

# USERS' MANUAL

to accompany  
The Bergen Corpus of London Teenage Language (COLT)

by

*Anna-Brita Stenström, Gisle Andersen,  
Kristine Hasund, Kristine Monstad and Hanne Aas*



Department of English, University of Bergen, Norway



## **Preface**

Thanks to a grant from the Norwegian Research Council we were able to collect the The Bergen Corpus of London Teenage Language (COLT) in the spring of 1993. Since then COLT has received financial support from the Norwegian Academy of Science, the Meltzer foundation, and the Faculty of Arts at the University of Bergen.

The project was initiated by Anna-Brita Stenström in collaboration with Leiv Egil Breivik and was carried through with the help of postgraduate students employed as research assistants, notably Gisle Andersen, Vibecke Haslerud, Kristine Hasund, Migle Miliauskaite, Kristine Monstad, Ingrida Strazdaite, Nina Sørli, Ingrid Thompson and Hanne Aas. In addition, Lars Johannessen was engaged for the preparation of the material for text-to-sound conversion, which was completed by Tony Robinson at SoftSound, St Albans.

We are extremely grateful to the Department of Education in London for suggesting suitable London schools for collecting the material; to the Longman Group, London, not only for letting us use the method of corpus collection that was used for the collection of the British National Corpus but also for carrying out the orthographic transcription; and finally to the researchers at Lancaster University, in particular Elizabeth Eyes, for doing the word class tagging.

The project could hardly have been carried through without the assistance of Knut Hofland at The Norwegian Computing Centre for the Humanities and, at a later stage, Manfred Thaller at the Centre for Humanities Information Technologies Research, both at the University of Bergen.

Finally, our heartiest thanks go to the recruits. Had it not been for their willingness to assist by recording the conversations, COLT would of course never had got off the ground.

## Contents

Preface	ii
<b>1 BACKGROUND</b>	<b>1</b>
1.1 AIM	1
1.2 CORPUS COMPILATION	1
1.3 FROM TAPE RECORDINGS TO CD-ROM	2
<b>2 THE COLT SPEAKERS: SOCIAL BACKGROUND AND CONVERSATIONAL SETTINGS</b>	<b>2</b>
2.1 SPEAKER-SPECIFIC INFORMATION	3
2.1.1 Age and gender	3
2.1.2 Social class	4
2.1.3 Names and anonymity	8
2.2 CONVERSATION-SPECIFIC INFORMATION	11
2.2.1 The London boroughs	11
2.2.2 Conversational settings	11
<b>3 HEADER INFORMATION, MARK-UP CONVENTIONS AND COMPUTER SEARCHING</b>	<b>12</b>
3.1 HEADER INFORMATION	12
3.2 MARK-UP AND INDEXING	14
3.2.1 Paralinguistic features and non-verbal sounds	14
3.3 THE PROSODIC VERSION	15
3.4 THE TAGGED VERSION	15
3.5 COMPUTER SEARCHING WITH TACT	15
3.5.1 KWIC - Key Words In Context	16
3.5.2 Variable Context Display	16
3.5.3 Distribution and Normalised Distribution	16
3.5.4 Word List	17
3.5.5 Regular expressions	18
3.5.6 Selecting certain texts and certain speakers	19
<b>4 COLT-BASED RESEARCH</b>	<b>19</b>
Appendix 1 COLT-based publications (as of December 1998)	21
Appendix 2 Survey of COLT text files	24
Appendix 3 Personal data sheet	30
Appendix 4 Personal data survey	31
Appendix 5 Paralinguistic features in COLT	32
Appendix 6 Non-verbal sounds in COLT	33
Appendix 7 COLT tagset (CLAWS 6)	35



## **1 Background**

As a prelude to the project, we organised a seminar in November 1992 with experienced corpus linguists as invited speakers: Jan Aarts (Nijmegen), Steve Crowdy (Cambridge), Sidney Greenbaum (London), Stig Johansson (Oslo), Jan Svartvik (Lund), and John Sinclair (Birmingham).

The collection of the material took place in 1993. The reason for compiling the corpus was simply that we realised that the fact that teenage language was largely unexplored could be remedied by the collection of a reasonably large corpus of teenage talk.

### **1.1 Aim**

The aim of the project has been to create a corpus of British English teenage talk and make it available for research, first on the internet, next as an orthographically and prosodically transcribed CD-ROM version, and finally as a CD-ROM version with both text and sound.

We are convinced that the study of spontaneous teenage talk will give us new insights into language development and language change, not least from the point of view of grammaticalisation. Much of what happens in teenage talk is likely to have an everlasting effect on adult speech and the language in general. The reason for restricting the corpus collection to London was the assumption that new trends predominate among teenagers in the capital, from where it can be expected to spread to the rest of the country, and even abroad.

### **1.2 Corpus compilation**

The techniques used for collecting COLT were modelled on the principles adopted for the collection of the British National Corpus (BNC), although with a much smaller corpus in mind. Our aim was to record half a million words in a limited area of the UK, while the aim of the BNC scheme was to collect ten million words, including both old and young speakers, in the entire country. Advised by the Department of Education, we contacted schools in five different school districts in London, trying to recruit students who were willing to carry a walkman and a lapel microphone for a few days and record all the conversations they took part in, preferably with friends of the same age who were not supposed to be aware of the recording. The recruits were also equipped with a logbook and instructed to write down information about the co-speaker(s) and the setting. Vibecke Haslerud, the first research assistant at the COLT project, administered the sampling of the Inner London recordings by

handing out equipment, instructing the recruits, collecting tapes and equipment and marking the tapes. The whole procedure, including the recording, took roughly three weeks and took place in April/March 1993. In September, this material, which constitutes the bulk of the corpus, was supplemented by recordings collected in a school in the Outer London area.

### **1.3 From tape recordings to CD-ROM**

The recordings were made by 31 volunteering 13-17 year old boys and girls from five socially different school boroughs, so-called 'recruits' equipped with a Sony Walkman, a lapel microphone and a log book.

The entire material of roughly half a million words was orthographically transcribed by trained transcribers employed by the Longman Group for transcribing The British National Corpus (BNC). A copy of this version of COLT was incorporated in the BNC. At the Bergen end, the orthographically transcribed material was subsequently submitted to careful editing, which involved correcting misinterpreted talk, reducing the number of <unclear> passages and adding untranscribed talk. The edited version was then tagged for word classes in the same way as the BNC by a research team at Lancaster university.

Since we aimed at a more spokenlike version of COLT, the bulk of the material has been subjected to a simplified prosodic analysis, which involved replacement of the orthographic version with 'sentences' beginning with a capital letter and ending with a punctuation mark, by marking pauses, tone unit boundaries and nuclear tone. Finally, the orthographic version of COLT was digitised in Bergen as a preparation for text-sound alignment, which was carried out by SoftSound, St Albans.

## **2 The COLT speakers: social background and conversational settings**

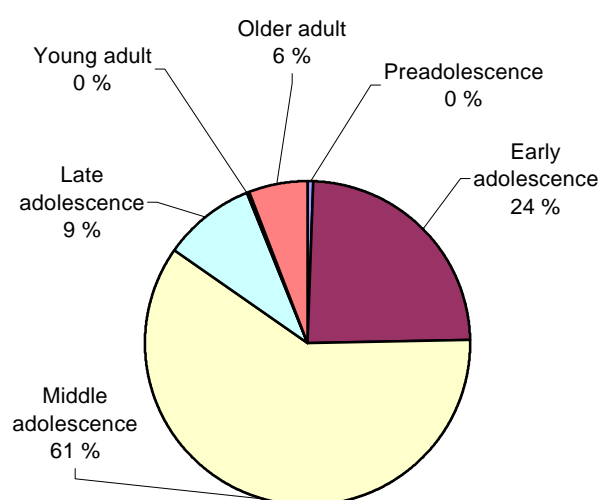
The current section contains a survey of the available background information concerning the speakers and conversations in COLT. This information is based on the the logbooks that the recruits where requested to use. For convenience, we distinguish between speaker-specific information (speakers' age, gender, social class etc) and conversation-specific information (location and setting).

For each conversation, the speaker/setting information is given in the text header. The significance of the various header codes is given in section 3.1.

## 2.1 Speaker-specific information

### 2.1.1 Age and gender

COLT is specifically designed to represent the language of a restricted age group in London, namely teenagers. Nevertheless, the speakers that are actually classified with respect to age range from 1 to 59 years old. This is due to the occasional presence of some of the recruits' younger and older family members and to the presence of teachers in some of the conversations. For most research purposes, it will probably be convenient to bundle together some of the occurring values of the age variable, as some age groups, eg two-year-olds (for natural reasons) are represented with very low word counts. We suggest a grouping into six different age groups: preadolescence (0-9), early adolescence (10-13), middle adolescence (14-16), late adolescence (17-19), young adults (20-29) and older adults (30+). The distribution of text across the various age groups can be visualised as follows:



*Figure 1: Distribution of COLT text material in the various age groups*

Only three of these age groups, early, middle and late adolescence, can be said to represent the 'core' of COLT-informants and the target group of the project. 85 per cent of the corpus material comes from speakers within these age groups. The other age groups are represented to varying degrees. The preadolescent group accounts for a very small amount of text (1,855 words, 0.46 %), and the same goes for the young adult group (1,138 words, 0.28 %). Hence, whatever linguistic features are found within these age group must be interpreted with caution, due to their low overall rate of contribution. The older adult group mainly comprises

the recruits' parents and, to a lesser degree, their teachers. This group contributed about six per cent of the corpus material, which amounts to 23,055 words.

As regards gender, girls and boys contributed roughly the same amount of text: the male speakers about 51.8 per cent (230,616 words) and the female speakers 48.2 per cent (214,215 words).

### **2.1.2 Social class**

The calculation of a social class index has been a matter of some controversy within the COLT research team. The eventual classification, to be presented below, divides the recruits into three different social groups, and is a compromise between two earlier versions, Andersen (1995) and Hasund (1996). As the information that constitutes the basis for the calculation of social class is somewhat scarce and to some extent also unreliable (a thirteen-year-old will not always be able to name the exact occupations of his/her parents), we find it reasonable to operate with a less fine-grained scale than the one that was originally applied (Andersen 1995). Originally we divided the recruits into five different social groups, but we have now opted for only three groups, conveniently labelled 'high', 'middle' and 'low'.

We have based the social class index on information that the 31 COLT recruits provided by filling out a personal data sheet (cf Appendix 3). Three pieces of information from the data sheet are used as indicators of social class: residential area, parents' occupation and whether the parents are employed or not. Residential area and parents' occupation constitute social indices in their own right, while the employed/unemployed distinction is used as a slight modification of the occupational index. As this information was provided for no other speakers than the recruits themselves, only the recruits and their families are classified with respect to social class.

