La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM

Antonio Moreno Sandoval

Laboratorio de Lingüística Informática
Universidad Autónoma de Madrid
sandoval@maria.lllf.uam.es

Resumen

El Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid (LLI-UAM) tiene entre sus objetivos la creación de recursos lingüísticos en formato electrónico: gramáticas y lexicones computacionales, bases de datos terminológicas, catálogos digitalizados y todo tipo de corpus, orales y escritos, diacrónicos y sincrónicos, monolingües y multilingües. Esta comunicación se centrará en la comparación entre el primer corpus que desarrollamos a principios de los 90 (CORLEC) y el que estamos recopilando actualmente (C-ORAL-ROM). La comparación servirá de base para establecer la evolución que se ha producido tanto en la metodología como en el formato y contenido de los corpus de habla espontánea.

1. Perspectiva histórica

El Corpus Oral de Referencia de la Lengua Española (CORLEC) fue el primer corpus de habla espontánea del español [1], [2], financiado parcialmente por IBM, con un presupuesto muy bajo (5 millones de pesetas). Se realizó entre 1991 y 1992 y tomó como esquema de codificación el propuesto por la TEI en esos años.

C-ORAL-ROM [3] es un corpus multilingüe (italiano, francés, portugués y castellano) desarrollado por un consorcio de 9 socios y financiado por la Unión Europea dentro de su V Programa Marco. El presupuesto del equipo español es cercano a los 35 millones de pesetas y se extiende a lo largo de 3 años (2001-2003). El LLI-UAM fue elegido para realizar el corpus español precisamente por su experiencia en CORLEC. El proyecto ha desarrollado su propio esquema de codificación a partir de los formatos originales de cada equipo. Emplea XML como lenguaje de marcación e intercambio. Además se proporcionará el alineamiento del texto con el sonido fuente, así como distintos niveles de etiquetado lingüístico y algunos estudios cuantitativos comparativos de las cuatro lenguas romances.

Ambos corpus proporcionan una transcripción ortográfica siguiendo las convenciones propias de la lengua española, pero en la que se registran las elisiones de sonidos, errores de pronunciación u otros casos similares frecuentes en la lengua oral. La característica más importante es la espontaneidad: los textos han sido grabados en contextos reales y sin ningún guión preestablecido.

2. Características de CORLEC

El CORLEC es una base de datos que contiene aproximadamente 1.100.000 palabras transliteradas, a partir de textos orales espontáneos grabados en cintas de audio analógicas. Los medios técnicos del momento (1991-1992) no le permitían a un proyecto con un presupuesto tan bajo utilizar tecnología digital para las grabaciones. Tampoco se dispuso de ningún programa informático de tratamiento de sonido. Los transcriptores trabajaban con un reproductor analógico y unos auriculares y transcribían directamente en un procesador de textos (WordPerfect). Entre los objetivos del proyecto nunca se planteó el alineamiento de sonido y transcripción, sino registrar por primera vez de manera exhaustiva la variedad oral del español. La calidad acústica de las cintas es deficiente en muchos casos. En la actualidad, disponemos de una copia digitalizada de los originales analógicos[1].

2.1. Criterios de transcripción

- La *fidelidad* a lo que dicen los hablantes: los segmentos fonéticos suprimidos, las interrupciones, las realizaciones repetidas, las autocorrecciones, las palabras inventadas o las palabras de otras lenguas se transcriben tal y como las pronuncia el hablante. Para recuperar la forma canónica, todos estos casos se anotan con etiquetas pertinentes.
- Utilización de signos de puntuación (comillas, puntos suspensivos, punto y aparte, etc.) para marcar situaciones discursivas. Las comillas se emplean para discurso directo, para resaltar palabras y para marcar títulos. Los puntos suspensivos sirven para marcar pausas, vacilaciones, cortes bruscos. Las comas y el punto y aparte se utilizan como marcadores de las unidades sintácticas.
- El transcriptor debe seguir las normas ortográficas para los textos escritos, "y habrá de ser marcada aunque potencialmente el hablante no se detenga" [1]. Probablemente esta decisión sea la más contradictoria con el primer requisito: por una parte se es fiel a la pronunciación; por otro lado, se siguen las convenciones ortográficas en cuanto a la sintaxis de la lengua escrita.

