactions

Unlike adults, teenagers can also use abusives without offending each other. Examples in the library extract are *Bloody idiots* in line 18 and You're *so thicky* in line 19. It is as if abusives were used just for the fun of it. This is even more apparent in the lunch-break extract below, involving three 14- and 15-year-olds who are sitting in a London park:

<u who=4 id 70>	They taste like brussel sprouts [<unclear>]
<u who=? id 71>	[<nv>laugh</nv>]
<u who=1 id 72>	Brussel sprouts?
<u who=4 id 73>	Fuck off!
<u who=? id 74>	<nv>laugh</nv>
<u who=4 id 75>	[Get out of here you tart.]
<u who=? id 76>	[Yeah that hurt didn’t it?]
<u who=1 id 77>	<nv>laugh</nv> Urgh they’re not very nice, no.

Transcript sample 6.

3.2. Choice of topics

The brief library extract contains only one real topic, the question of whether Billy and Catherine are going out together. It starts off, it seems, with Catherine‘s remark *Cor! Look at that bum now!*. A second topic is entered upon but never finished; Sue’s attempt to tell the others about a dream she had is firmly rejected.

Overall, it appears that the topics teenagers prefer to talk about are related to sex, drink and improper behaviour. The lunch-break interaction contains the following exchange:

<u who=1 id=25>	You know Karen? Karen would do something like that because apparently she‘s always walking down the streets and like taking her top off and showing her tits to everyone and sort of like pulling her trousers down.
<u who=2 id=26>	I know she goes to the <unclear> with the door open when all the other kids <unclear>

Transcript sample 7.

3.3. Taboo words

There is an immediate link between taboo topics and the use of taboo words. No wonder there are so many of them in teenage talk. In the library extract we find the exclaims *cor* and my *God,* the intensifier *bloody,* and the taboo slang word *screw* (‘to have sex with’).

The majority of the taboo words in the data were used as vulgar variants of ordinary accepted words, eg *arsehole, crap, cunt, dick, fuck, piss, shit* and *turd*. Some of these were used as abusives *(No, like I hate them. They're cunts.),* and some were used for swearing *(It's so bloody heavy., What a shit lighter!, Oh fuck!*) in addition to their intensifying or exclaiming function.

Unlike adults, especially the youngest teenagers scream and shout instead of using ordinary words to show approval, disapproval, surprise etc. In the library talk we have *Urgh! Urgh* ! (id 36). Other common 'sounds' are *aah, aargh, eeeh, oi, ouch, uhuu, uuuhu, wooooo* and *wraaa*. It remains to be seen whether the younger teenagers use this kind of nonverbal item as a substitute for swearwords.

3.4. Other tendencies

3.4.1. Lexis

A comparison with the London-Lund Corpus of Spoken English (LLC)[4]. shows that new uses of *go* and *like* have crept into British English. In COLT but not in LLC, *go* is used as a reporting verb equivalent to say (*Excuse me could I have a glass of water she* ***goes***)**,** while *like* is used for hedging (*Yeah but she might sort of* ***like*** *expose herself*), as a conjunction, equivalent to *as if* (*I mean it's* ***like*** *you can't help feeling sorry for Aron*), and to indicate approximation (*You could do* ***like*** *four cartwheels*).

This change is obviously due to influence from American English. Another manifestation of American influence in COLT is the expression *yeah man*.

3.4.2. Simplified pronunciation

The Longman transcribers found teenage language much more difficult to transcribe than adult language, not only because they did not always understand what the teenagers were talking about but also because of their often indistinct pronunciation. Some youngsters, for instance, have a tendency to swallow entire syllables. Others replace the voiceless stops p, t, k with glottal stop, as in Cockney pronunciation.

[4] LLC is composed of adult conversations collected mainly in the 1960s.

The reduced forms *dunno*, *gonna*, *gotta* and *wanna* are more common than *going to*, *got to* and *want to*, and there are numerous instances of *innit* for *isn't it*. Stretching it a bit further, the simplified pronunciation also seems to have an effect on the syntax, insofar as certain clause elements are just left out, especially the subject (*just couldn't resist it*) and auxiliaries (*What she say? What you on about?*).

4. Final remarks

The work on the corpus is proceeding according to plan. The COLT markup system has been finalised, and we are ready to enter the next phase in the process: the conversion of the Longman transcriptions to COLT transcriptions. What we foresee at this point is that converting the Longman mark-up to COLT mark-up will involve considerably more work than the next step from COLT to TEI mark-up.

The trends reported on here are based on only a fraction of the corpus. They will of course be studied in the entire corpus and in more detail. There is no way we can make any generalizations at this point, but we are convinced that this new exciting material will yield a tremendous amount of interesting findings.

References

Crowdy, S. (1991): 'Spoken Corpus Design and Transcription'.

Crystal, D. (1975): *The English Tone of Voice. Essays on intonation, prosody and paralanguage*. London: Edward Arnold.

Du Bois, J. W. (1991): 'Transcription Design Principles for Spoken Discourse Research'. In: *Pragmatics: Quarterly Publication of the International Pragmatics Association* 1(1), 71-106.

Edwards, J.A. (1993): 'Principles and Contrasting Systems of Discourse Transcription'. In J.A. Edwards and M.D. Lampert (eds) *Talking Data*. Hillsdale, NJ: Lawrence Earlbaum.

Haslerud, V. and A-B. Stenström. Forthcoming: 'The Bergen Corpus of London Teenager Language (COLT)'. In G. Leech (ed) *Spoken English on Computer*. London: Longman.

Milroy, L. (1987): *Observing & Analysing Natural Language*. Oxford: Basil Blackwell.

Sperberg-McQueen, C.M. and L. Burnard (1994): *Guidelines for Text Encoding and Interchange*. Oxford: Oxford University Computing Services.

Appendix A

COLT transcript sample

<1> oh H_I Billy#
<3> you alR/IGHT Cath#
<1> what are you doing here you C\UNT#
<2> <@>don't call my boyfriend a C/UNT</>#
<1> no it's my D\AD you're talking about#
<2> @@ . [@@]
<3> [S\UE]#
<1> cor look at that . B\UM [now]#
<2> [@**@@@**@]
<1> sh [quiet +sh +sh +sh] teacher +teacher +teacher==
<2> [I think you're R\EAlly s=]#
<2> *I think you're R\EAlly# S\/Exy Billy#
<1> too late he's M\INE#
<2> alright too late he's Y/OURS#
<1> [**@**@@]
<2> [what are] you going \OUT with Billy#
<1> yes I \AM#
<2> oh is she going \OUT with you#
<1> yes I \AM#
<?> [\ARe you]#
<2> [say S\AY]#
<3> I'm \ONly joking Cath# ... ah I've gotta like C\ATHerine now#
<1> @@@@@@ [@@]
<2> [<nv> snivel</>]
<3> {<2 sylls> too TH_ICky girls}#
<1> {<nv>+snivel .. [+snivel . +snivel</>]}
<2> [@@@@@@@@@@ .. you] dirty D\ICK#
<1> I see . it's screw me and L/EAVe me# W/ELL# . you KN/OW# . I underST/AND#