As is well known, there are major differences in social standards between the various boroughs of London. Area of residence is a significant constituent of a person's social background, and it is of prime importance that differences in area of residence are reflected in a description of the social profile of the recruits. The COLT material involves recruits from ten different residential areas. The Inner London boroughs are represented by recruits from Camden, Hackney, Islington, Tower Hamlets and Westminster. The Outer London boroughs are Barnet, Brent, Enfield and Richmond upon Thames. The final area represented in the corpus is Hertfordshire in the Greater London Metropolitan Area. Each of the areas was assigned a borough index on a scale ranging from 1 to 5, which reflects certain social class



features of the area. The index is a complex one, calculated by means of figures from the *Key statistics for local authorities, Great Britain* (Office of Population Censuses and Surveys:1994). Four components were used in the calculation of the borough index:

- Component 1* The percentage of the borough's population who are economically active in Social classes I-II.
- Component 2* The percentage of the borough's population who are economically active in Social classes IV-V.
- Component 3* The percentage of the borough's families comprising lone parents with dependent child(ren).
- Component 4* The percentage of the borough's population who live in a house rented from a local authority.

The effect of components 1 and 2 on the borough index is obvious. A high percentage of the population economically active in the two highest social classes, I and II, gives a high component score for the borough; a high percentage economically active in the two lowest social classes, IV and V, gives a low score. The last two components are perhaps more controversial. If an area has a high percentage of families consisting of lone parents with dependent children (single parent families), it will be perceived by most people as a low-status area. Single parents, and single mothers in particular, are in many ways financially unprivileged in today's Britain, and this counts negatively in terms of socioeconomic status. Therefore, a high percentage of single parent families gives a low component score for the borough. And finally, if a high percentage of the population live in houses rented from a local authority, such as council houses, this will yield a low score in the calculation of the borough index.

The four factors in the borough index were weighted equally, and an approximation of the average score constitutes the borough index. For comparison, the figures for Greater London and Britain are included in the calculation. The following Borough indices are attributed to the ten areas represented in the corpus (The highest score yields Borough index 1.):

*Table 1: COLT Borough index*

<b>BOROUGH/AREA</b>	<b>COMP 1</b>	<b>COMP 2</b>	<b>COMP 3</b>	<b>COMP 4</b>	<b>AVERAGE</b>	<b>BOROUGH INDEX</b>
Richmond	1	1	1	1	1	1
Barnet	2	1	2	1	1,5	2
Hertfordshire	2	3	1	2	2	2
Westminster	2	2	3	2	2,25	2
Camden	2	2	4	3	2,75	3

Enfield	4	3	2	2	2,75	3
Brent	4	3	4	2	3,25	3
Islington	3	4	5	5	4,25	4
Hackney	4	4	5	5	4,5	5
Tower Hamlets	5	5	5	5	5	5
Greater London	3	3	3	2	2,75	3
Britain	4	4	2	2	3	3

The COLT recruits represent a wide range of different boroughs in terms of social class. Indeed, as 1 is the highest score and 5 the lowest, all the possible borough categories are represented. There is, moreover, a fair degree of consistency within the boroughs with respect to the four components that the borough index is based on. Two boroughs, the very top and very bottom ones (Richmond and Tower Hamlets), have the same component scores throughout. No borough has a variation in component scores greater than 2 points.

The information regarding parents' occupation is treated in accordance with *The Standard Occupational Classification* (Office of Population Censuses and Surveys (OPCS): 1991). Each parent has been classified by the standard categories I-V, except for a single, unclassifiable recruit who did not provide any information regarding parents' occupation. The OPCS classification gives a detailed list of how to categorise each single occupation, and each profession falls into one of the following broad categories, known as 'social classes':

- I Professional etc occupations
- II Managerial and technical occupations
- III Skilled occupations
- IV Partly skilled occupations
- V Unskilled occupations (ibid:12)

Some recruits reported that the parents neither worked nor had a profession. They were given the same occupational score as those belonging to class V. Since recruits who gave the answer 'none' as to parents' profession consistently answered 'no' to the question about parents' employment, it seemed plausible to categorise them as members of class V.

There is a lot of controversy connected with the issue of how to weigh parents' occupational scores in social class index calculation. Commonly, sociolinguists use only the father's occupation as indicator of social class. Traditionally, the male adult of a family has been viewed as the breadwinner, and his occupational score has determined the social class of the rest of the family. More recently, however, the mother's occupation is also being taken into consideration, due to the increase in the number of families with both parents working, as

well as a gradual process of levelling of the sex roles. On this account, we found it natural to include the mother's occupation in the calculation of the social class index, and the two occupations have been weighed equally.

For the sake of simplicity, the scale of socioeconomic groups shown above has been reversed, so that, in the calculation of occupational index, the highest occupational category, 1, is assigned an occupational score of 5 points, while the lowest category, V, yields occupational score 1, etc. We calculated the First occupational score as the average of father's and mother's occupational scores in cases where information on both mother and father was available. In cases where only one parent is mentioned on the personal data sheet, this parent counts as breadwinner, and his/her occupational score counts as the recruit's First occupational score.

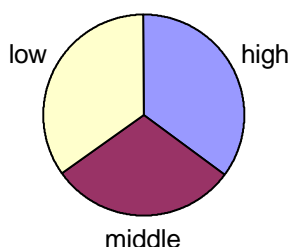
In most sociolinguistic studies, the factor of unemployment is ignored in the calculation of a social class index. In our opinion, this is a major drawback, because unemployment certainly has a severe effect on people's economic situation and thus on the socioeconomic status of the family. The number of people who are long-term unemployed has reached unacceptable levels in some parts of Britain, particularly in urban areas such as London. A social class index applied in a sociolinguistic description of an urban dialect ought to reflect this fact. We therefore chose to include the employed/unemployed distinction by including a 'non-working factor' in the calculation of the social index. Any recruit who has answered 'no' to the question 'Currently employed?' for one or both parents is assigned a non-working factor. The non-working factor was calculated as a relative figure of 30 per cent of the First occupational score and was subtracted from it. This yields a Second occupational score which reflects the financial situation of a family who must support itself on only one salary.

Of the two social factors presented thus far, we considered parents' occupation the most important one, because it presumably is a better indicator of social class than area of residence. Information regarding parents' occupation is specific to the single recruit, while area of residence is shared with other recruits. No borough of London is consistent as regards the social class of its inhabitants. We therefore weighted the two indicators differently by multiplying the occupational score by 2 before calculating the Recruit's total score. The weighting was done by applying the following formula for the calculation of the Recruit's total score:

$$\text{Recruit's total score} = \frac{2 \times (\text{Second occupational score}) + \text{Borough score}}{3}$$

The figures for Recruit's total score were then approximated and the scale was once again reversed. The result is the recruit's Social group index, which forms a scale from 1 to 3, where 1 indicates the highest social group.

As only recruits and their families are classified, only about 50 per cent of the corpus material can be assigned a social group value. The material that has been classified is evenly distributed across the three social groups:



*Figure 2: Distribution of COLT text material in the various social groups*

The classification of the individual recruits is given in the Personal data survey in Appendix 4.

### **2.1.3 Names and anonymity**

In corpus building, the issue of speakers' anonymity is an important one. In the current section we outline our policy for protection of participants' identity in COLT, taking legal, ethical, sociolinguistic, and computational factors into account.

In the transcripts, last names, addresses and telephone numbers have been deleted, while first names are authentic and have not been replaced by fictitious names. The final CD-ROM version of the corpus will include a digitised version of the audio recordings, where last names, addresses and telephone numbers are erased. No names have been erased from the original tapes, which are kept at the University of Bergen and are not generally available. We hold that first names ought not to be changed in the COLT transcripts, as they carry social and ethnic information which it is generally difficult to match with fictitious names. The speakers

represent a broad range of socioeconomic backgrounds and ethnic groups, which is reflected, to a certain extent, in their first names.

In the invitation to take part in the research project, the following promise was given to the COLT recruits: 'You and the people you have recorded are guaranteed full anonymity'. Lengthy discussions within the research team of what was implied in the term 'full anonymity', resulted in an agreement to delete all surnames and addresses in the transcription, but leave all first names unchanged. Having consulted the legal advisers at the Norwegian Social Science Data Services (NSD), who examined COLT from the point of view of the EU directive of 1995, where EU members are instructed to protect the fundamental rights to privacy of persons in connection with the processing of personal information, we were informed that COLT does not contain information that can be directly or indirectly traced back to an individual. This conclusion was reached on the basis of the following conditions:

- (a) The material consists of audiotaped recordings of arbitrary conversations between the COLT recruits and their families and friends.
- (b) The tapes are not labelled with any form of identification of the speakers.
- (c) The information that appears in the conversations and on the personal data sheet filled out by each recruit cannot be traced back to a specific participant, neither directly nor indirectly.
- (d) The transcribed material, which is electronically registered, does not contain any information that can be traced back to an individual respondent.
- (e) There exists no list of complete names that can be used to identify the speakers.

As COLT contains no systematic register of personal information, NSD's conclusion was that first names need not be changed in order for the personal identity of the informants to be regarded as fully anonymous. In other words, no promise was broken by keeping first names as they were. We also consulted expertise at the Faculty of Law, University of Oslo, who supported NSD's conclusion that the privacy of the COLT speakers can be regarded as sufficiently protected.

Some corpus builders use fictitious names with a similar stress pattern and number of syllables as the real name. We found that, although such a policy is commendable, it could not be successfully implemented on COLT. In the first place, we do not believe it is possible to fulfill all the criteria in a sufficient number of cases, not even with straightforward English names. Considering, then, that there is a large number of ethnic minority speakers in COLT, the task becomes even more demanding. Take, for instance, the names Chung Chew, Rashid, Mikesh, Surita, Miraq, Obina, Siobhan, Tan, and Waki. A replacement of these names that

would take phonological, social, ethnic and other aspects into account with a satisfactory result seems an impossible task.

A number of names create problems for automatic substitution, because replacement would make them incomprehensible in their context. This amounts, for instance, to names of celebrities such as Anthony Hopkins, Jodie Foster, Michael Jackson, David Copperfield. Similarly, names should not be replaced by fictitious ones when they are used to designate inanimate objects, fictional persons, abstract entities or place names, such as the computer games Adam, Guy, and Joe & Mack; and fictional characters such as Bertie Wooster and James Bond. Frequently, the participants in a conversation discuss names, and, in such conversations, there may be phonetic restrictions on the names that can be used as replacements, for instance, if they are used to create puns or rhymes:

- [1] Danny: Have you got a bit of a lisp? a wisp?  
James: I dunno I never used to, it's just erm some words I say.  
Danny: You can't pronounce [your]  
James: [No] it's not that bad, but it's [<unclear>]  
Danny: [Rough]  
James: I never used to.  
Danny: Oh.  
James: It's like on things like ...(3) Warren.  
Danny: Say Warren properly. He goes <??>Wawen</>  
James: And Darren.  
Danny: <??>Dawen and Wawen Wodgers</>  
James: Well apart from <unclear>  
Danny: <??>Wawen Wodgers</>  
? I say Darren can't say Warren. He's gotta lisp.
- [2] Dan: It's really weird my mu= erm one of my mum's friends who's a bit of goer, she's got herself a new bloke called Jim yeah and he looks exactly like Jimmi Hendrix and he plays the guitar.  
Cassie: <laughing>Wicked</>.
- [3] Danny: I haven't got them. Oh man did you see Reg peg it after me? I shouted out, Oi Fadge. <nv>laugh</nv> Fat chopper radge. <unclear> sawdust.  
Edward: Ee god I wonder how many times that's been said.

