COLA: A Spanish spoken corpus of youth language

Knut Hofland

Aksis/UNIFOB
University of Bergen
knut.hofland@aksis.uib.no

Annette Myre Jørgensen

Department of Romance Languages
University of Bergen
annette.myre@roman.uib.no

Eli-Marie Drange

Department of Romance Languages
University of Bergen
eli-marie.drange@roman.uib.no

Anna-Brita Stenström

Prof. eremitus
ab.stenstrom@telia.com

Abstract

In this paper we describe the COLA corpus. We give an overview of the contents and how the recording and transcription was done. We describe how a web-based system with access to sound is used for browsing and searching. In the end we give a summary of current research on the material.

1. Introduction

The aim of the COLA Corpus (Corpus Oral de Lenguaje Adolescente) is to build a corpus of informal Spanish youth language from Madrid and other capitals of Spanish speaking countries and to stimulate to research on youth language. The project started in 2002 by the initiative from Annette Myre Jørgensen at the Department of Romance languages at the University of Bergen and Anna-Brita Stenström at Department of English, and it is funded by the Faculty of Arts at the University of Bergen and the Meltzer fund.

The method used for recording the data follows the same pattern as the COLT Corpus of English adolescents and the UNO Corpus of Norwegian adolescents, which in turn is patterned on the Longman model used for collecting the British National Corpus (BNC) (Haslerud 1995:235; Crowdy 1995). The recruits were selected from schools in areas with different social status in order to create a balanced corpus with regards to gender, type of school and social status. The recruits are also between 13-18 years old. Each recruit was then equipped with a Minidisc recorder and a microphone, and asked to record his or her conversations with friends and at school for a few days. Some of the conversations are recorded at school, in breaks or during teamwork, and some of the conversations are recorded at home or at places where adolescents use to meet, as parks and so on. The recruits filled in a questionnaire with some personal information as place of birth, language spoken at home, etc, and they were also requested to write down some information about the other participants in their conversations.

This method is very useful to record informal conversations, and you avoid the “observer’s paradox” mentioned by William Labov: “to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation” (Labov 1972:209). The different topics of discussion and the casual way of speech support the idea that the participants in the conversations do not care about the presence of the microphone. They sometimes refer to the microphone in the conversations, but just to assure themselves that the recording devices are functioning without any problems. The recorded data is particularly suitable to study the communicative style and the vocabulary used in the youth language, and since we also receive information about the background of the recruits, it is possible to do sociolinguistic surveys on the corpus. Since the COLA Corpus follows the same pattern as the COLT Corpus and the UNO Corpus, it is also possible to do cross-linguistic studies between Spanish, English and Norwegian.

At this point more than 50 recruits of different age, gender and social background have participated in the COLA Project, and they have recorded about 76 hours of conversations with their friends and acquaintances in Madrid (Colam), Buenos Aires (Colaba) and Santiago Chile (Colas). The corpus will later be expanded by adding material from other places.

2. Recording and transfer to a PC

The recordings were done with a Sony NetMD minidisc (MD) recorder with a lapel microphone. The NetMD has a USB connection, but due to digital rights management reasons it is not possible to upload sound files from the MD to a PC through the USB port. Our plan was to use a MD deck connected with a digital link to a PC for transfer of the files (in real time). But we experienced dropouts in the sound (because the discs were written by portable units) and in the end we had to transfer the sound through the analogue output of the portable MD. The most recent version of the Sony MDs, the Hi-MD, has the possibily of uploading files through the USB port, as long as it has been recorded from an analogue source (minidisc.org 2005). But it is not possible to upload files from a disc recorded on a non Hi-MD recorder. The transferred files are stored in WAV format at our file server.

3. Transcription

When the project started we chose Transcriber as the software for transcription of the sound from the digital sound files (Barras et. al 1998). The Linguistic Data Corporation (LDC) has a good overview over current software for transcribing and annotation on their web-site (LDC 2005). Several papers at the last LREC conference covered different software for transcription (Garg 2004, Martin 2004, Schmidt 2004). The conversations are transcribed orthographically. The command “insert event” is used to indicate extralinguistic information as whether the text is said laughing, singing, shouting, etc. Further, Transcriber has a special command to show overlapping speech, but this is limited to two speakers. The transcript is synchronized to the audio file each time the return key is hit. These points can later be adjusted. When playing the sound file the text will scroll and when selecting a text line the player moves to the corresponding part of the sound file. A screen dump of the program is shown in figure 1 (after the references). Transcriber saves the information in a XML format and this format is shown in figure 2.