2.2. Formato de la transcripción

El texto se divide en *cabecera* y *transcripción* propiamente dicha. La cabecera identifica cada unos de los hablantes, con sus características sociolingüísticas (sexo, edad, profesión, nivel de estudios, procedencia) así como la localización y la

[1] Los interesados deben ponerse en contacto con francisco.marcos.marin@uam.es

fecha de la grabación, la fuente (en el caso de textos tomados de la radio o la televisión), la cinta donde está el sonido original, la clase de texto y el transcriptor.

La información proporcionada en la transcripción está enriquecida con todo tipo de etiquetas sobre elementos fáticos (sonidos emitidos por los hablantes que se interpretan como afirmaciones, interrogaciones, etc.), sobre ruidos (risas, aplausos, música, etc.) y especialmente las interacciones discursivas: se marcan los turnos de palabra y el solapamiento de los hablantes.

El sistema de codificación empleado sigue las normas TEI, aunque debido a que cuando se recopiló el corpus todavía estas [4] no habían sido publicadas oficialmente, la interpretación por parte de los autores presenta decisiones particulares. Hay que señalar, además, que el corpus no ha pasado por ningún proceso de validación de su formato (por ejemplo, validación mediante una DTD)[2].

2.3. Distribución del corpus

La distribución de los distintos tipos de textos es uno de los componentes definitorios de cualquier corpus. CORLEC está organizado por criterios temáticos (Tabla 1). Destacan los textos periodísticos y familiares, con un 38,6 % y un 24,5 % respectivamente. Sin embargo, no es representativo para la distribución ni el tipo discursivo (monólogo frente a diálogo/conversación) ni la fuente (grabación directa, a través de los medios de comunicación, telefónica). Tampoco es significativa la distinción esencial en la lengua oral entre el registro informal y el registro formal.

Tabla 1: Distribución de CORLEC

Clases	Nº Palabras	Porcentaje
Administrativos y políticos	61.200	5,6 %
Científicos	36.600	3,3 %
Familiares	269.500	24,5 %
Educativos	58.300	5,3 %
Humanísticos	61.200	5,6 %
Instrucciones (megafonía)	6.600	0,6 %
Jurídicos	35.200	3,2 %
Lúdicos (concursos, etc.)	61.200	5,6 %
Periodísticos		
Debates	93.500	8,5 %
Deportes	58.300	5,3 %
Documentales	28.600	2,6 %
Entrevistas	171.200	15,6 %
Noticiario	72.600	6,6 %
Publicitarios	30.800	2,8 %
Religiosos	12.100	1,1 %
Técnicos	43.100	3,9 %
TOTAL ESTIMADO	1.100.000	100 %

[2] Posteriormente, en 1997, el corpus fue revisado y codificado en SGML por un equipo de la Real Academia Española, que incorporó el CORLEC a la parte oral del CREA. En la actualidad, en el LLI estamos convirtiendo el CORLEC a un formato muy similar a C-Oral-Rom, en XML y validado con una DTD propia.

El corpus (la transcripción) está accesible, de manera gratuita, a través del servidor ftp del LLI-UAM (ftp://ftp.lllf.uam.es/pub/corpus/oral).

3. Características de C-Oral-Rom

Cada uno de los cuatro corpus que componen C-Oral-Rom consta de 300.000 palabras y los textos se organizan en función de una distribución acordada para permitir la *comparabilidad* entre las cuatro lenguas.

3.1. La comparabilidad entre las cuatro lenguas

El carácter multilingüe de C-Oral-Rom es uno de sus rasgos distintivos. La interacción entre las diferentes tradiciones en la recopilación de corpus orales, además de las ricas experiencias de cada grupo se concretan en su lado práctico en la necesidad de llegar a un acuerdo en dos puntos esenciales: la distribución del corpus y el formato unificado para la descripción del corpus.

