EL PROYECTO C-ORAL-ROM Y SU APLICACIÓN A LA ENSEÑANZA DEL ESPAÑOL

ANTONIO MORENO SANDOVAL
Universidad Autónoma de Madrid

JOHN URRESTI
DTL, Instituto Cervantes

1. Introducción

C-ORAL-ROM es un corpus de habla espontánea en cuatro lenguas romances: italiano, francés, español y portugués. El proyecto ha sido financiado por la Comisión Europea dentro del V Programa Marco con un presupuesto superior al millón de euros. Se ha desarrollado desde enero de 2001 hasta marzo de 2004 y en él han participado nueve socios. El grupo central de desarrollo lo han constituido las Universidades de Florencia (coordinadora del proyecto), de Provenza, de Lisboa y la Autónoma de Madrid, cada una de las cuales ha compilado, transcrito, anotado y documentado el corpus correspondiente a su lengua. Además, la empresa WinPitch France ha desarrollado la herramienta informática para el alineamiento de la transcripción con el sonido. El Instituto Cervantes ha realizado la exploración de las posibilidades de la aplicación del corpus a la enseñanza, al tiempo que ha participado en la diseminación de los resultados del proyecto, junto con *ELRA* (agencia europea para la distribución de recursos lingüísticos en formato electrónico). La editorial John Benjamins se encargará de la publicación del libro y el *DVD* con el corpus completo.

C-ORAL-ROM consta de un total de 772 textos orales, con más de 120 horas de habla espontánea, distribuidas de manera equivalente en las cuatro lenguas, con una extensión aproximada de 300.000 palabras en cada idioma. Todas las grabaciones incluyen cuatro tipos de información:

1. Los metadatos de la grabación, donde se especifica las características sociolingüísticas, el contexto situacional, la calidad acústica y el contenido. El objeto de esta sección es proporcionar una identificación clara de los distintos tipos de habla espontánea registrados.
2. La transcripción ortográfica siguiendo el formato CHAT, enriquecida con la anotación de marcas prosódicas terminales y no terminales.
3. La sincronización entre el sonido original y su correspondiente transcripción ortográfica.

ORALIA, Vol. 8, 2005, págs. 81-104.

4. La anotación de la información morfosintáctica para cada palabra transcrita, así como la compilación de inventarios de formas y lemas, organizados por frecuencia.

Dicho de otra manera, cada texto oral se presenta en su versión sonora, transcrita y anotada morfosintácticamente. Mediante dos programas especiales incluidos en el DVD, se puede acceder a la fuente a partir de la transcripción (empleando WinPitch Corpus) y a la anotación lingüística (con el programa Contextes[1]). De esta manera, los investigadores pueden realizar sus consultas sobre todo el corpus. La única limitación es la imposibilidad de exportar texto o sonido a otra aplicación informática (por ejemplo, copiar en el portapapeles para incorporarlo en un procesador de textos). Las figuras 1 y 2 muestran ejemplos de pantallas de los programas mencionados.

El objetivo de este artículo es presentar la experiencia de los dos equipos españoles que ha participado en el proyecto europeo *C-ORAL-ROM*. En la primera parte, se tratarán los aspectos relacionados con el desarrollo del corpus. En la segunda parte, se abordarán cuestiones sobre los usuarios y aplicaciones potenciales de este corpus, insistiendo en los aspectos de enseñanza del español como *LE*. En la tercera parte, se describen las investigaciones realizadas sobre los resultados y los proyectos de extensión del corpus.

2. El desarrollo del corpus *C-ORAL-ROM*

En esta sección se abordará la metodología: cómo se ha hecho C-ORAL-ROM. Primeramente, se expondrán los objetivos centrales, enmarcándolos históricamente dentro de los antecedentes. A continuación se desarrollarán los aspectos más innovadores de la metodología empleada.

2.1. *Antecedentes y objetivos del proyecto*

El proyecto surge de una propuesta de E. Cresti y M. Moneglia que exponen a viejos amigos suyos (C. Blanche-Benveniste, F. Bacelar y F. Marcos

[1] WinPitch Corpus ha sido desarrollado por Philippe Martin, profesor de la Univ. París 7. Se puede conseguir información completa en la dirección http://www.winpitch.com/. Contextes es un programa de concordancias escrito por Jean Veronis, profesor de la Univ. de Proveza en Aix, además de director del equipo DELIC. Para información adicional, consúltese http://www.up.univ-mrs.fr/~veronis/logiciels/Contextes/index-en.html