As the examples suggest, replacement is problematic in cases where it would ruin the punch line of a joke or destroy the effect of creative playing with sounds.

## 2.2 Conversation-specific information

In the COLT text files, the headers contain not only information about the individual speakers, but also more general features of the conversation, notably geographical location in London and setting.

### 2.2.1 The London boroughs

The COLT project involves five different schools located in five different school boroughs. The school boroughs represented are the Inner London boroughs Hackney, Tower Hamlets and Camden, the Outer London borough Barnet, and Hertfordshire, represented by a private boarding school. Hertfordshire is located within the London Metropolitan area, as defined in the *Key statistics for local authorities, Great Britain* (Office of Population Censuses and Surveys:1994). The borough classification incorporates information on borough of residence and school borough of the recruits. The residential boroughs include Barnet, Brent, Camden, Enfield, Hackney, Hertfordshire, Islington, Richmond, Tower Hamlets and Westminster. In most cases borough of residence is identical with school borough, but some recruits attend school in a borough different from their residential borough. This applies in particular to the Hertfordshire boarding school pupils. Since teenagers are generally assumed to adapt to peer group norms of linguistic behaviour, and are more likely to identify with classmates than their parents, the school borough classification appears to yield a more satisfactory and reliable indicator of group membership. The classification of each recruit with respect to school borough and residential borough is given in the Personal data survey in Appendix 4.

### 2.2.2 Conversational settings

Most of the COLT conversations take place in or near the school or the home of the recruit. Table 2 contains a complete survey of the settings that occur, and the list is identical with the searchable codes for this parameter (cf section 3.5.6):

*Table 2: Settings of the COLT conversations*

apartment	boarding_school_house	bus
car	Church_Street	classroom
class_room	entrance_to_home_flats	flat
games_hall	girls'_changing_room	gym_room_-school_playground_- _classroom
home	home/friend's_house	home_of_speaker_11
house	house_of_speaker_3	in_a_park
in_the_house_of_speaker_3	music_room	outside
outside_home	outside_pub	park

Peter's_house	playground	pub
respondent's_room_at_boarding_school	restaurant	school
school_classroom	school_dining_room	school_sixth_form_common_room
school_dinner_hall	school_form_room	school_library
school_outside	school_playground	
school_study	science_laboratory	shop
street	street_-_bus_-_classroom	study_room_at_boarding_school
walking_to_shop	way_to_school	

Although a wide variety of settings occur, settings in connection with school and home are the most common and account for about 48 per cent and 32 per cent of the material, respectively.

### 3 Header information, mark-up conventions and computer searching

#### 3.1 Header information

The speaker-specific and conversation-specific information described above is listed in the header of each of the COLT text files. The header information is identical in the orthographic and prosodic versions of the corpus, while the word class tagged version contains a slightly different set of codes. The following survey illustrates the header information given in the text files:

<b>&lt;REG&gt;</b>	<b>S</b>	<i>region: south</i>
<b>% reference number // title</b>		
<b>&lt;REF&gt;</b>	<b>B140402</b>	<i>number of text/file</i>
<b>&lt;TIT&gt;</b>	<b>?</b>	<i>title</i>
<b>% date // time</b>		
<b>&lt;DAT&gt;</b>	<b>?</b>	<i>date</i>
<b>&lt;TIM&gt;</b>	<b>?</b>	<i>time</i>
<b>% recording // input device</b>		
<b>&lt;DEV&gt;</b>	<b>wlk</b>	<i>walkman</i>
<b>% duration of conversation</b>		
<b>&lt;DUR&gt;</b>	<b>6.23</b>	<i>minutes.seconds in decimal code</i>
<b>&lt;WC&gt;</b>	<b>1324</b>	<i>word count</i>
<b>% details of conversation // locality - county &amp; town</b>		
<b>&lt;CTY&gt;</b>	<b>Greater London</b>	<i>county</i>
<b>&lt;LOC&gt;</b>	<b>Barnet</b>	<i>borough where conversation takes place</i>
<b>% setting // activity</b>		
<b>&lt;SET&gt;</b>	<b>classroom</b>	<i>setting</i>
<b>&lt;ACT&gt;</b>	<b>art lesson</b>	<i>activity</i>
<b>% type of discourse // superfield // subject // individual sub</b>		
<b>&lt;TYP&gt;</b>	<b>conversation</b>	<i>type of discourse</i>
<b>&lt;SUP&gt;</b>	<b>?</b>	<i>superfield</i>
<b>&lt;SUB&gt;</b>	<b>?</b>	<i>subject</i>
<b>&lt;IND&gt;</b>	<b>?</b>	<i>individual subject</i>
<b>% spontaneity factor // audience</b>		
<b>&lt;SPN&gt;</b>	<b>3</b>	<i>3 indicates highest spontaneity</i>
<b>&lt;AUD&gt;</b>	<b>4+</b>	<i>number of participants and/or people present</i>
<b>&lt;BOB&gt;</b>	<b>?</b>	<i>miscellaneous information regarding the conversation</i>



```

% details of participant // identifier // name // gender
<IDT1> 1 I is always the recruit carrying the walkman
<NAM1> Alex name of the recruit
<GEN1> m gender
% age // first language // dialect
<AGE1> 14 age
<LAN1> BrE language
<DIA1> London dialect
% occupation // education // social group
<OCC1> student occupation
<EDU1> still studying education
<SOC1> 1 social group
% relationship to respondent // other relationships
<REL1> respondent indicates that this speaker is the recruit (respondent)
<OTH1> friend2 friend8 pupil2 this speaker is a friend of speakers 2 and 8, and a pupil of speaker 12
<BOB1> ? miscellaneous information regarding the speaker
% details of participant // identifier // name // gender information about the next speaker
<IDT2> 2
<NAM2> Marc
<GEN2> m
% age // first language // dialect
<AGE2> 14
<LAN2> BrE
<DIA2> London
% occupation // education // social group
<OCC2> student
<EDU2> still studying
<SOC2> ?
% relationship to respondent // other relationships
<REL2> friend speaker 2 is a friend of the recruit
<OTH2> friend3 friend8 friend9 pupil12 - and a friend of speakers 3, 8, 9, also a pupil of speaker 12
<BOB2> ?
% details of participant // identifier // name // gender etc
<IDT8> 8
<NAM8> Daniel
<GEN8> m
% age // first language // dialect
<AGE8> 13
<LAN8> BrE
<DIA8> London
% occupation // education // social group
<OCC8> student
<EDU8> still studying
<SOC8> ?
% relationship to respondent // other relationships
<REL8> friend
<OTH8> friend2 pupil12
<BOB8> ?
% details of participant // identifier // name // gender
<IDT12> 12
<NAM12> ?
<GEN12> m
% age // first language // dialect
<AGE12> 35
<LAN12> BrE
<DIA12> London
% occupation // education // social group
<OCC12> teacher
<EDU12> ?
<SOC12> ?

```

% relationship to respondent // other relationships	
<REL12>	teacher
<OTH12>	teacher2 teacher8
<BOB12>	?

### 3.2 Mark-up and indexing

Table 3 describes the transcription conventions used in the orthographic corpus and the tagged corpus. Worth noting is the dual use of the full stop (.) and the comma (,), which can indicate either as a sentence boundary or a pause or both.

Table 3: COLT mark-up

ORTHOGRAPHIC	TAGGED	COMMENTS
, . ? !	see CLAWS6 tagset	sentence-like boundaries; also continuing, terminating, questioning, and exclamatory intonation
CAPS	CAPS	sentence beginnings
=	<trunc> text </trunc>	incomplete word
,	see CLAWS6 tagset	brief pause
.	see CLAWS6 tagset	medium pause
...	<pause dur=3>	long pause
... (5)	<pause dur=5>	pause 5 seconds
<nv> laugh </nv>	<vocal desc=laugh>	non-verbal sound
<name>	&name;&NP1;	name
<address>	&address;&NN1;	address
<unclear>	<unclear>	unintelligible speech
<unclear> (5)	<unclear dur=5>	unintelligible speech 5 seconds
[text]	<ptr=P001> text <ptr=P002>	single overlap
[{}]	non-existent	double overlap
<singing> text </>	<shift new=singing> text <shift>	paralinguistic features
(hairdryer on)	<event desc="hairdryer on">	contextual comment
<??> text </>	<uncertain> text </uncertain>	uncertain transcription
(sic)	<sic> text </sic>	awkward pronunciation

#### 3.2.1 Paralinguistic features and non-verbal sounds

A large number of original and innovative paralinguistic features are applied in COLT, examples being <mimicking a monkey>, <mimicking a yobbo hooligan>, <mimicking Yorkshire accent> and <teasing>. The markup convention for these is <paralinguistic feature> text </> in the orthographic and prosodic version, and in the tagged version <shift new=paralinguistic feature> text <shift>. A complete list of the paralinguistic features found in COLT is given in Appendix 5.

In addition, as a reflection of the humoristic and playful nature of many of the COLT conversations, COLT contains a relatively large number of non-verbal sounds that are,

presumably, less prevalent in corpora of adult English. These include <nv>blowing air through lips</nv>, <nv>mimicking gorilla noises</nv>, <nv>mimicking shaving noise</nv>, and so on. These are marked with the tag <nv> in the orthographic corpus, and the tag <vocal desc=> in the tagged corpus. A complete list of the non-verbal sounds is given in Appendix 6.

### 3.3 The prosodic version

A representative selection of COLT conversations, amounting to approximately 150,000 words, has been transcribed prosodically. A survey of the prosodically transcribed conversations that are available is given in Appendix 2. The texts were selected on the basis of three criteria: age, gender and social group. The following prosodic markers were used:

<b>bold type</b>	nucleus
\	fall
/	rise
V	fall-rise
^	rise-fall
-	level
#	tone-unit boundary

### 3.4 The tagged version

COLT has been automatically word class tagged at the University of Lancaster by means of the CLAWS 6 tagset, as developed for the BNC. Some editing and correcting has been done by research assistants at the University of Bergen. The tagset is presented in Appendix 7.

### 3.5 Computer searching with TACT

This section presents the use of the search program TACT on the COLT corpus. TACT enables the user to search in a database of spoken conversations for the location of words, word combinations and word formation patterns. In the COLT database, TACT is applied to give the distribution of an item in relation to certain non-linguistic variables.

Five different types of display systems are available for searches in the corpus: Key Words In Context (KWIC), Variable Context Display, Distribution, Normalised Distribution and Word List. In the following, the query syntax for each of these will be described in turn.

### 3.5.1 KWIC - Key Words In Context

A KWIC display lists all the occurrences of a word (or a regular expression; cf 3.5.5) with one line of context. Here is an example that shows the search output for the word *Bergen*:

```
bergen (10)
B140501 i=68      's funny innit? |w8 Where's Bergen? |w1
B141601 i=10      to? Er the er, university of Bergen or something.
B141601 i=15      Don't you think the university of Bergen is a bag of shite
B141601 i=44      |w2 Burgundy? |w4 Burgundy? |w1 Bergen. |w4 [Bergen,
B141601 i=45      |w4 Burgundy? |w1 Bergen. |w4 [Bergen, that was it yeah.]
B141601 i=46      [Bergen, that was it yeah.] |w2 [Bergen, Bergen yeah, I
B141601 i=46      that was it yeah.] |w2 [Bergen, Bergen yeah, I heard it
B141601 i=49      <name>,|w1 It's alright. |w4 like, Bergen university it was
B141707 i=4       Some students at the university of Bergen. |w9 Shit!
B141708 i=28      this lady from the, university of Bergen. |w8 So how
```

The number in parentheses in the top line shows the total number of occurrences of *Bergen* in the entire corpus. The numbers at the front of each line give the reference number of the conversation, followed by the turn number where the word can be found. The target word appears in the middle of the line. Clicking on the target word shows the full text, which allows a closer study of each occurrence. The KWIC display allows the user to quickly browse a large number of occurrences to see how a particular word is used.