4. Browsing and searching

The transcripts and the sound files are put into a web-based system made for the COLT project and further developed in several projects studying Norwegian dialects (Hofland 2003). An overview over this system is given in figure 3. The transcript is converted to a HTML file with links to sound fragments. There are links to turns, lines and fixed time fragments (10 seconds). On the server there is a small program extracting the sound fragment based on file name, start time and duration. This fragment can be delivered as a WAV file or as a compressed MP3 file (by the Lame program). The WAV fragment can automatically be entered into Praat for further phonetic analysis if the browser is configured to use Praat for playing WAV files. An examle of the HTML file is shown in figure 4.

In addition to the transcripts we have a file with information about every speaker, giving gender, age etc. The transcripts and information from this file are converted into the Corpus Workbench format (Christ 1994). We have a web interface to this search program and a sample search screen is given in figure 5. It is possible to search for a combination of words and personal information and also any linguistic information given to the words (either by automatic or manual tagging). An example of the KWIC result is given in figure 6. Each KWIC line has a link to a few seconds of sound context for the keyword. Since the text and sound are only synchronized at certain points, the time stamping for the words between these points is done by interpolation. The sound fragments from the concordance can also be delivered as WAV or MP3 files. A special version of the system exists where the user can make a classification of each concordance line. This information is saved and can be used for crosstabulation with personal information or any other information about the text.

The texts have been converted to the format used by version 4 of WordSmith (a stand-alone Windows concordance program), see figure 7. This version can play sound from the concordance and also have better support for XML-coded texts. But the search facilities are not as flexible as with the Corpus Workbench.

It will be easy to convert the transcripts to the SMIL format (Syncronized Mulitmedia Integration Language), a W3C recommendation, or the SAMI format proposed by Microsoft (Microsoft Syncronized Accessible Media Interchange).

Since we started developing our system, several open-source systems for time aligned data have been presented (ONZE 2005, Milde and Gut2002).

5. Research

Annette Myre Jørgensen has compared pragmatic markers from Buenos Aires and Madrid, they are quite different (Jørgensen forthcoming).

In her Ph.D. project *Linguistic Globalization: a Comparison between Norwegian and Chilean Adolescent Language* at the University of Bergen, Eli-Marie Drange compares the COLAs Corpus with the UNO Corpus. The aim of this study is to analyse the use of Anglicisms and other loanwords in adolescent informal language. Since this study concentrates on lexical items, these items are coded while transcribing the recorded conversations in Transcriber. When the transcriptions are made searchable through Corpus Workbench, it is possible to search for all the words that are coded. First of all, it is necessary to indicate whether a word is pronounced according to the sound system of the host language or if it has been adapted to the phonological system of the borrowing language. For this purpose, the following signs are used:

¡ = the word is pronounced with original pronunciation
! = the pronunciation is partially adapted, that is, some of the sounds are pronounced in the borrowing language

+ = the pronunciation is completely adapted to the borrowing language

The great majority of the loanwords are English loanwords, but there are also some words from other languages. The loanwords are therefore marked with a code related to their origin, using the following letters:

A = Anglicism
F = French word
I = Italian word
Q = Quechua word
Etc.

One of the aims of this study is to determine the pragmatic or communicative functions of the loanwords in the conversations. For this purpose, it is necessary to analyse each word in its context. To easy this analysis, the words are roughly divided in different categories with different codes, as indicated here:

1 = proper names: brand names, names of places, persons, etc.
2 = foreign customs: as Halloween, etc.
3 = different kinds of expressions: “I don’t know”, okei, etc.
4 = new inventions: for instance words related to new technology as internet, mail, etc.
5 = “old” loanwords: words that still maintain a foreign shape
6 = “pseudoloans”: invented words that seem to be of foreign origin, but are not
8 = other words: words and expressions that need to be analysed in more detail

Each word that is of foreign origin is coded with these codes in the following manner:

Scawm4j01: sabes que me doblé el dedo (¡A8) heavymente (¡A8) heavy
Scawm4g15: (¡A3) I’m sorry

In Corpus Workbench it is possible to narrow the search to one of these categories, and that is very useful for the purpose of the analysis.

In figure 6 there is a result of a search for all the English words of speaker scawm4j01.