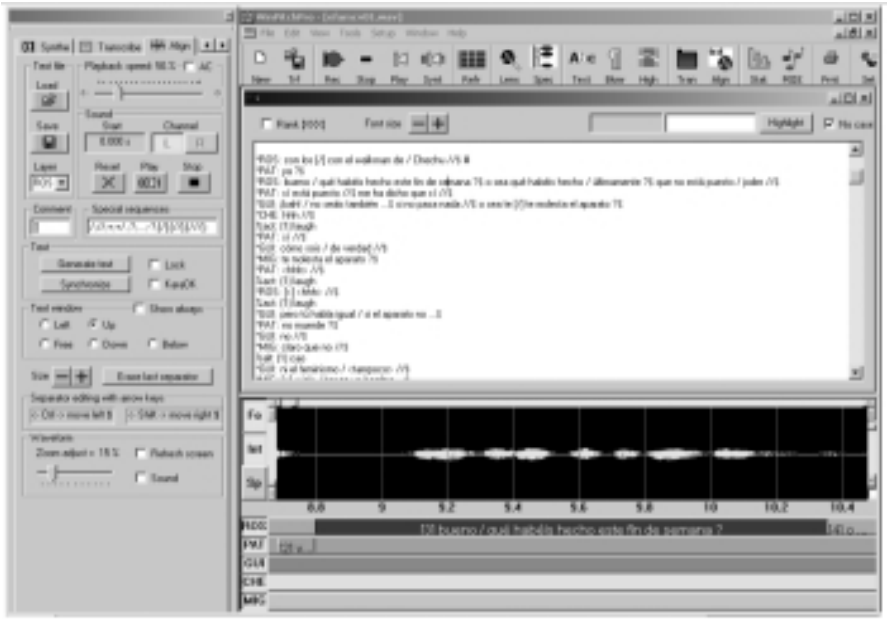


Figura 1. WinPitch y la transcripción alineada con el sonido fuente

Figura 2. Contextes y el corpus anotado morfosintácticamente

Marín): realizar un corpus oral cuatrilingüe comparable[2] a partir de la experiencia previa. Para ello, se reutilizarían textos que cada equipo ya tenía publicados[3] y se completaría con nuevas grabaciones empleando tecnología digital. Al núcleo central se integraría la empresa de software de análisis acústico WinPitch France para introducir innovaciones tecnológicas y dos instituciones académicas que funcionarían como usuarios evaluadores: el Instituto Cervantes para la aplicación del corpus a la enseñanza de L2 y el Instituto Trentino di Cultura para la aplicación a las tecnologías de reconocimiento automático de habla. La negociación con la Comisión Europea impuso dos nuevos socios: ELRA como agencia de transferencia y gestión de los resultados de la investigación y un editorial que se encargara de la publicación internacional en formato multimedia (originalmente se contó con Honoré Champion pero finalmente se encargará John Benjamins)[4].

Realizar un proyecto complejo como éste ha supuesto enfrentarse a problemas de todo tipo, pero previstos desde el comienzo. Podemos afirmar que uno de los éxitos del proyecto ha sido encontrar soluciones válidas a tres tipos de problemas:

1. Metodológicos: cómo conseguir un diseño común de la estructura del corpus y un formato común en la transcripción, anotación y alineamiento. Estos dos aspectos eran esenciales si quería obtener un corpus comparable. Al mismo tiempo, hay que considerar la dificultad de compaginar los criterios de cuatro tradiciones distintas.
2. Legales: la legislación europea sobre propiedad intelectual y derecha a la privacidad cambió a mediados de los noventa. Los corpus originales de cada equipo se recogieron en las décadas anteriores y no estaban sujetos a la obligatoriedad de solicitar el consentimiento por escrito de los participantes en las grabaciones. En nuestro caso, además estaba la obligación añadida de permitir a terceros su utilización en sus propios desarrollos. Es decir, el contrato con la Comisión Europea obligaba a que los resultados fueran públicos y reutilizables, de manera que se pudieran citar sin problemas en cualquier tipo de estudios o aparecer en desarrollos de tecnología lingüística como cursos de idiomas o programas multimedia. Por ello, todos los textos finalmente publicados han pasado la validación legal (realizada por ELRA y confirmada por la Comisión Europea). Cualquier fragmento puede reproducirse porque sus autores lo han autorizado por escrito. El modelo de carta de consentimiento se ofrece en la Figura 3.
3. Tecnológicos: un corpus hablado actual debe emplear los recursos tecnológicos disponibles. En primer lugar, las grabaciones deben ser digitales (bien DAT o minidisk) y con micrófonos especiales. Posteriormente, todo tratamiento de la fuente sonora tiene que ser digital, lo que implica herramientas informáticas apropiadas. Se ha aprovechado el proyecto para desarrollar una versión ampliada del programa WinPitch para la transcripción y alineamiento de texto y sonido. La principal innovación de esta herramienta es la posibilidad de alinear las transcripciones de los solapamientos en niveles separados para cada hablante. Otra de las innovaciones tecnológicas del proyecto es que se ha facilitado la anotación morfosintáctica empleando analizadores automáticos, apoyados con revisión manual por parte de lingüistas.

La información detallada acerca de todos estos aspectos se describe en las noventa páginas del capítulo primero del libro que acompaña al DVD. En los siguientes capítulos se describen las peculiaridades de cada subcor-

Yo, ..., DNI ....., a petición de los responsables del proyecto C-ORAL-ROM, doy mi autorización para:

- grabación de mi voz
- transcripción de la grabación
- tratamiento del sonido y de la transcripción
- publicación y comercialización del sonido y la transcripción

Mantengo el derecho a oír la grabación y a denegar mi autorización por cualquier motivo que considere pertinente

FIGURA 3. Modelo de autorización para C-ORAL-ROM

[2] Los corpus de más de una lengua pueden ser paralelos o comparables. Los corpus paralelos consisten en el mismo texto en las distintas lenguas. Normalmente uno es el original y los otros son traducciones, y generalmente son textos escritos. El equivalente en corpus oral sería, por ejemplo, las distintas versiones de una película. Sin embargo, es obvio que no se puede hacer un corpus paralelo con habla espontánea, pues se violaría la condición esencial de espontaneidad, entendida como habla sin guión previo. Por el contrario, lo apropiado es hacer un corpus comparable, donde se controle el tipo de texto (registro, dominio temático, característica sociolingüísticas de los hablantes) y se deje plena libertad a los participantes en su producción discursiva.

[3] El equipo francés partía de la experiencia de su corpus oral del francés empezado en 1978, que contenía más de 2.000.000 de palabras. El LabLIta de Florencia contaba con el Corpus di italiano parlato, iniciado en los setenta. Los portugueses del CLUL habían recogido más de un millón y medio de palabras, desde principios de los ochenta.. El LLI-UAM había compilado entre 1990-1992 el Corpus Oral de Referencia de la Lengua Española Contemporánea, con algo más de un millón de palabras.

[4] Dos de los promotores del proyecto, Claire Blanche-Benveniste y Francisco Marcos Marín, finalmente no se harían cargo de la dirección de sus respectivos equipos por ocupar responsabilidades académicas, pasando la dirección a Jean Veronis y Antonio Moreno Sandoval.

pus con respecto al formato y metodología común. Debido a las limitaciones propias de un artículo como éste, nos limitaremos a destacar los aspectos sobresalientes.