### 3.5.2 Variable Context Display

Whereas the KWIC display gives only one line of context, the Variable Context Display allows the user to control the amount of context in which a word is to be displayed. For example, one can ask for the word *Peter* to be displayed in a context of one line before and one line after the target word. This is done by setting the parameter "Query context" to "Lines" and typing 1 before and 1 after. This yields the following result:

```
----- B132901 i=21
... I'm not giving you <laughing>half!</> ... See you should be taping
everything, that's what I am. Let's just really bore them! ... Peter, I
think you're the most wonderful boy in the whole entire world.
----- B132901 i=333
|w1 Here, i= i= this one, they all the same. They gave us all
the same thing. It's Peter, er Grace, big Anthony, some other people, oh
yeah, Josie's doing it, er who else, what's the I dunno the other girls'
----- B133101 i=27
```

### 3.5.3 Distribution and Normalised Distribution

The Distribution display allows the user to search for the occurrence of a word as it is distributed across the non-linguistic parameters speaker identity, age, gender, social class,

location, setting, occupation, and number of participants. The various searchable factors are listed under "Distribution reference". Here is an example of how the word *wicked* is distributed according to age:

Table 4: Distribution of wicked with respect to age (TACT search result)

AGE	#	Graph
???	12	*****
11	1	*
12	5	***
13	35	*****
14	39	*****
15	20	*****
16	7	****
17	13	*****
18	2	*
19	2	*
39	1	*
45	1	*

The distribution of an item can also be given as a relative figure, ie frequency relative to the total number of words uttered by a particular age group ("Size"). This is done by means of the Normalised Distribution display. The outcome of this search is a table that gives absolute and relative frequencies of an item, as well as some basic statistics:

Table 5: Normalised distribution of wicked with respect to age (TACT search result)

AGE	#	Size	# / Size	Z-Score	Graph (# / Size)
???	12	65155	184.18E-6	-36.26136	*****
11	1	840	1.1905E-3	3.70991	*****
12	5	10526	475.01E-6	9.27099	*****
13	35	83182	420.76E-6	50.96295	*****
14	39	92704	420.69E-6	56.76458	*****
15	20	69825	286.43E-6	-3.57532	*****
16	7	77984	89.762E-6	-79.78786	*****
17	13	19198	677.15E-6	36.08717	*****
18	2	11370	175.90E-6	-6.79282	*****
19	2	6180	323.62E-6	0.81952	*****
39	1	2237	447.03E-6	1.66088	*****
45	1	4524	221.04E-6	-1.69353	*****

### 3.5.4 Word List

The Word List display gives a list of all the words that match a particular pattern. For instance, it is possible to produce a list of all words ending in a particular letter or sequence of letters. (For query syntax of regular expressions, see 3.5.5 below.) This is particularly useful for a researcher who is interested in the productivity of certain morphemes, such as *-able*:

Table 6: Word list and frequencies of morpheme -able (TACT search result)

able (77)	available (5)	believable (1)
cable (1)	capable (2)	changeable (2)
comfortable (9)	fashionable (1)	impressionable (1)
inequitable (1)	inimitable (1)	irritable (1)
malleable (6)	miserable (3)	noticeable (1)
portable (2)	predictable (1)	reasonable (7)
rechargeable (1)	reliable (1)	renewable (8)
reputable (1)	respectable (2)	sizable (1)
sociable (3)	suitable (1)	table (69)
timetable (1)	unable (1)	unavailable (1)
unbelievable (3)	uncomfortable (2)	unfuckingtouchable (1)
unreliable (1)	unscrewable (1)	unsociable (1)
up-gradable (2)	valuable (1)	vulnerable (4)

### 3.5.5 Regular expressions

The use of a so-called regular expression allows the researcher to search for an open-ended word or a combination of letters within a word. Table 6 above was retrieved by searching for words ending in *-able*, and the query code required for this is `".*able"`. The symbol `"."` (full stop) is used to replace any one letter in a regular expression, and the symbol `"*"` (asterisk) is used to replace any number of letters. The following is a list of examples of regular expressions. (The complete survey of query syntax can be found on the web page <http://kh.hit.uib.no/tactweb/doc/query.htm>)

Table 7: Regular expressions in TACT

Query	Output	Output in COLT's Word List mode
f.t	Any three-letter word beginning with the letter "f" and ending with the letter "t"	<i>fat fit</i>
f..t	Any four-letter word beginning with the letter "f" and ending with the letter "t"	<i>fact fart fast feet felt fist flat flit foot fort</i>
f.*t	Any word of any length beginning with the letter "f" and ending with "t"	<i>fact faggit faggot faint fart fast fat fault feet felt figgit fight first fist fit fittest etc</i>
fu...	Any five letter word beginning with the sequence "ba"	<i>fucks fuck's fudge fully fumes funky funny</i>
..cks..	Any seven-letter word containing, as the middle letter, the sequence "cks"	<i>jackson</i>
x.*	Any word of any length beginning with the letter "x"	<i>x xanadu xavier x-rayed x-rays</i>
.*x	Any word of any length ending with the letter "x"	<i>alex andrex box coax complex dex essex ex filofax fix fox fx halifax hendrix knox lax lennox etc</i>
.*x.*	Any word containing the letter "x" in any position	<i>alex alexandra alex's andrex annexe anorexic anxiety anxious auxiliary etc</i>
.*shit.*	Any word containing the sequence "shit" in any position.	<i>apeshit bullshit bullshitter shit shite shithole shits shit's shitted shittiest shittily shitting shitty shity</i>
.*less	Any word ending with the sequence "less"	<i>bless careless countless dickless endless giles's gormless heartless hopeless jules's less mateless etc</i>

### 3.5.6 Selecting certain texts and certain speakers

There are various ways of restricting the search to a certain part of the corpus. For instance, a user may want to restrict the search for instances of *like* to one particular text. This is done by typing "like ; when ref=b140810". The index "ref" picks out a particular text file, namely the conversation b140810. (For a complete list of texts, see Appendix 2.) One may modify the selection even further by restricting the search to a single utterance, by typing "like ; when ref=b140810 & id=122", where the index "id" picks out a particular utterance in a text.

It is also possible to restrict the search to certain speakers in the corpus. This is done by specifying the value for the index "who". The query "like ; when who=1" will select all occurrences of the word *like* when the speaker has speaker identification number 1, ie when the person carrying the tape recorder (the recruit) is speaking. Other possible selection refinements are, for instance:

Table 8: Query syntax for TACT

Parameter	Example	Possible value of parameter
Age of speaker	like ; when age=14	any integer
Location of conversation	like ; when location=hackney	Barnet, Brent, Camden, Enfield, Greater_London, Hackney, Hertfordshire, Islington, London, North_London, Tower_Hamlets, Westminster
Conversational setting	like ; when setting=classroom	See Table 2
Gender of speaker	like ; when gender=m	f, m
Class of speaker	like ; when class=3	1, 2, 3
Occupation of speaker	like ; when occupation=student	artist, bartender, credit_controller, disc_jockey, history_lecturer, housewife, lecturer, policeman, publican, pupil, receptionist, retired, shop_keeper, shopworker, student, studying, teacher, unemployed, unemployed_pubowner
Number of participants in conversation	like ; when number=5	any integer

Furthermore, one may combine several search restrictions by using the coordinator "&" (ampersand), as in "like ; when gender=f & class=3 & location=hackney".

## 4 COLT-based research

The COLT material has already been the source of a large number of studies. At the English department, University of Bergen, we have concentrated on what immediately struck us as characteristic as we began studying the teenage conversations: the frequent use of pragmatic particles (eg *cos, I think, like, sort of*), tags (*innit, okay, right*) and taboo words. Other studies

were devoted to nonstandard features of grammar (eg auxiliary deletion, negative concord, double negation) and interactional style (eg conflict talk). Nearly all the studies reflect a sociolinguistic approach. A recently updated survey of COLT-based research is given in Appendix 1.



## Appendix 1 COLT-based publications (as of December 1998)

- Andersen, G. 1995. 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, G. 1997a. *They gave us these yeah, and they like wanna see like how we talk and all that*. The use of *like* and other discourse markers in London teenage speech. In U-B. Kotsinas, A-B. Stenström & A-M. Karlsson (eds). *Ungdomsspråk i Norden*. Stockholm: MINS 43: 82-95.
- Andersen, G. 1997b. *They like wanna see like how we talk and all that*. The use of *like* as a discourse marker in London teenage speech. In M. Ljung (ed). *Corpus-based studies in English*. Amsterdam: Rodopi, 37-48.
- Andersen, G. 1997c. *I goes you hang it up in your shower, innit? He goes yeah*. The use and development of invariant tags and follow-ups in London teenage speech. Unpublished paper presented at the First Language Variation Workshop, Reading UK.
- Andersen, G. 1998a. The pragmatic marker *like* from a relevance theoretical perspective. In A. Jucker & Y. Ziv (eds). *Discourse markers: descriptions and theory*. Amsterdam: John Benjamins, 147-170.
- Andersen, G. 1998 b. Pragmatic markers and the semantics/pragmatics distinction. Unpublished paper. University of Bergen.
- Andersen, G. Forthcoming a. Are tag questions questions? Evidence from spoken data. Paper presented at the 18th ICAME Conference. Belfast, UK, May 1998.
- Andersen, G. Forthcoming b. The role of the pragmatic marker *like* in the identification of explicatures. Paper presented at the 6th International Pragmatics Conference. Reims, France, July 1998.
- Andersen, G. & Hasund, K. 1996. COLT on TACT: A demonstration of TACT as applied to the Bergen Corpus of Teenage Language. Poster and software demonstration at the ALLC/ACH Conference, University of Bergen. Published online, [http://www.ach.org/ACH\\_Posters/colt.html](http://www.ach.org/ACH_Posters/colt.html)
- Andersen, G. & A-B. Stenström. 1996. COLT: A progress report. *ICAME Journal* 20: 133-136.
- Berland, U. 1997. Invariant tags: Pragmatic functions of *innit*, *okay*, *right* and *yeah* in London teenage conversation. Unpublished MA thesis. Department of English, University of Bergen.
- Bynes, A. 1998. A corpus-based study of expletive use among London teenagers. Unpublished MA thesis. Department of English, University of Bergen.
- Erhardt, S. 1998. Strategien von Gesprächskontrolle bei jugendlichen männlichen und weiblichen Sprechern. Unpublished MA thesis. Department of English, University of Jena.
- Erman, B. 1996. *Just wear the wig innit!* From identifying and proposition-oriented to intensifying and speaker-oriented: grammaticalization in progress. In *Proceedings from the XVIth Scandinavian conference of linguistics*.
- Erman, B. 1997. *Guy's just such a dickhead* The context and function of *just* in teenage talk. In U-B Kotsinas, A-B. Stenström & A-M. Karlsson (eds). *Ungdomsspråk i Norden*. Stockholm: MINS 43: 96-110.
- Guenther, U. Forthcoming. What's funny to whom? Humour as a conversational strategy in everyday talk. A study based on the BNC and COLT. PhD thesis. University of Freiburg.