The preliminary study shows that the frequency of English loanwords is low, though the loanwords used cover a wide range of functions. The most common words refer to new inventions or foreign phenomenon. It is also common to use loanwords to catch the attention of the listener for example as intensifiers or in order to soften the message. Further, English loanwords are used as euphemisms or as swearwords. The study also shows that the new words usually are phonologically and morphologically integrated in the borrowing language.

AB Stenström has looked at some of the most prominent features of the teenage language in COLT and COLAm in a contrastive perspective, namely the use of intensifiers, tags, taboo words, pragmatic markers and slang, in addition to features of politeness

***He's well nice – Es mazo majo*. London and Madrid girls' use of intensifiers.** (Stenström 2005 a) compares the use of adjective intensifiers among Spanish and English teenage girls with special emphasis on English *well* and Spanish *mazo*, which have both undergone grammaticalization. The fact that the use of *mazo* (' a lot') in this new function has not yet been documented in the dictionaries points to a very recent trend. , probably a teenage innovation. Both *mazo* and *well* belong to the ten most common adjective intensifiers in the respective language, topped by *really, very* and *bloody* in COLT and *muy* ('very'), *super* (borrowed from English) and superlative forms ending in *ísimo/a in COLAm.* Taboo words, too, are also represented among the ten most frequent intensifiers in both corpora, notably *fucking* and *joder qué* ('fucking'). Overall, intensifiers are most frequent in the Madrid girls' conversations, which seems to indicate that the Spanish girls are more keen to show their attitude to what was being said and use more empathy than the English girls.

Taboo words in teenage talk. London and Madrid girls' conversations compared (Stenström in press b). Teenagers' use of taboo words in a contrastive perspective is a neglected area. This study, which compares the use of taboo words among middle/upper class teenage girls in COLT and COLAm, shows that taboo words are more often used by the English than by the Spanish girls. The taboo words used belong to the same domains, sex and excretion. The most popular ones turned out to be *fuck* and *joder*, which have the same meaning, but while *joder* is only used as a verb or as an interjection, *fuck,* which is extremely productive, with its inflected forms *fucked* and *fucking,* is used more widely. The second most common taboo words are *dick* and *coño* ('cunt'), both of which are used for both sexes despite their male vs female reference. Interestingly, the Spanish girls avoid the word *dios* ('god') altogether, while *god* is a frequent word in the English girls' talk. Generally speaking, taboo words are not offensive in teenage talk. They have an interactional function, intensifying the contact between the speakers and adding extra focus to what is being said.. .

Three studies are devoted to pragmatic markers, which prove to be particularly frequent in teenage talk. The markers described here are *pues, o sea* and *en plan* and their English equivalents (Stenström in press, b, c, forthcoming b).

Spanish *pues* serves a wider range of functions than corresponding English markers. It can be translated by *well* when functioning as a turn-taker, conversational restarter, discourse organizer, filler, topic transition marker, thematic link, a question initiator and as response initiator; by *cos* especially when used as a causal connector, but also as a discourse organizer, a thematic link and a reinforcing marker. In its role as a consecutive connector and punctuation marker, it corresponds to other markers in English, for instance *then* or *therefore*. But there are also cases where there is no correspondence in English. All in all, *well* was found to be the nearest English correspondence in the majority of cases, *cos* in a few cases and a different marker or none at all in other cases.

O sea has a narrower range of functions than *pues*, but like *pues* it is used on the discursive level, clarifying an utterance corresponding to *that is*, marking background corresponding to *cos,* concluding corresponding to *so,* organizing corresponding to *well*; on the pragmatic level as a hedging device, modifying the speasker's commitment to what is being said, corresponding to *I think, sort of,* a quotative device introducing direct speech, corresponding to BE *like*; and on the interactional level corresponding to eg *that is* and *in other words*..

Like *o sea, en plan* is used both as a hedge and as a quotative device corresponding to English BE *like*, which is very frequent in COLT. What makes *en plan* particularly interesting is the fact that, like *mazo*, it has not yet been observed in the linguistic literature; nor is it mentioned in dictionaries. It is apparently an innovation, which is just beginning to appear in the language of teenagers. So far, *en plan* is less frequent than the corresponding *like* (hedge) and BE *like* (quotative) in COLT. Unlike *like,* which is mostly found in clause-medial position, *en plan* is commonly found in clause-initial and even clause-final position.