1. Distribución del corpus: la idea central es recoger una muestra lo más representativa posible de los distintos registros de habla espontánea, dentro de la limitación de las 300.000 palabras[5]. Por ello, se dividió el corpus en dos grandes secciones de tamaño similar: registro informal frente a registro formal, aproximadamente unas 150.000 palabras cada uno. Además se incluyó un subconjunto de conversaciones telefónicas (25000 palabras), muy interesante para los investigadores de sistemas de reconocimiento del habla. Dentro del registro informal se distingue entre un ámbito familiar o privado y un ámbito público. De igual manera, el registro formal consta de grabaciones de los medios de comunicación y grabaciones en contextos formales como conferencias, homilías, sesiones políticas. En el registro informal, la subdivisión se realizó además por el estilo dialógico: monólogos, diálogos y conversaciones. En el registro formal la subclasificación fue temática: política, deportes, ciencia, etc. En todos ellos se cuidó que los participantes de ambos sexos estuvieran equilibrados, aunque por las peculiaridades de los textos hay un cierto desequilibrio a favor de las mujeres en los textos informales y a favor de los hombres en los textos formales. En la cabecera de cada grabación se registran los aspectos sociolingüísticos más relevantes como la edad, estudios, procedencia geográfica, profesión y papel en la conversación. Igualmente se describe el tema, el contexto de la grabación y la situación del investigador (oculto o presente) que realiza la grabación. La Figura 4 muestra el árbol con la distribución del corpus.
2. Formato común: El otro aspecto que se necesita para permitir 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.

[5] Todos los corpus anteriores de los equipos participantes eran mucho más extensos en número de palabras. Sin embargo, en esta ocasión las consideraciones de calidad acústica, transcripción perfectamente revisada, alineamiento manual de texto y sonido y, por último, anotación morfosintáctica, obligaron a concentrarse en un número más reducido de palabras, pero suficiente para hacer una muestra útil y representativa. En este caso la calidad y la complejidad de los textos era preferible a la cantidad de palabras.

FIGURA 4. Árbol de la distribución del corpus

3. Validación de los datos: 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 ha sometido a dos tipos de validación, una interna y otra externa.
   - La validación interna: 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. 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 ha sido realizada por expertos externos al final del proyecto. Un equipo independiente de la empresa LOQUENDO ha evaluado el grado de acuerdo entre los anotadores en cuanto a la asignación de marcas prosódicas (véase informe en Cresti y Moneglia (eds) 2005). El final del proyecto y como requisito para que la Comisión Europea aprobara la resolución del contrato, ELRA

se encargó de auditar los datos y los permisos de utilización, de manera que las cifras que se proporcionan son reales y verificadas.

4. La anotación morfosintáctica: los cuatro corpus están completamente anotados con la categoría y lema de cada palabra, además de la información gramatical pertinente (género, número, tiempo, etc.). Esta anotación ha sido realizada automáticamente con analizadores morfológicos para cada lengua. Sólo una parte de cada corpus ha sido verificada por lingüistas para evaluar el grado de fiabilidad del etiquetado automático. En el caso del corpus español, se empleó el programa GRAMPAL (Moreno 1991, Guirao y Moreno 2003, Moreno y Guirao 2004) y se verificaron a mano 50.000 palabras (aproximadamente un sexto del corpus) con un grado de acierto del programa del 95,6 %[6]. Esta información es esencial para comparar los cuatro corpus en cuanto a la distribución por frecuencia de lemas y categorías.

Moreno (2002) analiza la evolución de los dos corpus orales realizados por el LLI-UAM y a modo de resumen lo presentamos en una tabla comparativa[7]:

	CORLEC	C-ORAL-ROM
Fecha de compilación	1990-1992	2001-2004
Número de palabras	1.100.000 (aprox.)	312.000 (aprox.)
Tipo de grabaciones	Analógicas	Digitales
Tipo de transcripción	Transliteración siguiendo las convenciones de la lengua escrita	Transliteración ortográfica, pero sin emplear signos de puntuación y otras convenciones de la lengua escrita
Niveles de anotación	Transcripción con marcas de fenómenos propios del habla	Transcripción con marcas de fenómenos del habla, anotaciones prosódicas de unidades entonativas, anotación morfosintáctica.

[6] La evaluación de la precisión se realizó tomando como modelo el corpus de 50000 verificadas por los lingüistas. Se contrastaron los análisis del modelo y del proporcionado por GRAMPAL y se marcaron las discrepancias. Las etiquetas asignadas por GRAMPAL coincidieron con el modelo en un 95,6% de los casos. Dado que la muestra seleccionada tiene un tamaño considerable con respecto al total, se puede considerar esa tasa de precisión como representativa de la fiabilidad de la anotación del corpus español completo.

[7] El trabajo de A. Moreno Sandoval ha sido financiado por la Unión Europea (IST 2000-26228) y por el MEC (CICYT TIN2004-07588-C03-02).

	CORLEC	C-ORAL-ROM
Alineamiento texto y sonido fuente	NO	SÍ, por unidades entonativas.
Autorización por escrito de los participantes	NO	SÍ
Validación	NO	SÍ, interna y externa

3. Usuarios potenciales y aplicaciones

3.1. *Evaluación del corpus*

Con el fin de evaluar la aplicabilidad de C-ORAL-ROM a la investigación y la enseñanza en los campos de la fonética, las tecnologías lingüísticas y las lenguas románicas, el Departamento de Tecnología y Proyectos Lingüísticos del Instituto Cervantes desarrolló una base de datos con información sobre usuarios potenciales del corpus, cuya experiencia en el campo de la investigación y la enseñanza da especial relevancia y fiabilidad a sus opiniones acerca de la aplicabilidad de esta herramienta.

El objetivo de final de la base de datos ha sido enviar a una muestra representativa de posibles usuarios el corpus, acompañado de un cuestionario, para que lo probasen y después enviasen el cuestionario con sus opiniones acerca de la aplicabilidad del corpus y sus sugerencias para mejorar la herramienta.

Los más de 1.000 expertos que integran esta base de datos desarrollan su actividad en las siguientes instituciones:

- Centros y aulas ELE del Instituto Cervantes
- Universidades españolas
- Universidades europeas
- Universidades hispanoamericanas
- Universidades norteamericanas
- Otras instituciones (fundaciones, asociaciones, empresas privadas, etc.)

