COLA: UN CORPUS ORAL DE LENGUAJE ADOLESCENTE

ANNETTE MYRE JØRGENSEN

Universidad de Bergen, Noruega

0. INTRODUCCIÓN

El corpus COLA es una base de datos de habla informal, transcrita y acoplada al sonido accesible en Internet para la investigación. En las líneas expuestas a continuación se describe el corpus oral COLA; Corpus Oral de Lenguaje Adolescente. Trataré el modo en el que se ha procedido para grabar y transcribir los datos así como sus contenidos, las investigaciones realizadas en español y en inglés utilizando este corpus.

El proyecto de la construcción de un corpus oral juvenil, COLA, empezó en el año 2001. Ha sido subvencionado por la Facultad de Humanidades de la Universidad de Bergen y por la fundación Meltzer de la misma universidad.

1. EL CORPUS COLA

Los objetivos del proyecto COLA (www.colam.org) es reunir el lenguaje juvenil hablado informal de Madrid y de otras capitales de habla española, construyendo un corpus para promover la investigación del lenguaje y el estilo comunicativo de los jóvenes de habla española, como, por ejemplo la influencia de otros idiomas en el lenguaje juvenil. Dado que el corpus COLA sigue el mismo patrón que los corpora COLT y UNO, se pueden realizar análisis contrastivos entre los diferentes idiomas inglés, español y noruego. El material recogido es especialmente apto para el análisis del léxico particular, propio del lenguaje juvenil. Se pueden hacer estudios sociolingüísticos también.

El corpus COLA, que, por el momento, está en fase de construcciòn, consta de varios subapartados, dependiendo de las capitales en las que se realizan grabaciones. Tiene actualmente alrededor de 350.000 palabras transcritas. De Madrid, Colam, hay 175.000 palabras, Santiago de Chile, Colas, hay 150.000, Buenos Aires, Colaba, 30.000 palabras transcritas. Actualmente se está ampliando el corpus COLA con material de La Habana y en Guatemala ciudad.

El habla transcrita de Madrid y de Santiago de Chile se halla a disposición de quien lo quiera investigar en: http://gandalf.aksis.uib.no/cola/

1.1. ¿Por qué estudiar el lenguaje juvenil?

La juventud constituye un grupo social nuevo y es el grupo social que marca los ideales que pràcticamente todo el mundo quiere imitar (Jørgensen 2006: 2) Se ha estudiado desde muchos ángulos, pero no desde el lingüístico (Rodríguez, 2002: 14). Los jóvenes tienen un argot propio, consecuencia de su búsqueda de identidad. Las palabras nuevas entran con gran facilidad en el lenguaje juvenil, y pasan luego al lenguaje adulto. (Zimmermann 2002: 138)

Los jóvenes no respetan la normativa como lo hace un adulto. (Rodríguez, 2002: 22). En añadidura a esto, los jóvenes tienen un situación comunicativa especial: su inseguridad y cambiante competencia comunicativa conlleva uso frecuente de: malas palabras, palabras tabús, marcadores del discurso, (Rodríguez, 2002: 24, Stenström 2005: 1). La influencia del inglès es patente en el lenguaje y es interesante ver el impacto que tiene en el juvenil. (Lorenzo, 1996: Rodriguez, 2002)

1.2. Métodos de recolección de datos del corpus

Los métodos de recolección de datos siguen los del modelo COLT, el Corpus de lenguaje juvenil inglés (www.uib.hf.aksis/colt), y el corpus UNO de lenguaje juvenil noruego (www.uib.hf.aksis/uno). Estos corpora a su vez ha usado el modelo Longman para la recolección de los datos del Corpus Nacional Británico BNC, (British National Corpus) (Haslerud 1995:235, Crowdy, 1995).

Los *reclutas*, así denominamos a los chicos reclutados para llevar el minidisk y de hacer la grabación, siguiendo el ejemplo de la terminología del corpus COLT, son escogidos de distintos colegios, de diferentes ambientes sociales de las capitales en cuestión, en orden a tener datos equilibrados en cuanto a género, edad y nivel social. Los reclutas tienen entre 13 y 19 años[1].