It's very good eh – Está muy bien eh. **Teenagers' use of tags: London and Madrid compared** (Stenström in press e). Tags are generally defined as tag-questions, ie short phrases/clause added to the end of a sentence/statement, which make it a question or appeal for agreement. However, tags have additional functions in teenage language.They often have a reinforcing effect, or the opposite, a softening effect. And they are often used as a kind of punctuation marker. Two of the five most frequent tags together dominate largely in both corpora, Spanish *eh* and *no* (94%) and English *yeah* and *right* (91%). The Spanish tag *sabes* ('you know') turned out to be relatively more common than English *you know.* It should be noticed that there is no direct equivalent to English *yeah* in Spanish, nor to Spanish *no* in English. *Yeah* is often found to serve as a kind of punctuation marker, marking boundaries in the discourse, a function that is said to be realized by *pues* in Spanish. More than half of the Spanish *eh*-tags were found within the turn, with an expressive-phatic function, while the English *eh*-tags were generally found in final position, appealing for feedback or yielding the turn. The appealing force of the turn-final was found to vary from very weak to very strong. When uttered in a separate tone unit in final position, the tag seems to function as a question, ie a speech act.

A matter of politeness? A contrastive study of phatic talk in teenage conversation (Stenström and Jørgensen forthcoming). Phatic talk was originally defined as loose talk with no informative value, the function of which is to establish and maintain social interaction. Such talk has a wide range of realizations, from entire encounters (such as chats) to connective linkers, and verbal fillers (*oh, er*) appellatives and insults (in their capacity as communicative devices). The starting point for this study was Leech's Phatic Maxim 'keep talking'. The question was to what extent the phatic strategies adopted by the teenagers in COLT and COLAm can be characterized as polite. The fact that everybody knew that they were being recorded often had the effect that entire recordings can best be described as phatic, ie loose talk for the sake of talking. Conventionalized phatic words or standardized expressions are rare, but appealers for feedback and reaction signals are common, in COLT realized by *right* and *mm/mhm*; in COLAm

by *no* and *sí*. One gets the impression that the turns are shorter, that there are more interruptions, that there is more laughter and more use of appellatives and taboo words in the Spanish than in the English conversations. The question is whether this points to stronger involvement on the part of the Spanish teenagers, or is a cultural feature. The study shows that the teenagers' use of expressions that have been criticized by adults is highly motivated for phatic purposes and can rightfully be considered to represent polite behaviour, notably the use of encouraging feedback and reaction signals, face-saving hedges, the macro-structural pragmatic markers and the use of rapport-creating taboo words.

A comparative study of London and Madrid slang (Stenström forthcoming a), which concentrates on taboo slang in COLT and COLAm indicates that taboo words are twice as frequent in the Spanish teenagers' talk as in the English teenagers' talk. This unfinished paper will consist of three parts, a short discussion about the problems with defining slang is followed by a description of the use of taboo slang in COLT and COLAm, with special emphasis on what types of taboo words are used and to what effect.

Several Master thesis are based on the COLA Corpus, and some Doctoral studies are in process.

6. References

Web site of project: http://www.colam.tk/

Barras, C., Geoffrois, E., Wu, Z. and Liberman, M. (1998) Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. *First International Conference on Language Resources and Evaluation (LREC), Granada, 1994* (Paris: ELRA), 1373-1376. Availabe on-line from http://www.ldc.upenn.edu/mirror/Transcriber/articles/Transcriber-LREC1998.pdf (accessed June 21st, 2005)

Christ, O. (1994) A modular and flexible architecture for an integrated corpus query system. *COMPLEX'94, Budapest, 1994.* Available on-line from http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ:complex94.ps.gz (accessed June 21st, 2005)

Crowdy, S. (1995). The BNC spoken corpus. In G. N. Leech, G. Myers and J. Thomas (eds.) *Spoken English on Computer: Transcription, Mark-up and Application*, (New York: Longman) 224-234.

Drange, E.-M. (forthcoming) La globalización lingüística: Una comparación entre el lenguaje juvenil noruego y chileno. Doctoral thesis, (Bergen: University of Bergen).

Garg, S., Martinovski, B., Robinson, S., Stephan, J., Tetreault, J. and Traum, D. (2004) Evaluation of Transcription and Annotation tools for a Multi-modal, Multi-party dialogue corpus. *Proceedings of Fourth International Conference on Language Resources and Evaluation, Lisbon, 2004* (Paris: ELRA), 2163-2166. Available online from http://www.ict.usc.edu/~traum/Papers/tools6.pdf (accessed June 21st, 2005)
Haslerud, V. and Stenström, A-B. (1995). The Bergen Corpus of London Teenager Language (COLT) In G. N. Leech, G. Myers and J. Thomas (eds.) *Spoken English on Computer: Transcription, Mark-up and Application*, (New York: Longman) .