- Haslerud, V. & A-B. Stenström. 1994. COLT: Mark-up and trends. *Hermes Journal of Linguistics*, 55-70.
- Haslerud, V. & A-B. 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.
- Hasund, K. 1996. COLT conflicts: Reflections of gender and class in the oppositional turn sequences of London teenage girls. Unpublished MA thesis. Department of English, University of Bergen.
- Hasund K. 1998a. From Woman's Place to Women's Places: Class-determined variation in the verbal disputes of London teenage girls. In A. Despard (ed). *A Woman's Place: Women, Domesticity and Private Life*. Kristiansand: Norwegian Academic Press.
- Hasund, K. 1998b. Protecting the innocent: The issue of informants' anonymity in the COLT corpus. In A. Renouf (ed): *Explorations in Corpus Linguistics*. Amsterdam: Rodopi, 13-27.
- Hasund, K. & A-B. Stenström. 1997. Conflict talk: A comparison of the verbal disputes of adolescent females in two corpora. In M. Ljung (ed). *Corpus-based studies in English*. Amsterdam: Rodopi, 119-133.
- Kotsinas, U.-B., Karlsson, A.-M. & A.-B. Stenström. 1997. (eds). *Ungdomsspråk i Norden*. Stockholm: MINS 43.
- Monstad, K. 1998. The grammaticalisation of *sort of* and *kind of* in young and old Londoner's speech. Unpublished MA thesis. Department of English, University of Bergen.
- Mosaker, H-M. 1998. Qualifying utterances: A study of epistemic phrases in COLT and the BNC. Unpublished MA thesis. Department of English, University of Bergen.
- Paradis, C. Forthcoming. *It's well weird*. Degree modifiers of adjectives: the nineties. Paper presented at the 18th ICAME conference. Belfast, UK, May 1998.
- Stenström, A-B. 1995. Taboos in teenage talk. In G. Melchers & B. Warren (eds). *Studies in Anglistics*. Stockholm: Almqvist & Wiksell International, 71-80.
- Stenström, A-B. 1997a. Tags in Teenage Talk. In U. Fries, V. Müller & P. Schneider (eds). *From Ælfric to the New York Times. Studies in English Corpus Linguistics*. Amsterdam: Rodopi, 139-148.
- Stenström, A-B. 1997b. *Can I have a chips please? - Just tell me what one you want*: Nonstandard grammatical features in London teenage talk. In J. Aarts et al (eds). *Studies in English Language and Teaching*. Amsterdam: Rodopi, 141-152.
- Stenström, A-B. 1998a. From sentence to discourse: *cos(because)* in teenage talk. In A. Jucker & Y. Ziv (eds). *Discourse markers: descriptions and theory*. Amsterdam: John Benjamins, 127-146.
- Stenström, A-B. 1998b The intensifier *well* - a recent development? Poster presented at the 6th International Pragmatics Conference. Reims, France, July 1998.
- Stenström, A-B. 1999a. Teenage talk: Chatting about boys and discussing taboo habits. In R. Horowitz (ed). *Developing understanding of the world through talk and text*. The University of San Antonio, Texas.
- Stenström, A-B. 1999b. Chatting about boys and other important things. *In Proceedings from the 16th International Congress of Linguistics in Paris 1997*.
- Stenström, A-B. 1999c. *He was really gormless - She's a bloody crap*. Girls, boys and intensifiers. *Out of Corpus*. Amsterdam: Rodopi.
- Stenström, A-B & Andersen, G. 1996. More trends in teenage talk: A corpus-based investigation of the discourse items *cos* and *innit*. In C. Percy, C. Meyer & I. Lancashire (eds). *Synchronic corpus linguistics*. Amsterdam: Rodopi, 189-203.
- Stenström, A-B & L. E. Breivik. 1993. The Bergen Corpus of Teenage Talk. *ICAME Journal* 17, 128.

- Stenström, A-B & V. Haslerud. 1994. Transcribing COLT: Mark-up and trends. *Hermes. Journal of linguistics* 13, 55-70.
- Straume, A. 1998. The variable (t) in London adolescent speech. A sociolinguistic study of a phonological variable. Unpublished MA thesis, Department of English, University of Bergen.
- Tandberg, A. 1996. *Innit* from a grammatical and pragmatic point of view. Unpublished MA thesis, Department of English, University of Bergen.
- Aas, H. J. 1998. Conversational storytelling in COLT: A study of three stories from the perspective of Systemic Functional Linguistics. Unpublished MA thesis. Department of English, University of Bergen.

## Appendix 2 Survey of COLT text files

Conversation number	File name; orthographic transcription	File name; word class tagged transcription	File name; prosodic transcription
132401	B132401.cor	B132401.tag	B132401.pro
132402	B132402.cor	B132402.tag	B132402.pro
132403	B132403.cor	B132403.tag	B132403.pro
132404	B132404.cor	B132404.tag	B132404.pro
132405	B132405.cor	B132405.tag	B132405.pro
132406	B132406.cor	B132406.tag	B132406.pro
132407	B132407.cor	B132407.tag	B132407.pro
132408	B132408.cor	B132408.tag	B132408.pro
132409	B132409.cor	B132409.tag	B132409.pro
132501	B132501.cor	B132501.tag	B132501.pro
132502	B132502.cor	B132502.tag	B132502.pro
132503	B132503.cor	B132503.tag	B132503.pro
132504	B132504.cor	B132504.tag	B132504.pro
132601	B132601.cor	B132601.tag	B132601.pro
132602	B132602.cor	B132602.tag	B132602.pro
132603	B132603.cor	B132603.tag	B132603.pro
132605	B132605.cor	B132605.tag	B132605.pro
132606	B132606.cor	B132606.tag	B132606.pro
132607	B132607.cor	B132607.tag	B132607.pro
132609	B132609.cor	B132609.tag	B132609.pro
132610	B132610.cor	B132610.tag	B132610.pro
132611	B132611.cor	B132611.tag	B132611.pro
132612	B132612.cor	B132612.tag	B132612.pro
132613	B132613.cor	B132613.tag	B132613.pro
132614	B132614.cor	B132614.tag	B132614.pro
132615	B132615.cor	B132615.tag	B132615.pro
132616	B132616.cor	B132616.tag	B132616.pro
132617	B132617.cor	B132617.tag	B132617.pro
132701	B132701.cor	B132701.tag	B132701.pro
132702	B132702.cor	B132702.tag	B132702.pro
132703	B132703.cor	B132703.tag	B132703.pro
132704	B132704.cor	B132704.tag	B132704.pro
132705	B132705.cor	B132705.tag	B132705.pro
132706	B132706.cor	B132706.tag	B132706.pro
132707	B132707.cor	B132707.tag	B132707.pro
132708	B132708.cor	B132708.tag	B132708.pro
132709	B132709.cor	B132709.tag	
132801	B132801.cor	B132801.tag	
132802	B132802.cor	B132802.tag	
132803	B132803.cor	B132803.tag	
132804	B132804.cor	B132804.tag	
132901	B132901.cor	B132901.tag	
132902	B132902.cor	B132902.tag	
132903	B132903.cor	B132903.tag	
132905	B132905.cor	B132905.tag	
132906	B132906.cor	B132906.tag	
132907	B132907.cor	B132907.tag	
132908	B132908.cor	B132908.tag	
132909	B132909.cor	B132909.tag	
132911	B132911.cor	B132911.tag	
132912	B132912.cor	B132912.tag	
132913	B132913.cor	B132913.tag	
133101	B133101.cor	B133101.tag	B133101.pro
133201	B133201.cor	B133201.tag	B133201.pro
133202	B133202.cor	B133202.tag	B133202.pro
133203	B133203.cor	B133203.tag	B133203.pro
133301	B133301.cor	B133301.tag	B133301.pro
133302	B133302.cor	B133302.tag	B133302.pro
133401	B133401.cor	B133401.tag	B133401.pro
133501	B133501.cor	B133501.tag	B133501.pro
133602	B133602.cor	B133602.tag	B133602.pro
133701	B133701.cor	B133701.tag	B133701.pro
133702	B133702.cor	B133702.tag	B133702.pro
133703	B133703.cor	B133703.tag	B133703.pro

133704	B133704.cor	B133704.tag	B133704.pro
133705	B133705.cor	B133705.tag	B133705.pro
133801	B133801.cor	B133801.tag	B133801.pro
133901	B133901.cor	B133901.tag	B133901.pro
133902	B133902.cor	B133902.tag	B133902.pro
133903	B133903.cor	B133903.tag	
133904	B133904.cor	B133904.tag	
133905	B133905.cor	B133905.tag	
133906	B133906.cor	B133906.tag	
134101	B134101.cor	B134101.tag	
134102	B134102.cor	B134102.tag	
134103	B134103.cor	B134103.tag	
134201	B134201.cor	B134201.tag	
134202	B134202.cor	B134202.tag	
134401	B134401.cor	B134401.tag	
134601	B134601.cor	B134601.tag	
134602	B134602.cor	B134602.tag	
134801	B134801.cor	B134801.tag	
134802	B134802.cor	B134802.tag	
134803	B134803.cor	B134803.tag	
134804	B134804.cor	B134804.tag	
134901	B134901.cor	B134901.tag	
134902	B134902.cor	B134902.tag	
134903	B134903.cor	B134903.tag	
135001	B135001.cor	B135001.tag	
135003	B135003.cor	B135003.tag	
135004	B135004.cor	B135004.tag	
135201	B135201.cor	B135201.tag	
135202	B135202.cor	B135202.tag	
135203	B135203.cor	B135203.tag	
135204	B135204.cor	B135204.tag	
135205	B135205.cor	B135205.tag	
135206	B135206.cor	B135206.tag	
135207	B135207.cor	B135207.tag	
135301	B135301.cor	B135301.tag	
135302	B135302.cor	B135302.tag	
135303	B135303.cor	B135303.tag	
135304	B135304.cor	B135304.tag	
135305	B135305.cor	B135305.tag	
135306	B135306.cor	B135306.tag	
135601	B135601.cor	B135601.tag	
135602	B135602.cor	B135602.tag	
135603	B135603.cor	B135603.tag	
135701	B135701.cor	B135701.tag	
135702	B135702.cor	B135702.tag	
135703	B135703.cor	B135703.tag	
135704	B135704.cor	B135704.tag	
135705	B135705.cor	B135705.tag	
135706	B135706.cor	B135706.tag	
135801	B135801.cor	B135801.tag	
135802	B135802.cor	B135802.tag	
135803	B135803.cor	B135803.tag	
135804	B135804.cor	B135804.tag	
135805	B135805.cor	B135805.tag	
135806	B135806.cor	B135806.tag	
135807	B135807.cor	B135807.tag	
135808	B135808.cor	B135808.tag	
135809	B135809.cor	B135809.tag	
135901	B135901.cor	B135901.tag	B135901.pro
135902	B135902.cor	B135902.tag	B135902.pro
135903	B135903.cor	B135903.tag	B135903.pro
135904	B135904.cor	B135904.tag	B135904.pro
135905	B135905.cor	B135905.tag	B135905.pro
135906	B135906.cor	B135906.tag	B135906.pro
135907	B135907.cor	B135907.tag	B135907.pro
136101	B136101.cor	B136101.tag	B136101.pro
136102	B136102.cor	B136102.tag	B136102.pro
136103	B136103.cor	B136103.tag	B136103.pro
136104	B136104.cor	B136104.tag	B136104.pro
136105	B136105.cor	B136105.tag	B136105.pro
136106	B136106.cor	B136106.tag	B136106.pro
136107	B136107.cor	B136107.tag	B136107.pro