El cuestionario se envió en enero y febrero de 2004 a una muestra suficientemente representativa que estaba integrada por 277 expertos pertenecientes a diversos países. Para ello se desarrollaron dos versiones del cuestionario, una en inglés y otra en español. Ambas contenían nueve preguntas cuyo fin era obtener la siguiente información:

- Perfiles de usuarios
- Otras herramientas parecidas disponibles en el mercado y las necesidades de sus usuarios
- Ventajas e inconvenientes de la herramienta con relación a su forma, contenidos y funcionamiento general
- Posibles aplicaciones
- Sugerencias para mejorar la herramienta
- Posibilidades existentes para integrar la herramienta en actividades de enseñanza del español, italiano, francés y portugués como lenguas extranjeras

Paralelamente a este cuestionario, un equipo del Departamento de Tecnologías y Proyectos Lingüísticos del Instituto Cervantes realizó las pruebas con la finalidad de poder contrastar sus conclusiones con las emitidas por los expertos a los que se consultó.

3.2. *Resultados de la evaluación*

Los resultados de la evaluación han sido en general positivos y demuestran la existencia de un interés real en el mercado hacia una herramienta de estas características.

Todos los expertos que participaron en la evaluación mostraron su interés por el corpus y solicitaron que se les mantuviese informados sobre la evolución del proyecto y también acerca del momento de poner el corpus al alcance del público.

3.2.1. *Potencial y aplicabilidad*

La valoración general fue positiva, concluyéndose que C-ORAL-ROM ofrece un gran volumen de recursos originales y muy variados que, efectivamente, resultarán útiles a los investigadores y formadores que trabajan en fonética, tecnologías lingüísticas, lingüística aplicada, traducción y ELE.

Los expertos que participaron en la encuesta sugirieron un gran número de usos potenciales que resumimos a continuación:

1. Potencial y aplicabilidad al campo de la investigación lingüística (fonética, lexicología, pragmática y sintaxis):
   - Estudio del uso y de la evolución del lenguaje
   - Estudio de los enlaces de las estructuras del lenguaje, estrategias pragmáticas, estructuras sintácticas, pronunciación

   - Investigación en el campo de la pragmática y del análisis del discurso
   - Análisis de conversaciones
   - Análisis de contextos
   - Estudio de la terminología y del significado
   - Análisis de la fonética de un determinado lenguaje y para establecer comparaciones
   - Análisis de aspectos fonéticos, transcripciones, ortología y errores de pronunciación
   - Análisis del uso del sujeto pronominal en dos o más lenguajes
   - Análisis de aspectos del discurso tales como la anáfora, la coherencia, etc.
2. Potencial y aplicabilidad al campo de la investigación sobre tecnologías lingüísticas:
   - Análisis del habla espontánea
   - Estimación de pautas acústicas
   - Predicción de pautas lingüísticas
   - Reconocimiento automático del habla
   - Utilización como referencia para compilar nuevos corpus orales
   - Eliminación de ambigüedades
3. Potencial y aplicabilidad en el campo de la traducción
   - Estudio de la terminología
   - Extracción de términos y oraciones
   - Compilación de bancos de datos sobre materias específicas: economía, medicina, etc.
   - Comparación entre lenguajes
4. Potencial y aplicabilidad en el campo de ELE
   - Es un material básico para la enseñanza asistida por ordenador
   - Es un material básico para un aula multimedia
   - Es un material importante para el autoaprendizaje y para autodidactas
   - Es un material útil para practicar la comprensión oral y la interacción
   - Es un material esencial para aplicar un enfoque comunicativo a la enseñanza y al aprendizaje del lenguaje
   - Es útil para desarrollar habilidades orales
   - Es útil para practicar con distintos registros del lenguaje y situaciones de comunicación
   - Es útil para analizar el discurso oral en distintos contextos
   - Es útil para practicar con enlaces, estructuras sintéticas, estrategias pragmáticas y pronunciación
   - Es útil para analizar aspectos contextuales para comprender alófonos

- Es útil para estudiar vocabulario y oraciones
- Es útil para analizar el uso variable del sujeto pronominal desde un punto de vista pragmático
- Es un material multimedia que contribuye a aumentar el interés de los alumnos

5. Potencial y aplicabilidad en el campo de la edición y la publicación
   - Es un material de referencia para realizar consultas sobre lenguaje hablado en esos cuatro idiomas
   - Es un material de referencia para analizar muestras orales para producir materiales didácticos
   - Es una referencia importante para mejorar y difundir recursos multimedia

3.2.2. *Sugerencias de los expertos*

Los expertos que participaron en la encuesta sugirieron las siguientes mejoras encaminadas, en su gran mayoría, a incrementar el potencial del corpus como una herramienta de ELE:

1. Incluir un motor de búsqueda común a todos los idiomas o uno específico para cada idioma, que permita organizar por temas, contenidos funcionales y gramaticales, entre otros criterios. Este motor resultaría especialmente útil para emplear la herramienta en el marco de la enseñanza de idiomas.
2. Deberían explicarse los criterios por los que se determina si una muestra de lenguaje, formal o informal, se puede incluir en el corpus.
3. Las referencias de las muestras no son suficientemente descriptivas: deberían utilizarse palabras del título o indicarse el tema con el fin de hacer que el uso de la herramienta resulte más sencillo.
4. Sería muy útil incluir guías, pautas o sugerencias acerca de cómo pueden los formadores utilizar el corpus en sus clases. En este sentido, sería muy práctico clasificar las muestras según los niveles que se establecen en el "Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación".
5. Sería útil añadir archivos y campos de tipo científico y tecnológico, para facilitar su aplicación en determinadas materias, como la traducción
6. Además de permitir la alineación de los archivos de sonido con una transcripción fonética existente, sería útil que se añadiese la correspondiente transcripción fonética a los archivos de lengua hablada. Esta mejora sería de especial utilidad para los formadores, así como para los alumnos a distancia y los autodidactas.
7. La inclusión de corpus complementarios con grabaciones de otras áreas geográficas en las que se habla cualquiera de las cuatro lenguas contempladas, sería útil para investigadores y formadores de los campos de la dialectología.
8. Se podría incluir un corpus rumano para extender C-ORAL-ROM a otras lenguas románicas. De hecho ya existen corpus rumanos compilados que podrían utilizarse.