Las grabaciones se han hecho con una grabadora SONY-Net, minidisco, MD, con un micrófono de solapa. Se les pide a los jóvenes grabar su charla informal con sus compañeros durante tres o cuatro días, sin presencia de adultos. Algunas conversaciones se han grabado en colegios, en los recreos, en casa, otras en lugares diversos como en parques, por la calle, en el metro, etc. Los reclutas han rellenado cuestionarios con datos sobre lugar de nacimiento, lengua de los padres e idioma hablado en casa, etc., y, también han recogido los datos de los compañeros que participaban en la conversación de la cinta.

El método de recopilación mencionado resulta muy eficaz a la hora de grabar conversaciones informales, ya que se evita la “paradoja del observador” que W. Labov formula asi: *detectar como habla la gente cuando no esta siendo observada; cuando solo es mediante la observación que uno lo puede llegar a saber*. (Labov 1972:209)

En este momento, más de 50 reclutas de diferente sexo, edad y clase social han participado en el proyecto COLA, y se han grabado 76 horas de conversación con amigos y conocidos en Madrid, Buenos Aires y Santiago de Chile.

1.3. Dificultades en la recopilación del corpus Cola

No siempre se logra obtener los permisos necesarios para realizar las grabaciones según los criterios éticos establecidos en Noruega por la NSD[2] para las grabaciones. Estos criterios prohiben las grabaciones del habla sin informar al interesado. Se exige permiso escrito por parte del hablante (de sus padres si son menores de edad), así como garantizar la anonimidad a los que se grabe, cambiando los nombres de las personas y de los lugares propios. A pesar de ello, suele haber jóvenes que no reciben el permiso de sus para hacer las grabaciones.

Las chicas que no tienen problema alguno en ponerse a hablar largo y tendido. Sin embargo, esto no les es natural a los chicos más jóvenes. Un joven noruego comenta:

det pleier egentlig vi aldri å gjøre, bare sitte og snakke (ØSVGGU1)
es algo que no hacemos nunca, lo de estar asi como hablando

Puede ser una dificultad el hecho de que los jóvenes sepan que estàn siendo grabados. Los diferentes temas de conversación y el modo casual de hablar sugiere que la presencia del micrófono no influye en la conversación de los jóvenes. Si hacen referencia al micrófono a lo largo de sus conversaciones, solamente es para ver si funciona.

2. LAS GRABACIONES Y LAS TRANSCRIPCIONES INFORMATIZADAS

Las conversaciones grabadas se pasan de los minidiscos a discos CD. Para la transcripción del sonido en el CD elegimos el programa Transcriber. (Barras et al. 1998). El programa Transcriber transcribe tiene un comando especial para marcar el habla solapada, un comando “insertar evento” se usa para indicar información extra lingüística del modo de enunciar el texto como riéndose, gritando, cantando, etc.

Cada hablante recibe su propio còdigo en el que se hallan los datos de gènero, edad y clase social.

El material transcrito està sincronizado con el archivo de sonido cada vez que se de a la tecla de return/enter. Cuando se pone en marcha el archivo del sonido, el archivo del texto baja en consonancia con este, y, cuando se cliquea el texto, sale el sonido.

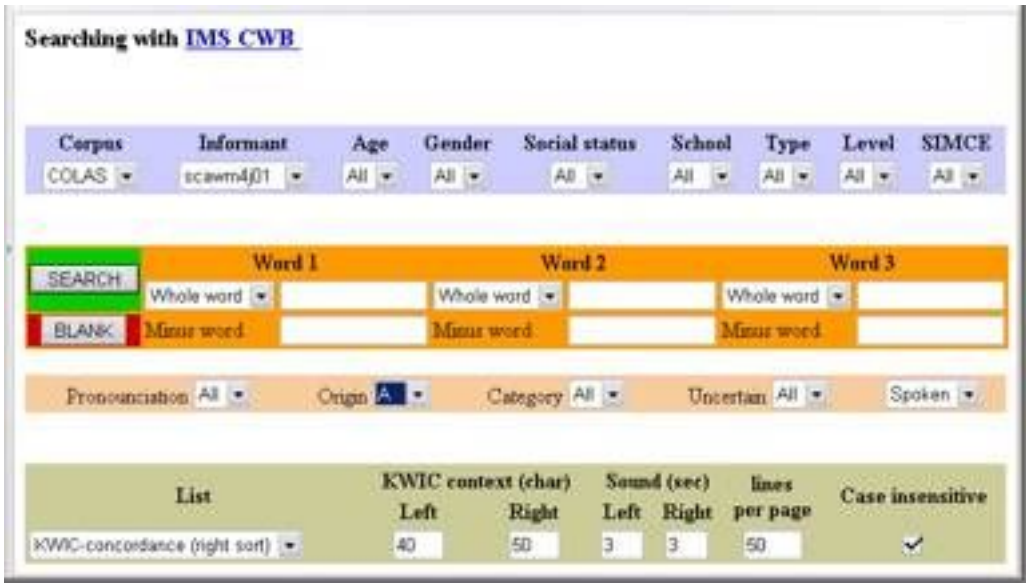
Se muestra una parte de la pantalla en la imagen que sigue:

2.1. Las búsquedas y las variables del material

La transcripción y los archivos de sonido están insertados en un sistema Web hecho para el proyecto COLT y COLA (Hofland 2003). La transcripción se convierte en un archivo HTML, con enlaces a fragmentos de sonido. Un ejemplo del archivo HTML se halla en siguiente imagen:

T	L	10	scawm4j01:	aaaaayyyy (risa/)
T	L	10	scawm4jx2:	sabes que me doblé el dedo pero
T	L	10	scawm4j01:	(¡A8) heavymente (¡A8) heavy
T	L	10	scawm4jx1:	1[ese/]
			scawm4jx2:	1[este]
T	L	10	scawm4jx1:	2[lo tienes hinchado]
			scawm4jx2:	2[sí]
T	L	10	scawm4j01:	como morado
T	L	10		(risa/)

Hay un archivo con la información de cada hablante que da la edad, sexo y nivel social, etc. Las transcripciones y la información mencionada están convertidas en un formato Corpus Workbench (Christ 1994). Tenemos una imagen en pantalla para el programa de búsqueda:



El texto y el sonido están sincronizados en varios puntos/momentos. Los fragmentos de sonido de la concordancia se pueden presentar en archivos WAV o MP3.

2.2. Las concordancias

Hay una versión especial donde el usuario puede hacer una clasificación de cada línea de concordancia:

Los textos han sido convertidos al formato usado por la versión 4 de Wordsmith, que es un programa único de concordancia de ventanas. Esta versión puede emitir el sonido de la concordancia y también tener un mejor soporte para los textos descifrados XML, aunque las facilidades de la búsqueda no son tan flexibles como con el Corpus Workbench.

3. LAS INVESTIGACIONES CON EL CORPUS COLA

La Dra. Anna-Brita Stenström compara los marcadores pragmáticos y las palabras tabùs inglesas y españolas. Hay gran creatividad en cuanto a los marcadores pragmáticos, cosa que se ha corroborado tento en el materal inglés como en el noruego, y los marcadores adquieren diferentes funciones en la oración. Ocurre lo mismo en español y en inglés. La tendencia es que hay màs palabras malsonantes, risas e interrupciones en el material español que en el inglès. También ha analizado el modo en que se dan retroalimentación los jóvenes ingleses y españoles, y resulta que

La Dra. Annette Myre Jørgensen compara marcadores pragmáticos del lenguaje juvenil de Madrid con el de Buenos Aires y ha llegado a la conclusión de que son muy diferentes. En Madrid se usan las siguientes palabras como marcadores: *tío, tía, tronco, tronca, sabes, pues, eso, vale, o sea, y tal, entonces,* etc. En Buenos Aires, sin embargo, se usan los siguientes marcadores: *boludo, nene, viste, ché, este, ay, pucha, chico, chica*.

La Dra. Anna-Brita Stenström y la Dra. Annette Myre Jørgensen han hecho un estudio de la cortesia analizando el habla juvenil de Londres y Madrid en el artìculo "¿Una cuestiòn de cortesía? Estudio contrastivo del lenguaje fàtico en la conversaciòn juvenil." En ambos corpora consta que las palabras consideradas como insultos, no lo son en el habla juvenil. Tienen una función fática, y por lo tanto de cortesìa.