*The Bergen Corpus of London Teenage Language*

136108	B136108.cor	B136108.tag	B136108.pro
136301	B136301.cor	B136301.tag	
136302	B136302.cor	B136302.tag	
136303	B136303.cor	B136303.tag	
136304	B136304.cor	B136304.tag	
136401	B136401.cor	B136401.tag	B136401.pro
136402	B136402.cor	B136402.tag	B136402.pro
136403	B136403.cor	B136403.tag	B136403.pro
136404	B136404.cor	B136404.tag	B136404.pro
136405	B136405.cor	B136405.tag	B136405.pro
136406	B136406.cor	B136406.tag	B136406.pro
136407	B136407.cor	B136407.tag	B136407.pro
136408	B136408.cor	B136408.tag	B136408.pro
136409	B136409.cor	B136409.tag	B136409.pro
136410	B136410.cor	B136410.tag	B136410.pro
136411	B136411.cor	B136411.tag	B136411.pro
136501	B136501.cor	B136501.tag	B136501.pro
136502	B136502.cor	B136502.tag	B136502.pro
136503	B136503.cor	B136503.tag	B136503.pro
136601	B136601.cor	B136601.tag	
136602	B136602.cor	B136602.tag	
136603	B136603.cor	B136603.tag	
136604	B136604.cor	B136604.tag	
136701	B136701.cor	B136701.tag	
136901	B136901.cor	B136901.tag	
136902	B136902.cor	B136902.tag	
136903	B136903.cor	B136903.tag	
137101	B137101.cor	B137101.tag	
137102	B137102.cor	B137102.tag	
137103	B137103.cor	B137103.tag	
137104	B137104.cor	B137104.tag	
137201	B137201.cor	B137201.tag	
137202	B137202.cor	B137202.tag	
137701	B137701.cor	B137701.tag	B137701.pro
137801	B137801.cor	B137801.tag	B137801.pro
137802	B137802.cor	B137802.tag	B137802.pro
137803	B137803.cor	B137803.tag	B137803.pro
137804	B137804.cor	B137804.tag	B137804.pro
137901	B137901.cor	B137901.tag	B137901.pro
137902	B137902.cor	B137902.tag	B137902.pro
137903	B137903.cor	B137903.tag	B137903.pro
137904	B137904.cor	B137904.tag	B137904.pro
138001	B138001.cor	B138001.tag	B138001.pro
138102	B138102.cor	B138102.tag	
138201	B138201.cor	B138201.tag	
138301	B138301.cor	B138301.tag	
138501	B138501.cor	B138501.tag	
138502	B138502.cor	B138502.tag	
138503	B138503.cor	B138503.tag	
138504	B138504.cor	B138504.tag	
138601	B138601.cor	B138601.tag	
138602	B138602.cor	B138602.tag	
138603	B138603.cor	B138603.tag	
138604	B138604.cor	B138604.tag	
138901	B138901.cor	B138901.tag	B138901.pro
138902	B138902.cor	B138902.tag	B138902.pro
138903	B138903.cor	B138903.tag	B138903.pro
138904	B138904.cor	B138904.tag	B138904.pro
138905	B138905.cor	B138905.tag	B138905.pro
138906	B138906.cor	B138906.tag	B138906.pro
138907	B138907.cor	B138907.tag	B138907.pro
139001	B139001.cor	B139001.tag	B139001.pro
139002	B139002.cor	B139002.tag	B139002.pro
139003	B139003.cor	B139003.tag	B139003.pro
139201	B139201.cor	B139201.tag	B139201.pro
139301	B139301.cor	B139301.tag	B139301.pro
139302	B139302.cor	B139302.tag	B139302.pro
139303	B139303.cor	B139303.tag	B139303.pro
139304	B139304.cor	B139304.tag	B139304.pro
139305	B139305.cor	B139305.tag	B139305.pro
139306	B139306.cor	B139306.tag	B139306.pro
139307	B139307.cor	B139307.tag	B139307.pro

139308	B139308.cor	B139308.tag	B139308.pro
139401	B139401.cor	B139401.tag	
139402	B139402.cor	B139402.tag	
139403	B139403.cor	B139403.tag	
139501	B139501.cor	B139501.tag	
139502	B139502.cor	B139502.tag	
139503	B139503.cor	B139503.tag	
139504	B139504.cor	B139504.tag	
139505	B139505.cor	B139505.tag	
139506	B139506.cor	B139506.tag	
139601	B139601.cor	B139601.tag	
139602	B139602.cor	B139602.tag	
139603	B139603.cor	B139603.tag	
139604	B139604.cor	B139604.tag	
139605	B139605.cor	B139605.tag	
139606	B139606.cor	B139606.tag	
139607	B139607.cor	B139607.tag	
139608	B139608.cor	B139608.tag	
139609	B139609.cor	B139609.tag	
139610	B139610.cor	B139610.tag	
139611	B139611.cor	B139611.tag	
139612	B139612.cor	B139612.tag	
139613	B139613.cor	B139613.tag	
139614	B139614.cor	B139614.tag	
139701	B139701.cor	B139701.tag	
139702	B139702.cor	B139702.tag	
139703	B139703.cor	B139703.tag	
139704	B139704.cor	B139704.tag	
139705	B139705.cor	B139705.tag	
139706	B139706.cor	B139706.tag	
139707	B139707.cor	B139707.tag	
139708	B139708.cor	B139708.tag	
139709	B139709.cor	B139709.tag	
139801	B139801.cor	B139801.tag	
139802	B139802.cor	B139802.tag	
139803	B139803.cor	B139803.tag	
139804	B139804.cor	B139804.tag	
139805	B139805.cor	B139805.tag	
139806	B139806.cor	B139806.tag	
139807	B139807.cor	B139807.tag	
139808	B139808.cor	B139808.tag	
139809	B139809.cor	B139809.tag	
140201	B140201.cor	B140201.tag	
140202	B140202.cor	B140202.tag	
140301	B140301.cor	B140301.tag	
140302	B140302.cor	B140302.tag	
140303	B140303.cor	B140303.tag	
140401	B140401.cor	B140401.tag	
140402	B140402.cor	B140402.tag	
140403	B140403.cor	B140403.tag	
140501	B140501.cor	B140501.tag	
140502	B140502.cor	B140502.tag	
140503	B140503.cor	B140503.tag	
140504	B140504.cor	B140504.tag	
140505	B140505.cor	B140505.tag	
140601	B140601.cor	B140601.tag	
140602	B140602.cor	B140602.tag	
140603	B140603.cor	B140603.tag	
140604	B140604.cor	B140604.tag	
140605	B140605.cor	B140605.tag	
140606	B140606.cor	B140606.tag	
140607	B140607.cor	B140607.tag	
140701	B140701.cor	B140701.tag	
140702	B140702.cor	B140702.tag	
140703	B140703.cor	B140703.tag	
140704	B140704.cor	B140704.tag	
140705	B140705.cor	B140705.tag	
140706	B140706.cor	B140706.tag	
140707	B140707.cor	B140707.tag	
140708	B140708.cor	B140708.tag	
140709	B140709.cor	B140709.tag	
140801	B140801.cor	B140801.tag	

*The Bergen Corpus of London Teenage Language*


140802	B140802.cor	B140802.tag	
140803	B140803.cor	B140803.tag	
140804	B140804.cor	B140804.tag	
140805	B140805.cor	B140805.tag	
140806	B140806.cor	B140806.tag	
140807	B140807.cor	B140807.tag	
140808	B140808.cor	B140808.tag	
140809	B140809.cor	B140809.tag	
140810	B140810.cor	B140810.tag	
140811	B140811.cor	B140811.tag	
140901	B140901.cor	B140901.tag	
140904	B140904.cor	B140904.tag	
141001	B141001.cor	B141001.tag	
141002	B141002.cor	B141002.tag	
141003	B141003.cor	B141003.tag	
141101	B141101.cor	B141101.tag	B141101.pro
141102	B141102.cor	B141102.tag	B141102.pro
141103	B141103.cor	B141103.tag	B141103.pro
141104	B141104.cor	B141104.tag	B141104.pro
141105	B141105.cor	B141105.tag	B141105.pro
141106	B141106.cor	B141106.tag	B141106.pro
141107	B141107.cor	B141107.tag	B141107.pro
141201	B141201.cor	B141201.tag	
141202	B141202.cor	B141202.tag	
141203	B141203.cor	B141203.tag	
141204	B141204.cor	B141204.tag	
141301	B141301.cor	B141301.tag	
141302	B141302.cor	B141302.tag	
141303	B141303.cor	B141303.tag	
141401	B141401.cor	B141401.tag	
141402	B141402.cor	B141402.tag	
141403	B141403.cor	B141403.tag	
141404	B141404.cor	B141404.tag	
141405	B141405.cor	B141405.tag	
141501	B141501.cor	B141501.tag	
141601	B141601.cor	B141601.tag	B141601.pro
141602	B141602.cor	B141602.tag	B141602.pro
141603	B141603.cor	B141603.tag	B141603.pro
141604	B141604.cor	B141604.tag	B141604.pro
141605	B141605.cor	B141605.tag	B141605.pro
141606	B141606.cor	B141606.tag	B141606.pro
141701	B141701.cor	B141701.tag	B141701.pro
141702	B141702.cor	B141702.tag	
141703	B141703.cor	B141703.tag	
141704	B141704.cor	B141704.tag	
141705	B141705.cor	B141705.tag	
141706	B141706.cor	B141706.tag	
141707	B141707.cor	B141707.tag	
141708	B141708.cor	B141708.tag	
141801	B141801.cor	B141801.tag	
141802	B141802.cor	B141802.tag	
141803	B141803.cor	B141803.tag	
141804	B141804.cor	B141804.tag	
141805	B141805.cor	B141805.tag	
141806	B141806.cor	B141806.tag	
141901	B141901.cor	B141901.tag	
141902	B141902.cor	B141902.tag	
141903	B141903.cor	B141903.tag	
141904	B141904.cor	B141904.tag	
141905	B141905.cor	B141905.tag	
141906	B141906.cor	B141906.tag	
141907	B141907.cor	B141907.tag	
142001	B142001.cor	B142001.tag	
142002	B142002.cor	B142002.tag	
142003	B142003.cor	B142003.tag	
142004	B142004.cor	B142004.tag	
142005	B142005.cor	B142005.tag	
142101	B142101.cor	B142101.tag	B142101.pro
142102	B142102.cor	B142102.tag	B142102.pro
142103	B142103.cor	B142103.tag	B142103.pro
142104	B142104.cor	B142104.tag	B142104.pro
142105	B142105.cor	B142105.tag	B142105.pro