3.3. *Conclusiones sobre la aplicabilidad de C-ORAL-ROM a la enseñanza de LE*

Los comentarios y sugerencias de los expertos, permiten llegar a las siguientes conclusiones con respecto al potencial que tiene este corpus:

1. Se trata de un recurso lingüístico de gran valor y múltiples aplicaciones en el campo de la investigación y la formación lingüística y de las tecnologías lingüísticas.
2. C-ORAL-ROM tiene muchas aplicaciones directas a la enseñanza de lenguas extranjeras, así como un gran potencial para ser mejorado mediante nuevos desarrollos que pudiesen darse en el futuro. Solventa en parte la carencia de muestras que puedan ser utilizadas para el estudio y la práctica del lenguaje coloquial hablado.
3. Sin embargo, es necesario facilitar la tarea de los formadores, proporcionándoles guías y sugerencias acerca de cómo utilizar la herramienta, así como referenciar las muestras de una manera más representativa y proporcionarles un motor de búsqueda. En gran medida, la importancia de incluir estas mejoras para facilitar la aplicación de C-ORAL-ROM en el mundo de la formación estriba en la escasez de tiempo de que disponen los formadores.
4. Un corpus de archivos orales alineados con sus correspondientes archivos de texto, que permite realizar análisis lingüísticos detallados y establecer comparaciones mediante aplicaciones técnicas innovadoras, genera un amplio abanico de oportunidades para ser utilizado tanto en la enseñanza convencional como en la formación a través de Internet.

4. Algunas investigaciones realizadas sobre el corpus español

El corpus ha sido la fuente para una serie de publicaciones realizadas por el equipo del LLI-UAM. A continuación resumimos las conclusiones de los estudios realizados.

4.1. *Resultados de la comparación de los cuatro corpus*

El volumen publicado por John Benjamins incluye un capítulo dedicado a comparar las estadísticas extraídas a partir de los datos anotados. Las tablas y los gráficos se ofrecen también en el DVD junto con los corpus para que los investigadores puedan acceder fácilmente a los datos. Se han computado los siguientes parámetros:

Longitud media de las proferencias (utterances) medida en palabras y por tipos de texto.

1. Longitud media de los turnos, en palabras y por tipos de texto.
2. La tasa de velocidad, medida en palabras por segundo.
3. La longitud media de las unidades tonales, en palabras.
4. La fragmentación del discurso, entendida como el número medio de palabras que ocurre una interrupción.

Una observación interesante es que en general el portugués, el español y el italiano están más próximos en cuanto a estas medidas que el francés. Los datos no son suficientes para extraer conclusiones significativas pero alumbran tendencias para investigaciones futuras.

A nuestro juicio, lo más destacable del material publicado son las listas, ordenadas por frecuencias, de las formas y los lemas que aparecen en cada corpus. Estos inventarios han podido elaborarse a partir de la anotación morfosintáctica y suponen un hito en el estudio del léxico de las cuatro lenguas, en su variante de habla espontánea, por la fiabilidad y representatividad de la muestra. A modo de ejemplo, citamos algunos datos básicos del inventario español:

Nº de formas distintas	21536
Nº de lemas distintos	19315
Nº de lemas nominales	4583
Nº de lemas verbales	1779
Nº de lemas adjetivales	2023

4.2. *La relación entre unidades lingüísticas e información socio-contextual*

Sin duda uno de los aspectos más interesantes que puede aportar C-ORAL-ROM a los estudios lingüísticos, gracias a la gran cantidad y variedad de información anotada, es relacionar unidades lingüísticas (léxico, categorías sintácticas) con variables sociolingüísticas y contextuales. Un ejemplo de ello es el experimento que realizó el equipo de investigación y presentó en el congreso internacional Corpus Linguistics 2003, celebrado en la Universidad de Lancaster. La comunicación fue seleccionada para aparecer en el libro Corpus Linguistics across the World, publicado por Rodopi en 2005.

El experimento consistió en aplicar recientes técnicas estadísticas (el test de Dunning) e informáticas (metadatos y anotación en xml) a todo el corpus español para encontrar las diferencias representativas entre las diferentes partes (o subcorpus). De esta manera, se pueden asociar los elementos léxicos o gramaticales más distintivos con un rasgo socio-contextual concreto.

El método, expuesto en Guirao et al. (2005), consiste en los siguientes pasos:

1. Relacionar los metadatos con los rasgos lingüísticos empleando la anotación en xml. Efectivamente, las cabeceras de cada texto contienen una completa información del contexto sociolingüístico, desde la clasificación del tipo de registro hasta las características de cada participante (sexo, edad, formación, profesión, lugar de origen). Mediante un programa informático, se crearon automáticamente distintos subcorpus. Por ejemplo, el subcorpus de las mujeres y de los hombres, subcorpus por profesiones, subcorpus por registros (medios de comunicación, teléfono, etc.). Por otra parte, la transcripción de cada texto está también anotada con etiquetas morfosintácticas, de las que tomamos solo en cuenta la categoría sintáctica y el lema. De esta manera, para cada subcorpus, contamos todas las ocurrencias para cada tipo (ya sea éste una forma flexionada, un lema o una categoría).
2. Extracción de agrupamientos de unidades. Si hubiéramos calculado las estadísticas directamente sobre cada palabra, el resultado no sería correcto, ya que aquellas unidades formadas por más de una palabra (marcadores discursivos tan frecuentes y tan significativos en la lengua oral como "por ejemplo", "o sea" o "es decir") no saldrían en el cómputo. Por supuesto, tampoco saldrían las locuciones prepositivas ("a pesar de"), las adverbiales ("ya mismo") o los compuestos ("Lingüística de Corpus"). Para resolver este problema, desarrollamos un programa de extracción de candidatos a unidad pluriverbal (nuestra traducción del término inglés *multi-word*). De los candidatos a unidad pluriverbal se seleccionaron a mano los que pasarían a la lista definitiva, de manera que para nuestro método toda unidad pluriverbal es considerada un unidad léxica, equivalente a una unidad basada en una única palabra gráfica.
3. Aplicación de la Estadística de la Sorpresa. Esta técnica desarrollada por Dunning (1993), también conocida en inglés por *log-likelihood ratio test*, consiste en asumir una distribución binomial de las unidades