En su proyecto de tesis doctoral en la Universidad de Bergen, *Globalización linguistica: una comparación entre el lenguaje juvenil chileno y noruego*, Eli-Marie Drange compara el uso de anglicismos en los corpora COLA con el UNO. El objeto de este estudio es una preocupación global, la influencia del inglés. Muchas palabras inglesas se introducen en la lengua por la música, las películas, la tecnología y las noticias. Por ello, analiza el uso de anglicismos y otros prestamos en el lenguaje informal juvenil.**:**

1. íbamos a chatear allá (#1-1-1m)

1. dice *ah I love you* (#2-2-1m)

Llega a una inesperada conclusión: *En mi material no he encontrado muchos extranjerismos al igual que: Gómez Capuz 2001, Sharp 2001, Graedler & Johanson 2002*. (Drange 2002,)

Se están haciendo Tesis de Máster sobre *en plan, o sea* como marcadores pragmáticos, las palabras tabùs, las interupciones, etc., así como estudios de cambio de código, *codeswitching* en el lenguaje juvenil y la presencia de anglicismos.

BIBLIOGRAFIA:

La página web de las transcripciones de COLA: http://gandalf.aksis.uib.no/cola/

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.. 2005. La globalización lingüística: una comparación entre el lenguaje juvenil noruego y chileno. *Interlingüística XV*. Valencia: Universidad de Valencia. 373–380.

DRANGE, E.-M. (forthcoming) La globalización lingüística: Una comparación entre el lenguaje juvenil noruego y chileno. Tesis doctoral, Universidad de 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.

HOFLAND, K. , JØRGENSEN, A. M., DRANGE, E.M., (Forthcoming) Proceedings of the CL 2005 Conference, Birmingham

JØRGENSEN, A. M. (Forthcoming) "Los marcadores pragmaticos del habla juvenil madrileño y bonaerense".

JØRGENSEN, A. M. (2004) Cola-prosjektet: "En korpusbasert undersøkelse av spansk tenåringsspråk." *Tribune* 15. Universitetet i Bergen.

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)

LORENZO, E.: (1996) *Los anglicismos hispànicos*. Madrid. Gredos.

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)

RODRIGUEZ, F. (2002) *El lenguaje de los jòvenes*. Madrid, Ariel

RODRIGUEZ, F. (2002) *Comunicaciòn y cultura juvenil*. Madrid, Ariel

SCHMIDT, T. (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. 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.2005 b *It is very good eh – Está muy bien eh*. Teenagers' use of tags – London and Madrid compared. In K Mc Cafferty, T. Bull and K. Killie (eds). *Contexts – Historical, Social, Linguistic. Studies in Celebration of Toril Swan.* Pieterlen: Peter Lang AG.

STENSTRÖM, A-B. (2005 c) with K. Aijmer (eds). Approaches to spoken interaction. Special Issue of *Journal of Pragmatics.* Elsevier.

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

STENSTRÖM, A-B. (In press c) The Spanish discourse markers *o sea* and *pues* and their English correspondences. To appear in K. Aijmer and A-M. Simon-Vandenbergen (eds). *Pragmatic Markers in Contrast.* Elsevier.

STENSTRÖM, A-B. (In press d) Teenage talk: A London-based chat and discussion compared. In R. Horowitz (ed). *Talking Texts: How Speech and Writing Interact in School Learning.* New Jersey: Lawrence Erlbaum.

STENSTRÖM, A-B. (Forthcoming b) "Teenagers' use of taboo slang. London and Madrid teenagers compared. Slang: a linguistic universal?" International Slang Conference Budapest 14-16 November 2005.

STENSTRÖM, A-B. and JØRGENSEN, A. M (Forcoming) "¿Una cuestiòn de cortesía? Estudio contrastivo del lenguaje fàtico en la conversaciòn juvenil." *Journal of Pragmatics.*

STENSTRÖM, A-B and JØRGENSEN, A. M. (Forthcoming) "A matter of politeness? A contrastive study of phatic talk in teenage conversation." *Proceedings from the 9th IPRA conference*.

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/borrar/ (accessed June 21st, 2005)

ZIMMERMANN, K. (2002): La variedad juvenil y la interacción verbal entre jóvenes." en: Rodríguez, F. (2002): *El lenguaje de los jóvenes*, Barcelona, Ariel

---

[1] La Organización de Naciones Unidas (ONU) emplea una definición entre 15 y 24, donde distingue los adolescentes (13-19) a los adultos jóvenes (19-24)[1]. COLAm opera con una definición de 13 a 19 años.

[2] Norsk Samfunnsvitenskapelig Datatjeneste. Servicio Social Noruego de Datos informáticos.