142106	B142106.cor	B142106.tag	B142106.pro
142201	B142201.cor	B142201.tag	B142201.pro
142202	B142202.cor	B142202.tag	B142202.pro
142301	B142301.cor	B142301.tag	B142301.pro
142302	B142302.cor	B142302.tag	B142302.pro
142303	B142303.cor	B142303.tag	B142303.pro
142304	B142304.cor	B142304.tag	B142304.pro
142305	B142305.cor	B142305.tag	B142305.pro
142306	B142306.cor	B142306.tag	B142306.pro
142307	B142307.cor	B142307.tag	B142307.pro
142601	B142601.cor	B142601.tag	B142601.pro
142602	B142602.cor	B142602.tag	B142602.pro
142603	B142603.cor	B142603.tag	B142603.pro
142604	B142604.cor	B142604.tag	B142604.pro
142606	B142606.cor	B142606.tag	B142606.pro
142607	B142607.cor	B142607.tag	B142607.pro
142608	B142608.cor	B142608.tag	B142608.pro
142701	B142701.cor	B142701.tag	B142701.pro
142702	B142702.cor	B142702.tag	B142702.pro
142703	B142703.cor	B142703.tag	B142703.pro
142704	B142704.cor	B142704.tag	
142705	B142705.cor	B142705.tag	
142706	B142706.cor	B142706.tag	
142801	B142801.cor	B142801.tag	
142802	B142802.cor	B142802.tag	

### Appendix 3 Personal data sheet

UNIVERSITETET I BERGEN  
ENGELSK INSTITUTT  
Sydnesplass 9 - 5007 Bergen  
Tlf.: (05) 21 30 50  
Innv.: (05) 21 23 60  
Telefax: (05) 23 18 97



UNIVERSITY OF BERGEN  
DEPARTMENT OF ENGLISH  
Sydnesplass 9  
N-5007 Bergen  
Norway

## Personal data

All information will be treated confidentially. Recruits, their family and conversation partners are guaranteed full anonymity.

Recruit number 18

A) Area of residence (London borough) Brent  
Postcode NW10 4AJ

B) Have you ever lived in any other part of England? If so, where? — and for how long? —

C) Mother's occupation Teacher  
Is she currently employed? Y/N Yes

D) Father's occupation Graphic designer  
Is he currently employed? Y/N Yes

Please remember to take this form with you when you are giving back the personal stereo etc. to Ms Haslerud

Thanks very much for your co-operation!

## Appendix 4 Personal data survey

Recruit number	Recruit name	Text files	Sex	Age	School borough	Residential borough	Postcode	Also lived in	Mother's occupation	Currently employed?	Father's occupation	Currently employed?	Social group
1	Peter	B132401-B132504	M	14	Hackney	Hackney	E97HT		Teacher	Yes	Doctor(GP)	Yes	2
2	Josie	B132601-B132913	F	14	Hackney	Hackney	6AP		Housewife	No	Plumber	Yes	3
3	Robert	B133101-B133602	M	15	Hackney	Hackney	N16 OEU		Receptionist	Yes	Carpenter	Yes	3
4	Cassie	B133701-B134202	F	15	Hackney	Hackney	N16		Yoga teacher/gardener	Yes	Gardener	Yes	3
5	Anthony	B134401-B134602	M	15	Hackney	Hackney	OJR		College student	Yes	British telecom	Yes	3
6	Grace	B134801-B135207	F	14	Hackney	Stoke Newington	OSX			No		No	3
7	Richard	B135301-B135306	M	13	Tower Hamlets	Tower Hamlets	E3 5LG			No		Yes	
8	Michael	B135601-B135603	M	13	Tower Hamlets	Bow	G3 4NS	Essex (2 years)	Credit control	No	Lorry driver	Yes	3
9	Craig	B135701-B135809	M	13	Tower Hamlets	Tower Hamlets	E2		Advertiser	No	Pub owner	No	3
10	Anthony	B135901-B135907	M	13	Tower Hamlets	Bow	45E		Admin officer	Yes	Student	No	2
11	Barry	B136101-B136108	M	14	Tower Hamlets	Bow	2RS		Waitress	No	Market trader	Yes	3
12	Mark	B136301-B136304	M	13	Tower Hamlets	Bow	2EN		Social worker	Yes	Taylor	Yes	2
13	Carla	B136401-B136503	F	13	Camden	Westminster	W9 2NU			No	Science engineer	Yes	1
14	Sarah	B136601-B136903	F	13	Camden	Brent	NW6 6SB	Devon (1year)	Housewife	No	Computer work	Yes	2
15	Leon	B137101-B137202	M	13	Camden	Camden	NW1 9BJ		Teacher	Yes			1
17	Selum	B137701-B138301	M	13	Camden	Camden	NW0 2BA	Rowleyway (2 yrs)	Secretary	Yes	Taylor	Yes	2
18	Matthew	B138501-B138604	M	13	Camden	Brent	NW10 4AJ		Teacher	Yes	Graphic designer	Yes	2
19	Jonathan	B138901-B139003	M	17	Camden	Camden	NW3 3SV				Trade finance	Yes	1
21	Skonev	B139201-B139308	M	12	Camden	Islington	N19		Housewife	No	History lecturer	Yes	1
22	Terry	B139401-B139614	M	14	Barnet	Barnet	EN4 8PN		Housewife	No	Dress cutter	Yes	2
23	Robin	B139701-B139809	M	14	Barnet	Southgate	N14 5RY		Student (course)	No	Financial consultant	Yes	1
24	Eddie	B140201-B140303	M	15	Barnet	Barnet	N14 5RG		Teacher	Yes	Retired teacher	No	1
25	Alex	B140401-B140607	M	14	Barnet	Barnet	N14		Managing director	Yes	Engineer	Yes	1
26	Caroline	B140701-B140904	F	14	Barnet	Hackney	N4 2LX	Southgate (4 yrs)	Housewife	No	Computer programmer	Yes	2
27	Cassie	B141001-B141003	F	15	Barnet	Enfield	2 RG		Clerk	Yes	Civil servant	Yes	2
28	Danny	B141101-B141501	M	13	Hertfordshire	Islington	3JS	Enfield(6 years)	Pubowner	Yes	Pubowner	Yes	2
29	Jock	B141601-B142005	M	16	Hertfordshire	Richmond		Scotland(14 yrs)	Housewife	No	Loss control coordinator	Yes	1
30	Alistair	B142101-B142202	M	15	Hertfordshire	Essex	COg 4NX		Caterer	Yes	Insurance broker	Yes	1
31	Kath	B142301-B142307	F	17	Hertfordshire	Hertfordshire	EN5 4RL		Farmer	Yes	Farmer	Yes	1
33	Catriona	B142601-B142802	F	16	Hertfordshire	North London	N6 5BY		Artist	Yes	Artist	Yes	1

## **Appendix 5 Paralinguistic features in COLT**

- <hand over mouth>
- <laughing>
- <mimicking a man's voice>
- <mimicking a monkey>
- <mimicking a sexy woman's voice>
- <mimicking a stupid man's voice>
- <mimicking a woman's voice>
- <mimicking a yobbo hooligan>
- <mimicking African accent>
- <mimicking American accent - from Wayne's World>
- <mimicking American accent>
- <mimicking baby voice>
- <mimicking baby's voice>
- <mimicking Chinese accent>
- <mimicking Cockney accent>
- <mimicking crying>
- <mimicking dark voice from chewing gum advert>
- <mimicking dark voice>
- <mimicking drunken voice>
- <mimicking foreign (Pakistani?) accent>
- <mimicking foreign accent>
- <mimicking German accent>
- <mimicking girlie voice>
- <mimicking girl's voice>
- <mimicking her sister's accent>
- <mimicking Indian accent>
- <mimicking Jamaican accent>
- <mimicking Josie>
- <mimicking lisp>
- <mimicking Liverpoolian accent>
- <mimicking male voice>
- <mimicking man>
- <mimicking mentally handicapped>
- <mimicking Mr Bean's voice>
- <mimicking nasal speech>
- <mimicking Northern accent>
- <mimicking old man>
- <mimicking old woman's voice>
- <mimicking Pakistani accent>
- <mimicking police siren>
- <mimicking posh accent>
- <mimicking refined accent>
- <mimicking Romanian accent>
- <mimicking Scottish accent>
- <mimicking speaker 15's voice>
- <mimicking stupid person>
- <mimicking stutter>
- <mimicking Swedish accent>
- <mimicking Swedish chef from Muppet Show>
- <mimicking teacher's voice>
- <mimicking upper class person>
- <mimicking West Indian accent>
- <mimicking witch voice>
- <mimicking woman's voice>
- <mimicking Yorkshire accent>
- <mimicking+distorted voice>
- <mimicking+laughing>
- <mimicking+whining>
- <mimicking>
- <raised voice>
- <raising voice>
- <reading>
- <screaming>
- <shouting+mimicking American accent>
- <shouting+mimicking>
- <shouting>
- <sighing>
- <singing>
- <speaking French>
- <speaking quietly just for the tape>
- <speaking quietly>
- <speaking Spanish>
- <speaking with mouth full>
- <talking slowly>
- <teasing>
- <translating>
- <whinging>
- <whining>
- <whispering>
- <yawning>