en un corpus, en lugar de la distribución normal. De esta manera, se favorecen aquellas unidades que aparecen poco pero que son muy significativas de un registro particular. El test compara las unidades que aparecen en el subcorpus seleccionado con las unidades del subcorpus complementario (en este caso el resto de los subcorpus) y otorga un valor mayor a las unidades que aparecen sólo en el registro analizado. Así, por ejemplo, podemos destacar las palabras, lemas o categorías que, en nuestro corpus, emplean los hombres y las mujeres, diferentes profesiones o registros. Para comprobar la capacidad del test para encontrar unidades distintivas en dominios particulares, empleamos como hipótesis nula que si tomamos dos subcorpus cualquiera las unidades resultantes serán las mismas. Escogimos dos subcorpus bien definidos, los partes meteorológicos y grabaciones del dominio jurídico, y los resultados fueron los siguientes:

- Lemas más frecuentes (por orden de mayor a menor) en el dominio meteorológico:
  norte, fuerza, viento, en, componente, temperatura, noroeste, oeste, nube, zona
- Lemas más frecuentes (por orden de mayor a menor) en el dominio jurídico:
  policía, persona, derecho, contrato, judicial, delito, delincuente, ley, determinar, cometer

Estos resultados nos animaron a aplicar el test a otras unidades y otros dominios. Por ejemplo, cuáles son los lemas verbales más característicos de los hombres y las mujeres en nuestro corpus:

- 10 verbos más característicos de las mujeres en C-ORAL-ROM:
  Ir, decir, saber, venir, decir, mirar, comprar, gustar, quedar, contar
- 10 verbos más característicos de los hombres en C-ORAL-ROM:
  escuchar, recordar, aparecer, llegar, contemplar, caminar, intentar, amar, juntar, superar

En resumen, el método correlaciona datos lingüísticos y socio-contextuales mediante la Estadística de la Sorpresa de Dunning. Para conseguir esto, es esencial contar con un corpus ricamente anotado y utilizar xml como lenguaje de marcación. Los experimentos realizados por el equipo del LLI-UAM demuestran que el procedimiento sirve para validar empíricamente hipótesis sociolingüísticas así como para clasificar textos de acuerdo con una tipología textual. Sin embargo, extraer conclusiones e interpretaciones de los datos mostrados es prematuro, ya que el tamaño del corpus no es estadísticamente suficiente.

4.3. *Problemas de transcripción de habla espontánea*

La transcripción de grabaciones espontáneas presenta diferentes dificultades dependiendo del tipo de texto. Para los investigadores del C-ORAL-ROM, poder determinar qué interacciones comunicativas provocan más dificultades en la transcripción resulta un conocimiento útil para la compilación de nuevos corpus de habla espontánea: los recursos humanos y materiales son limitados y la estimación del grado de dificultad de una determinada compilación es esencial para calcular los costes y diseñar de acuerdo con ellos la distribución de las grabaciones.

Por ello, el equipo del LLI-UAM realizó el siguiente experimento (González et al. 2004), presentado en el taller sobre compilación y anotación de corpus orales de la IV Linguistic Resources and Evaluation Conference (LREC-2004). Partimos de la hipótesis de que los problemas de transcripción están asociados a la frecuencia de aparición de dos clases de fenómenos lingüísticos típicos del habla espontánea:

- Rasgos de producción, como palabras fragmentadas, apoyos vocálicos, rectificaciones, etc.
- Rasgos de interacción, como el número de turnos o el solapamiento.

Esta hipótesis es intuitiva y reconocida por diferentes investigadores. La novedad del experimento consistía en que esos fenómenos estaban anotados en la transcripción y, por tanto, su recuento era automático y completo. Nuestra tarea era analizar los resultados para comprobar empírica y cuantitativamente la hipótesis. Queremos insistir en el hecho de que este análisis se realizó después de la transcripción, revisión y anotación del corpus y, por tanto, los datos son completos y finales (sobre el corpus C-ORAL-ROM). La extrapolación a otros corpus dependerá de su cercanía a los tipos de grabaciones y al sistema de transcripción y anotación de los fenómenos reseñados.

Para evaluar la hipótesis decidimos crear dos escalas que nos permitieran situar los fenómenos en función de su dificultad:

1. Grado de Formalidad: en los dos extremos se situarían los textos informales privados (los más difíciles de transcribir) y los textos formales públicos. Se asume que cuanto más formal es la comunicación (por ejemplo, una conferencia en la universidad o una homilía) más cercana estará a la norma escrita y más fácil será su transcripción. Los textos de los medios de comunicación se sitúan en la posición intermedia, donde hay algunos más informales (los programas deportivos o los talk shows) y otros más formales (los reportajes científicos o las entrevistas políticas).

2. Número de hablantes: este parámetro sólo lo hemos evaluado en los textos clasificados como informales (los más complejos según la escala anterior). Cuanto mayor sea el número de participantes mayor la complejidad de la transcripción. Así los más difíciles son las conversaciones, seguidas de los diálogos y los monólogos.

Gráficamente, podemos representar ambas escalas de la siguiente manera:

Grado de formalidad

informal media formal

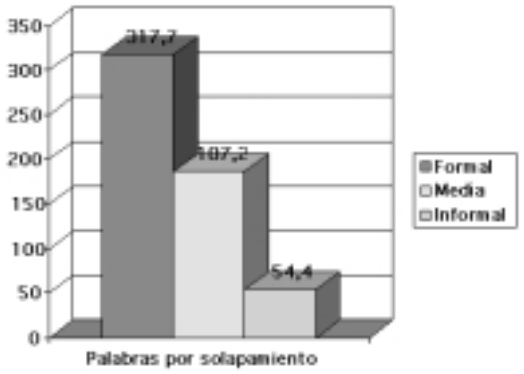
+ difícil ——————————————— – difícil

Número de hablantes

conversación diálogo monólogo

+ difícil ——————————————— – difícil

Para el grado de formalidad seleccionamos tres rasgos: palabras fragmentadas (es decir, no acabadas de pronunciar), apoyos vocálicos (expresiones como "eh", "um") y reinicios (cuando se interrumpe la proferencia y se comienza de nuevo). Para la escala de número de hablantes seleccionamos otros tres rasgos: número de turnos dialógicos, solapamiento y la tasa de velocidad en la producción de habla. Como toda esta información estaba marcada en la transcripción, se empleó un programa que calculaba la media de palabras que había entre dos etiquetas del mismo rasgo, salvo en el caso de la tasa de velocidad que se calculaba la media de palabras por segundo. Todas estas frecuencias se calcularon para cada tipo de texto. De esta manera, cuanto mayor sea el número de palabras por fenómeno, menos frecuencia e importancia tendría dicho rasgo para la clase textual concreta. Dado que se asume que la presencia de esos rasgos representa un problema en la transcripción, menor frecuencia implica menor dificultad.