Hofland, K. (2003) A web-based concordance system for spoken language corpora, in D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.) *Proceedings of the Corpus Linguistics 2003 conference*. (Lancaster: Computing Department), 330-331.

Jørgensen, A. M. (Forthcoming) Los marcadores pragmaticos del habla juvenil madrileño y bonaerense.
Labov, W. (1972) *Sociolinguistic Patterns* (Philadelphia: University of Pennsylvania Press)

LDC (2005) Linguistic Annotation. Available online from http://www.ldc.upenn.edu/annotation/ (accessed June 21st, 2005)

Martin, P. (2004) WinPitch Corpus. A text to Speech Alignment Tool for Multimodal Corpora, *Proceedings of Fourth International Conference on Language Resources and Evaluation, Lisbon, 2004* (Paris: ELRA), 537-540. Available on-line from http://lablita.dit.unifi.it/coralrom/papers/Philippe_Martin.pdf (accessed June 21st, 2005)

Milde, J.-T. and Gut, U. B. (2002). The TASX-environment: an XML-based toolset for time aligned speech corpora. *Proceedings of the third international conference on language resources and evaluation (LREC 2002), Gran Canaria*. Available online from http://coli.lili.uni-bielefeld.de/Texttechnologie/Forschergruppe/pdfs/tasx-environment.pdf (accessed June 21st, 2005)

Minidisc.org. (2005) Minidisc FAQ: Hi-MD Topics. Available on-line from http://www.minidisc.org/hi-md_faq.html (accessed June 21st, 2005)

ONZE (2005) ONZE Miner. Available on-line from http://www.ling.canterbury.ac.nz/jen/onzeminer/ (accessed June 21st, 2005)

Schmidt, Thomas (2004): Transcribing and annotating spoken language with EXMARaLDA. *Proceedings of the LREC-Workshop on XML based richly annotated corpora*, *Lisbon 2004*. (Paris: ELRA). Available on-line from http://www.rrz.uni-hamburg.de/exmaralda/Daten/4D-Literatur/Paper_LREC.pdf (accessed June 21st, 2005)

Stenström, A-B. (2005 a) *He's well nice – Es mazo majo*. London and Madrid girls' use of intensifiers. In S. Granath, J. Millander and E. Wennö (eds). *The Power of Words. Studies in Honour of Moira Linnarud.* (Karlstad: Karlstad University)

Stenström, A-B. (In press a) Teenage talk: A chat and discussion compared. In R. Horowitz (ed). *Talking Texts*. (New Jersey: Lawrence Erlbaum)

Stenström, A.-B. (In press b) Taboo words in teenage talk: London and Madrid girls' conversations compared. To appear in *Spanish in Context* Vol 3. 2006.

Stenström, A.-B. (In press c) The Spanish pragmatic marker *pues* and its English equivalents. To appear in *Proceedings from ICAME 2003, Guernsey*.

Stenström, A.-B. (In press d) The Spanish discourse markers *o sea* and *pues* and their English correspondences. To appear in proceedings from *Pragmatic Markers in Contrast. (Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten), Brussels, 22-23 May 2003.*

Stenström, A.-B. (In press e) *It is very good eh – Está muy bien eh*. Teengaers' use of tags – London and Madrid compared. Festschrift article.

Stenström, A.-B. (Forthcoming a) A comparative study of London and Madrid slang. for 'Slang: a linguistic universal? *International Slang Workshop. Budapest 14-16 November 2005.*

Stenström, A.-B. (Forthcoming b) Pragmatic markers in contrast: Spanish *en plan* and English *like* in teenage talk. To be presented at *Congreso del Españool en la Sociedad. University of Strathclyde, Glasgow, April 2006.*

Stenström, A.-B. and Jörgensen, A. M. (Forcoming) A matter of politeness? A contrastive study of phatic talk in teenage conversation. To be presented at the *9th International Pragmatics Conference. Riva del Garda, Italy.*

Stenström, A.-B. and Aijmer, K. (eds) (In press) Approaches to spoken interaction. In Special Issue of *Journal of Pragmatics.* (Exeter: Elsevier)

Stockdale, A. (2004) An Approach to Recording, Transcribing, and Preparing Audio Data for Qualitative Analysis. (Newton: Educational Development Center). Available on-line http://caepp.edc.org/QualAudio.pdf (accessed June 21st, 2005)

7. Figures

The figures for this document is missing.