## Appendix 6 Non-verbal sounds in COLT

<nv>barking</nv>  
 <nv>belch</nv>  
 <nv>blowing air</nv>  
 <nv>blowing air through lips</nv>  
 <nv>blowing nose</nv>  
 <nv>blurgh</nv>  
 <nv>bomb sounds</nv>  
 <nv>boomph</nv>  
 <nv>breathe</nv>  
 <nv>breathing heavily</nv>  
 <nv>burp</nv>  
 <nv>chewing</nv>  
 <nv>choke</nv>  
 <nv>clap</nv>  
 <nv>clapping</nv>  
 <nv>clears throat</nv>  
 <nv>clicking fingers</nv>  
 <nv>clicks tongue</nv>  
 <nv>cough</nv>  
 <nv>coughing</nv>  
 <nv>crying</nv>  
 <nv>deep breath</nv>  
 <nv>drinking</nv>  
 <nv>drinking sound</nv>  
 <nv>drowning noises</nv>  
 <nv>eating crips</nv>  
 <nv>gasp</nv>  
 <nv>giggle</nv>  
 <nv>hissing sound</nv>  
 <nv>howl</nv>  
 <nv>humming</nv>  
 <nv>humming+singing</nv>  
 <nv>imitates cat licking</nv>  
 <nv>imitating computer beep</nv>  
 <nv>imitating gun noise</nv>  
 <nv>imitating gun fire</nv>  
 <nv>imitates vomiting</nv>  
 <nv>kiss</nv>  
 <nv>kissing microphone</nv>  
 <nv>laugh</nv>  
 <nv>makes car sounds</nv>  
 <nv>makes creaking door noise</nv>  
 <nv>makes drunken sounds and a pretend  
 belch</nv>  
 <nv>makes incantation sounds</nv>  
 <nv>makes running noises</nv>  
 <nv>making odd noises</nv>  
 <nv>making sucking noises</nv>  
 <nv>making vomiting noises</nv>  
 <nv>meow</nv>  
 <nv>mimicking</nv>  
 <nv>mimicking banjo</nv>  
 <nv>mimicking blowing  
 handkerchief</nv>  
 <nv>mimicking bringing up phlegm</nv>  
 <nv>mimicking buzzing sound</nv>  
 <nv>mimicking car sounds</nv>  
 <nv>mimicking chicken sound</nv>  
 <nv>mimicking clearing throat</nv>  
 <nv>mimicking cry</nv>  
 <nv>mimicking dog barks</nv>  
 <nv>mimicking engine revving</nv>  
 <nv>mimicking explosion</nv>  
 <nv>mimicking gorilla noises</nv>  
 <nv>mimicking kissing sound</nv>  
 <nv>mimicking licking sound</nv>  
 <nv>mimicking microphone noises</nv>  
 <nv>mimicking person speaking with  
 mouth full</nv>  
 <nv>mimicking Scandinavian accent from  
 the Muppet Show</nv>  
 <nv>mimicking sound effect</nv>  
 <nv>mimicking sound of biting</nv>  
 <nv>mimicking shaving noise</nv>  
 <nv>mimicking squeaking</nv>  
 <nv>mimicking squeaking noise</nv>  
 <nv>mimicking vomiting</nv>  
 <nv>moan</nv>  
 <nv>moaning</nv>  
 <nv>noise for paws</nv>  
 <nv>noise in microphone</nv>  
 <nv>panting</nv>  
 <nv>purring noises</nv>  
 <nv>raspberry</nv>  
 <nv>running sound</nv>  
 <nv>scream</nv>  
 <nv>sigh</nv>  
 <nv>singing</nv>  
 <nv>singing+humming</nv>  
 <nv>sharp intake of breath</nv>  
 <nv>shout</nv>  
 <nv>slurps</nv>  
 <nv>slurping noises</nv>  
 <nv>shouting</nv>  
 <nv>sneeze</nv>  
 <nv>sniff</nv>  
 <nv>sound effect</nv>  
 <nv>sound of drinking with a straw</nv>

<nv>spitting cat sound</nv>  
<nv>spitting sound</nv>  
<nv>squawking sound</nv>  
<nv>squeals</nv>  
<nv>sucking noise</nv>  
<nv>sucking then purring noises</nv>  
<nv>sulking</nv>  
<nv>vomiting noises </nv>  
<nv>whine</nv>  
<nv>whining</nv>  
<nv>whinge</nv>  
<nv>whinging</nv>  
<nv>whistling</nv>  
<nv>yawn</nv>  
<nv>yelping sound</nv>

**Appendix 7 COLT tagset (CLAWS 6)**

- APPGE possessive pronoun, pre-nominal (*e.g. my, your, our*)  
   AT article (*e.g. the, no*)  
   AT1 singular article (*e.g. a, an, every*)  
   BCL before-clause marker (*e.g. in order (that), in order (to)*)  
   CC coordinating conjunction (*e.g. and, or*)  
   CCB adversative coordinating conjunction (*but*)  
   CS subordinating conjunction (*e.g. if, because, unless, so, for*)  
   CSA as (*as conjunction*)  
   CSN than (*as conjunction*)  
   CST that (*as conjunction*)  
   CSW whether (*as conjunction*)  
   DA after-determiner or post-determiner capable of pronominal function (*e.g. such, former, same*)  
   DA1 singular after-determiner (*e.g. little, much*)  
   DA2 plural after-determiner (*e.g. few, several, many*)  
   DAR comparative after-determiner (*e.g. more, less, fewer*)  
   DAT superlative after-determiner (*e.g. most, least, fewest*)  
   DB before determiner or pre-determiner capable of pronominal function (*all, half*)  
   DB2 plural before-determiner (*both*)  
   DD determiner (*capable of pronominal function*) (*e.g. any, some*)  
   DD1 singular determiner (*e.g. this, that, another*)  
   DD2 plural determiner (*e.g. these, those*)  
   DDQ wh-determiner (*which, what*)  
 DDQGE wh-determiner, genitive (*whose*)  
 DDQV wh-ever determiner (*whichever, whatever*)  
   EX existential there  
   FO formula  
   FU unclassified word  
   FW foreign word  
   GE germanic genitive marker (*' or 's*)  
   IF for (*as preposition*)  
   II general preposition  
   IO of (*as preposition*)  
   IW with, without (*as prepositions*)  
   JJ general adjective  
   JJR general comparative adjective (*e.g. older, better, stronger*)  
   JJT general superlative adjective (*e.g. oldest, best, strongest*)  
   JK catenative adjective (*able in be able to, willing in be willing to*)  
   MC cardinal number, neutral for number (*two, three*)  
 MCGE genitive cardinal number, neutral for number (*two's, 100's*)  
 MCMC hyphenated number (*40-50, 1770-1827*)  
   MC1 singular cardinal number (*one*)  
   MC2 plural cardinal number (*e.g. sixes, sevens*)  
   MD ordinal number (*e.g. first, second, next, last*)  
   MF fraction, neutral for number (*e.g. quarters, two-thirds*)  
   ND1 singular noun of direction (*e.g. north, southeast*)  
   NN common noun, neutral for number (*e.g. sheep, cod, headquarters*)

- NN1 singular common noun (*e.g. book, girl*)
- NN2 plural common noun (*e.g. books, girls*)
- NNA following noun of title (*e.g. M.A.*)
- NNB preceding noun of title (*e.g. Mr., Prof.*)
- NNJ organisation noun, neutral for number (*e.g. council, department*)
- NNJ2 organisation noun, plural (*e.g. governments, committees*)
- NNL1 singular locative noun (*e.g. island, street*)
- NNL2 plural locative noun (*e.g. islands, streets*)
- NNO numeral noun, neutral for number (*e.g. dozen, hundred*)
- NNO2 numeral noun, plural (*e.g. hundreds, thousands*)
- NNT1 temporal noun, singular (*e.g. day, week, year*)
- NNT2 temporal noun, plural (*e.g. days, weeks, years*)
- NNU unit of measurement, neutral for number (*e.g. in, cc*)
- NNU1 singular unit of measurement (*e.g. inch, centimetre*)
- NNU2 plural unit of measurement (*e.g. ins., feet*)
- NP proper noun, neutral for number (*e.g. IBM, Andes*)
- NP1 singular proper noun (*e.g. London, Jane, Frederick*)
- NP2 plural proper noun (*e.g. Browns, Reagans, Koreas*)
- NPD1 singular weekday noun (*e.g. Sunday*)
- NPD2 plural weekday noun (*e.g. Sundays*)
- NPM1 singular month noun (*e.g. October*)
- NPM2 plural month noun (*e.g. Octobers*)
- PN indefinite pronoun, neutral for number (*none*)
- PN1 indefinite pronoun, singular (*e.g. anyone, everything, nobody, one*)
- PNQO objective wh-pronoun (*whom*)
- PNQS subjective wh-pronoun (*who*)
- PNQV wh-ever pronoun (*whoever*)
- PNX1 reflexive indefinite pronoun (*oneself*)
- PPGE nominal possessive personal pronoun (*e.g. mine, yours*)
- PPH1 3rd person sing. neuter personal pronoun (*it*)
- PPHO1 3rd person sing. objective personal pronoun (*him, her*)
- PPHO2 3rd person plural objective personal pronoun (*them*)
- PPHS1 3rd person sing. subjective personal pronoun (*he, she*)
- PPHS2 3rd person plural subjective personal pronoun (*they*)
- PPIO1 1st person sing. objective personal pronoun (*me*)
- PPIO2 1st person plural objective personal pronoun (*us*)
- PPIS1 1st person sing. subjective personal pronoun (*I*)
- PPIS2 1st person plural subjective personal pronoun (*we*)
- PPX1 singular reflexive personal pronoun (*e.g. yourself, itself*)
- PPX2 plural reflexive personal pronoun (*e.g. yourselves, themselves*)
- PPY 2nd person personal pronoun (*you*)
- RA adverb, after nominal head (*e.g. else, galore*)
- REX adverb introducing appositional constructions (*namely, e.g.*)
- RG degree adverb (*very, so, too*)
- RGQ wh-degree adverb (*how*)
- RGQV wh-ever degree adverb (*however*)
- RGR comparative degree adverb (*more, less*)
- RGT superlative degree adverb (*most, least*)
- RL locative adverb (*e.g. alongside, forward*)
- RP prep. adverb, particle (*e.g. about, in*)



RPK	prep. adv., catenative ( <i>about in be about to</i> )
RR	general adverb
RRQ	wh- general adverb ( <i>where, when, why, how</i> )
RRQV	wh-ever general adverb ( <i>wherever, whenever</i> )
RRR	comparative general adverb ( <i>e.g. better, longer</i> )
RRT	superlative general adverb ( <i>e.g. best, longest</i> )
RT	quasi-nominal adverb of time ( <i>e.g. best, longest</i> )
TO	infinitive marker ( <i>to</i> )
UH	interjection ( <i>e.g. oh, yes, um</i> )
VB0	be base form ( <i>finite i.e. imperative, subjunctive</i> )
VBDR	were
VBDZ	was
VBG	being
VBI	be infinitive ( <i>To be or not... It will be...</i> )
VBM	am
VBN	been
VBR	are
VBZ	is
VD0	do base form ( <i>finite</i> )
VDD	did
VDG	doing
VDI	do infinitive ( <i>I may do... To do...</i> )
VDN	done
VDZ	does
VH0	have base form ( <i>finite</i> )
VHD	had ( <i>past tense</i> )
VHG	having
VHI	have infinitive
VHN	had ( <i>past participle</i> )
VHZ	has
VM	modal auxiliary ( <i>can, will, would, etc.</i> )
VMK	modal catenative ( <i>ought, used</i> )
VV0	base form of lexical verb ( <i>e.g. give, work</i> )
VVD	past tense of lexical verb ( <i>e.g. gave, worked</i> )
VVG	-ing participle of lexical verb ( <i>e.g. giving, working</i> )
VVGK	-ing participle catenative ( <i>going in be going to</i> )
VVI	infinitive ( <i>e.g. to give... It will work...</i> )
VVN	past participle of lexical verb ( <i>e.g. given, worked</i> )
VVNK	past participle catenative ( <i>e.g. bound in be bound to</i> )
VVZ	-s form of lexical verb ( <i>e.g. gives, works</i> )
XX	not, n't
ZZ1	singular letter of the alphabet ( <i>e.g. A, b</i> )
ZZ2	plural letter of the alphabet ( <i>e.g. A's, b's</i> )

### Punctuation tags

YEX	punctuation tag - exclamation mark (!)
YQUO	punctuation tag - quotes (")
YBL	punctuation tag - left bracket (()
YBR	punctuation tag - right bracket ())

YCOM punctuation tag - comma (,)  
YDSH punctuation tag - dash (-)  
YSTP punctuation tag - full-stop (.)  
YLIP punctuation tag - ellipsis (...)  
YCOL punctuation tag - colon (:)  
YSCOL punctuation tag - semicolon (;)  
YQUE punctuation tag - question mark (?)