3.1.1. La distribución de los textos

Conseguir una distribución similar en los diferentes corpus es esencial para poder hacer estudios comparables entre lenguas. A la dificultad de diseñar una distribución significativa en un corpus oral, debido a la inherente variabilidad del habla espontánea, hay que unir la dificultad de conjuntar distintas tradiciones en la transcripción.

En C-Oral-Rom se ha llegado al siguiente compromiso: los dos factores que influyen decisivamente en la variación son las *características de los hablantes* y el *contexto de uso* (no la temática, como era el eje organizador de CORLEC). Cada uno de estos dos parámetros se define por una serie de rasgos. Los hablantes se caracterizan según el sexo, la edad, la educación, la profesión y el lugar de origen. En el contexto de uso, distinguimos, por una parte, la estructura dialógica (es decir, monólogos frente a diálogos/conversaciones). Por otra parte, el dominio privado o público. Por último, diferenciamos las producciones tomadas de medios de comunicación y a través del teléfono.

Si cada uno de estos parámetros de variación está bien representado en el corpus, podemos considerar que la variación lingüística está bien recogida. Para nosotros, la distinción esencial es *habla formal* frente a *habla informal*. Cada uno de ellos está representado por el 50 % de los textos. Dentro del habla informal, incluimos un porcentaje determinado para los monólogos y para los diálogos/conversaciones. Por su parte, el criterio temático lo hemos incluido solo para crear una subclasificación dentro de las grabaciones de los medios. Finalmente, asignamos un pequeño fragmento del corpus a las grabaciones telefónicas, que son un subtipo en sí mismo por la naturaleza del medio. La distribución acordada para los cuatro corpus es la que se muestra en la Tabla 2.

Tabla 2: Distribución de C-Oral-Rom

Informal 150.000 pal.		Formal 150.000 pal
Familiar 113.000	Público 37.000	Formal en contexto natural 65.000

Mono	Conv	Mono	Conv	Formal en los medios 60.000 Grabaciones telefónicas 25.000
33.000	80.000	6000	31000	

3.1.2. El formato común

El otro aspecto que se necesita para conseguir la comparabilidad es emplear el mismo esquema de anotación. El consorcio ha acordado un formato C-Oral-Rom, que se basa esencialmente en el modelo italiano (cuyo origen es el formato CHAT). Para conseguir la plena reutilización de la transcripción, se utiliza XML como lenguaje de marcación, lo que garantiza la fácil interpretación mediante la correspondiente DTD. El LLI-UAM ha desarrollado para el proyecto el programa de conversión del formato C-ORAL-ROM a la versión en xml.

3.2. Otros aspectos

C-Oral-Rom presenta una serie de innovaciones muy importantes con respecto a su predecesor CORLEC. Estas innovaciones son producto de la evolución de los requisitos en la elaboración de recursos lingüísticos, que afectan muy especialmente a los corpus de habla espontánea.

3.2.1. La legalidad

A mediados de los 90 la legislación sobre Derechos de Autor y sobre la Intimidad cambió en los países europeos. Esto nos afecta y obliga directamente cuando realizamos grabaciones de individuos o utilizamos documentos sonoros tomados de los medios de comunicación.

En unos casos, los hablantes mantienen su derecho a guardar la intimidad, ya que sus palabras van a ser transcritas y publicadas. Por tanto, estarán a disposición de la comunidad científica y docente, no solo como documento para la investigación del grupo que lo recopila. Sin embargo, para mantener el rasgo de espontaneidad, capital para nuestros objetivos, el procedimiento que se sigue es pedir el permiso por escrito de los participantes en la grabación una vez realizada. Si los hablantes se niegan a firmar su consentimiento, entonces no se utiliza la grabación. El derecho a la intimidad afecta a todas las grabaciones realizadas en un contexto familiar o privado, pero no a las obtenidas en una situación pública.