Veamos un ejemplo para ilustrar el método. Estas son las cifras del número medio de palabras en las que nos encontramos un solapamiento en cada una clase, formal, media e informal: El gráfico nos muestra que se produce un solapamiento cada 54,4 palabras de media en los textos informales, mientras que en los textos formales es de 317,7. Es decir, es aproximadamente seis veces más frecuente el solapamiento en los textos informales que en los formales.

Así aplicamos el procedimiento a los 6 rasgos seleccionados en las dos escalas (en total 12 estadísticas) y obtuvimos las siguientes conclusiones:

1. Los rasgos de producción (palabras fragmentadas, apoyos vocálicos y reinicios) se comportan de manera contraria a la hipótesis intuitiva: son mucho más frecuentes en los textos formales que en los informales. Esto puede explicarse porque los hablantes en una situación formal tienden a cuidar su expresión y dedican más tiempo a producir sus proferencias. Al aumentar el tiempo de producción, hay más necesidad de llenar los huecos mientras se busca la expresión adecuada (los apoyos vocálicos) y también hay más probabilidad de que se cambie la expresión mientras se está emitiendo (palabras fragmentadas y reinicios). En cambio, en los textos informales parece más importante la necesidad de expresión y comunicación que la de emitir un discurso correcto y preciso.
2. Los rasgos de interacción (palabras por turno, solapamientos y tasa de velocidad) sí se comportan según la hipótesis intuitiva: son más frecuentes en los textos informales, donde la comunicación es mucho más rápida que en los formales. La velocidad de producción explica que los turnos sean más cortos pero más frecuentes, que haya solapamientos continuos y que se digan más palabras por segundo.

Como conclusión principal del estudio, podemos decir que los rasgos realmente influyentes en la dificultad de transcripción son los de interacción, mientras que los rasgos de producción no son un problema para el transcriptor: simplemente tiene que anotarlos, pero no le obligan a escuchar repetidamente el fragmento para describir lo que dice cada hablante en un solapamiento, por ejemplo. Una conclusión secundaria es que si combinamos ambos tipos de rasgos, los textos menos problemáticos de transcribir son los de los medios de comunicación, ya que, aunque situándose en el medio de la escala, están siempre muy cerca de los casos más sencillos. Efectivamente, la profesionalidad de los locutores televisivos o radiofónicos se refleja en que su producción de apoyos vocálicos, reinicios o palabras fragmentadas es casi similar a la de los textos informales y, aunque hablan casi tan rápido como en los textos informales, el solapamiento es considerablemente menor.

Finalmente, en este estudio dejamos abierta la comparación entre la dificultad para un transcriptor humano y la dificultad de "transcripción" para un sistema automático de reconocimiento de habla. En colaboración con el Laboratorio de Comunicación Hombre-Máquina de la UAM hemos realizado un experimento que se describe en Torre et al. (2004) y que fue presentado en las III Jornadas de Tecnologías del Habla organizadas por la Universidad Politécnica de Valencia. El método es el siguiente: se toman las transcripciones como corpus de evaluación, de manera que el sistema tiene que idealmente reconocer las mismas palabras que el transcriptor humano. Además, para reducir la enorme complejidad de la tarea para el sistema, la señal se suministra fragmentada por segmentos entonativos, tal y como están marcados en el corpus.

En este experimento se llega a conclusiones bastante parecidas: los textos con mejor tasa de reconocimiento por el sistema son los de los medios de comunicación, en concreto, los de reportajes científicos, que superan el 60% de reconocimiento. Sin embargo, todos los informales están en torno al rango de los 30-40 % de reconocimiento.

4.4. *Investigaciones en curso*

La extensión de un corpus, de cualquier tipo, es una tarea siempre inacabada. A continuación describiremos brevemente los proyectos en los que está embarcado el equipo español de C-ORAL-ROM.

4.4.1. *Herramientas para explotación del corpus*

Aunque el proyecto C-ORAL-ROM distribuye dos programas excelentes para consultar información sobre los textos (WinPitch y Contextes), el grupo LLI-UAM considera necesario desarrollar su propia tecnología de explotación para no depender de estos programas. La dependencia tecnológica se manifiesta en dos formas: por un lado el mantenimiento y actualización de los programas está condicionado a las relaciones futuras con el proveedor y, por otro lado, el acceso a la información está condicionado por las características de los programas, que no permiten todas las búsquedas interesantes para el grupo. En dos comunicaciones presentadas en LREC 2004 (Lisboa) y TALC-6 (Granada), Antonio Moreno y José M. Guirao exponen las herramientas informáticas en proceso de desarrollo, que se dividen en dos tipos:

- Herramientas para revisión y anotación del corpus: aquí se integran distintos programas. Por una parte, está el analizador morfológico y los editores que permiten modificar el lexicón y la gramática de desambiguación. Por otra parte, está el editor de textos para hacer las revisiones y modificaciones a los textos anotados.
- Herramientas para la consulta del corpus: se ha desarrollado un programa de consulta a los textos a través de la información asociada en los metadatos. De esta manera, se puede pedir que nos muestre todos los textos en los que aparezcan, por ejemplo, mujeres con características sociológicas concretas, o todos los textos donde aparezcan hablantes de determinada profesión, etc. Por otra parte, se está desarrollando un programa de concordancias con acceso a texto y sonido basado en xml. La funcionalidad de este programa es similar a la de Contextes pero permite explotar las posibilidades de la anotación en xml y, en concreto, integrarlo en un interfaz de consulta por Internet.