Por otra parte, muchos textos del corpus tienen derechos de autor, no sólo las grabaciones de los medios sino también las conferencias, las clases magistrales, las homilías, los consejos profesionales y cualquier otro texto oral donde el hablante proporcione ideas propias y con una estructura creativa. Todos los textos que aparecen en nuestro corpus cuentan con los permisos por escrito de sus autores o de los propietarios de los derechos.

3.2.2. La validación

Verificar la fiabilidad de los datos se ha convertido en uno de los temas de moda en los últimos años. Los usuarios de recursos lingüísticos (de la industria o de la universidad) quieren conocer cómo se han recopilado y el grado de exactitud de lo ofrecido.

C-Oral-Rom se somete a dos tipos de evaluación, una interna y otra externa.

La *validación interna* la realiza el propio equipo: cada texto recibe cinco pasadas (transcripción, revisión de la transcripción, etiquetado prosódico, revisión del etiquetado y alineamiento texto-sonido). Al menos tres lingüistas diferentes transcriben el texto.

Para verificar la consistencia entre los anotadores, se ha realizado un experimento [5] en el que se demuestra que el acuerdo a la hora de asignar etiquetas es considerable.

Por último, el programa de conversión a xml se encarga de verificar los errores en el formato (huecos en blanco, errores tipográficos, etiquetas malformadas, etc.). Por tanto, el contenido y la forma son validados de manera exhaustiva, garantizando a los usuarios del corpus que lo transcrito es fiel reflejo de lo que se dice. En este punto debemos destacar que el alineamiento del texto con el sonido es la garantía más sólida para validar un texto oral: cualquier discrepancia entre lo dicho y lo transcrito será fácilmente detectada.

La *validación externa* se realizará por expertos al final del proyecto. ELRA se encargará de auditar los datos y su distribución, de manera que las cifras que se proporcionan son reales y verificadas.

3.2.3. La calidad acústica

El corpus español de C-Oral-Rom se ha hecho completamente nuevo, aunque los otros equipos han reutilizado parte de sus textos antiguos. Las dos razones que nos empujaron a hacer grabaciones nuevas son, por una parte, que no contábamos con ningún permiso escrito de los textos de CORLEC (y nos iba a costar bastante conseguir las firmas de los hablantes) y, por otra parte, la calidad acústica de las cintas analógicas era muy deficiente.

Los textos de C-Oral-Rom han sido grabados con una grabadora digital DAT Tascam (modelo DA-P1) con dos micrófonos unidireccionales. El sonido original ha sido convertido a un fichero WAV mono, 16 bit, 22.050 Hz, a través del puerto SPDIF de un Sound Blaster Live Platinum 5.1. utilizando el software Creative Recorder. En las grabaciones en espacios públicos, cuando ha sido posible, las grabaciones se han realizado conectando la grabadora al equipo de sonido de la sala.

Las grabaciones de los medios o bien nos las han proporcionado directamente ellos o bien las hemos registrado con un ordenador conectado al receptor.

3.2.4. La anotación lingüística

Cualquier corpus aumenta su valor en función de los niveles de anotación que proporcione. La anotación de corpus de habla espontánea es algo diferente a la de los corpus escritos [6].

La diferencia no está tanto en la información que hay que anotar (categoría sintáctica, lema, rasgos morfosintácticos) como en que los recursos empleados en los corpus escritos no son tan efectivos en los corpus orales. Por ejemplo, los etiquetadores morfosintácticos (POS taggers) están entrenados sobre textos que tienen un orden sintáctico bastante estable y determinado. Sin embargo, los textos de

habla espontánea se caracterizan por una enorme flexibilidad. Además estos textos no se transcriben con las convenciones ortográficas, donde los signos de puntuación dan mucha información para desambiguar categorías.

El léxico de la lengua espontánea es más innovador y flexible. Encontramos muchas palabras que no aparecen en los diccionarios, porque son innovaciones, porque pertenecen a un registro informal o porque son simplemente errores de pronunciación.

Los objetivos de C-Oral-Rom no persiguen una anotación exhaustiva y completa sino simplemente proporcionar un primer acercamiento: lematización de las palabras con contenido léxico. Además se proporciona el etiquetado prosódico realizado a mano por los anotadores.

4. Conclusión: corpus de primera y segunda generación

Nuestra exposición ha mostrado la clara evolución en los corpus orales, partiendo de la base de que mantienen la esencia: registro del habla espontánea en su contexto de uso. Otros aspectos que se mantienen son las características del texto, que contiene no sólo la transcripción sino una cabecera con una rica información asociada.

Sin embargo, han cambiado muchas cosas, producto de la experiencia y de los avances tecnológicos, pero también del presupuesto y de la legislación. Los corpus de segunda generación tiene que proporcionar una calidad validada, tanto en la fiabilidad de la transcripción/anotación como en la fuente sonora. Además, ahora se debe exigir el alineamiento entre texto y sonido. Para que un corpus de habla espontánea pueda ser utilizado libremente por la comunidad científica (es decir, que se trate de un *corpus de referencia*) es necesario que cuente con las autorizaciones de los participantes, de manera que se preserven tanto su intimidad como sus derechos de autor.

Queremos terminar con nuestra previsión de futuro sobre estos corpus, que se concreta en dos aspectos:

1. Los corpus de habla espontánea se harán cada vez más necesarios pues el futuro de la telefonía móvil pasa por sistemas de procesamiento de habla efectivos en situaciones de uso real no restringido[3]. Para entrenar estos sistemas, hacen falta corpus ricos y variados.
2. La anotación de la información, en múltiples niveles (fonológico, morfológico, sintáctico, semántico), se convertirá en el principal rasgo diferenciador de cualquier corpus en el futuro cercano. Dado que las características del habla espontánea difieren de la lengua escrita, habrá que adaptar o construir herramientas nuevas para la anotación semiautomática, sin olvidar que toda codificación de calidad pasa por una verificación por parte de un especialista humano.

[3] Algo similar les ocurrió a los sistemas de procesamiento del lenguaje escrito cuando el acceso a enormes cantidades de información gracias a la universalización de internet obligó a desarrollar sistemas más flexibles y robustos.

La página oficial del proyecto (http://lablita.dit.unifi.it/coralrom/) contiene muestras de textos de distintos tipos en las cuatro lenguas, con la fuente sonora y la transcripción (fragmentos de unos 30 segundos).

Este trabajo ha sido parcialmente financiado con una ayuda de la CICYT (TEL-1999-1073-C02).

5. Referencias

[1] Marcos Marín, F., ``El Corpus Oral de Referencia de la Lengua Española Contemporánea", Informe del proyecto. Madrid, 1992. Accesible a través de ftp://ftp.lllf.uam.es/pub/corpus/oral.

[2] Llisterri, J., "Transcripción, etiquetado y codificación de corpus orales". Seminario de Industrias de la Lengua. Fundación Duques de Soria. Soria, 1997. Documento accesible a través de http://liceu.uab.es/~joaquim/publicacions/FDS97.html.

[3] Cresti, E. Et al. "The C-ORAL-ROM project. New methods for spoken language archives in a multilingual romance corpus". Proceedings of LREC 2002. Las Palmas de Gran Canaria.

[4] Sperberg-McQueen, C.M y Burnard, L. (eds.): *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. TEI, Chicago, Oxford, 1994.

[5] Moneglia, M., Scanaro, A. y Spinu, M. "Validation by expert transcribers of the C-ORAL-ROM prosodic tagging criteria on Italian, Spanish and Portuguese corpora of spontaneous speech". Documento del proyecto, accesible en http://lablita.dit.unifi.it/coralrom/papers/.

[6] Uchimoto, K., Nobata, C., Yamada, A., Sekine, S. y Isahara, H.: "Morphological Analysis of the Spontaneous Speech corpus". Proceedings of Conference of Computational Linguistics (COLING 2002); Taipei, Taiwan