Estos recursos constituyen una parte importante del proyecto RILARIM (Recursos de Ingeniería Lingüística Aplicados a la Recuperación de Información Multilingüe) financiado por el MEC (TIN2004-07588-C03-02) y que se desarrolla entre 2005 y 2007.

4.4.2. *Aplicación de la metodología a otros dominios*

C-ORAL-ROM es obviamente un corpus incompleto, a pesar del intento de recoger una muestra lo más representativa posible de la lengua hablada actual dentro de las limitaciones de tiempo y recursos humanos del proyecto. Sin duda, el aspecto peor representado en C-ORAL-ROM son las variantes dialectales. Aunque hay hablantes de distintas zonas geográficas, incluyendo América, lo cierto es que la mayoría pertenecen a la variante del centro peninsular (Madrid y Segovia). El equipo del LLI-UAM no cuenta con especialistas en dialectología, necesarios para el diseño de las grabaciones y su anotación. Sin embargo, creemos que la metodología y los recursos tecnológicos se aplican directamente a la recopilación de un corpus dialectal.

Otro de los registros sin cubrir en C-ORAL-ROM es del lenguaje infantil. Se tomó la decisión de no incluir este dominio por la dificultad de conseguir las autorizaciones y por no contar con especialistas en el tema. Afortunadamente, el LLI-UAM ha podido contar con la colaboración de Elena Garayzábal, profesora del Departamento de Lingüística de la UAM, para el diseño de la recogida de muestras y para obtener los permisos de los padres en distintos colegios de la Comunidad de Madrid. Miguel Pérez Milans se encargó de realizar la mayoría de las grabaciones y de su transcripción, subvencionado por una beca del LLI-UAM. Los resultados preliminares del proyecto CHIEDE (Corpus de Habla Infantil Espontánea Del Español) se ha pre-

sentado en el simposio de la SEL 2004 (Pérez Milans y Moreno Sandoval 2004). Este nuevo proyecto confirma que la metodología desarrollada por C-ORAL-ROM es extensible a corpus de dominios específicos.

Por último, también estamos trabajando en la compilación de un corpus de grabaciones de catas profesionales de vino, en este caso en colaboración con la Universidad de Castilla-La Mancha (campus de Ciudad Real). La recogida de este corpus forma parte de la tesis doctoral de Teresa de Cuadra, de la UCLM. De nuevo, aquí se integran la experiencia del LLI-UAM en la grabación de los datos y en la transcripción y anotación de los textos con la aportación del especialista en un dominio especializado. El especialista es esencial para decidir qué tipos de grabaciones hay que hacer y contactar con los hablantes idóneos.

4.4.3. *Extensión de la anotación a otros niveles lingüísticos*

La utilidad de un corpus aumenta con la diversidad de anotaciones que contiene, asumiendo que la calidad de las anotaciones está garantizada mediante la validación interna y externa, como se ha explicado anteriormente. El corpus C-ORAL-ROM contiene anotaciones en los siguiente niveles: léxico (palabras y lemas), prosódico (unidades tonales) y morfosintáctico (categorías e información gramatical). De estos tres niveles sólo la transcripción ortográfica y prosódica ha sido revisada y validada al completo. La anotación léxica y morfosintáctica se ha realizado semiautomáticamente y sólo están revisados a mano 50.000 palabras. Lógicamente, nuestra primera tarea pendiente es completar la revisión del todo el corpus. En la actualidad ya hemos alcanzado las 150.000 palabras revisadas a mano. La versión completa se hará pública en su momento.

Paralelamente y en colaboración con el Laboratorio de Comunicación Hombre-Máquina de la UAM hemos comenzado el proceso de transcripción fonológica de todo el corpus. Para ello, estamos empleando un transcriptor fonológico automático desarrollado originariamente por Doroteo Torre y mejorado con las aportaciones de los lingüistas del LLI. Este transcriptor fonológico realizará automáticamente la anotación que posteriormente será verificada y validada por lingüistas. Al final se obtendrá una transcripción fonológica (que no fonética) y una segmentación en sílabas. El programa desarrollado se empleará en futuros proyectos.

El último de los niveles de anotación que estamos acometiendo es el nivel semántico. En este caso es el tema de la tesis de Manuel Alcántara Plá que será defendida en julio de 2005. En concreto su "anotación y recuperación de información semántica eventiva de corpus" propone un sistema de anotación general aplicable tanto a corpus orales como escritos basado en el análisis de estructuras eventivas. Los fundamentos teóricos de esta tesis descansan en Pustejovsky y especialmente en Moreno Cabrera (veáse Moreno Cabrera 2004). La aportación de Alcántara ha consistido en diseñar el sistema de etiquetas y realizar la anotación manual de más de 50.000 palabras del corpus C-ORAL-ROM, así como herramientas informáticas para automatizar la operación. Los resultados que aporta son de gran alcance considerando la gran cantidad de datos anotados.

BIBLIOGRAFÍA CITADA

ALCÁNTARA PLÁ, M. (2005): *Anotación y extracción de información semántica eventiva en corpus. Tesis Doctoral*, Universidad Autónoma de Madrid.

CRESTI, E. y MONEGLIA, M. (eds.) (2005*): C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam, John Benjamins.

DUNNING T. (1993) 'Accurate methods for the statistics of surprise and coincidence'. *Computational Linguistics* 19(1): 61-74.

GONZÁLEZ, A. DE LA MADRID, G., ALCÁNTARA, M., DE LA TORRE, R, MORENO, A. (2004): "Orality and difficulties in the transcription of a spoken corpus", en los Proceedings of the Worshop on Compiling and Processing Spoken Corpora, LREC-2004, Lisboa.

GUIRAO, J.M, MORENO, A., GONZÁLEZ, A., DE LA MADRID, G. & ALCÁNTARA, M. (in press) "Relating linguistic units to socio-contextual information in a spontaneous speech corpus of Spanish", in *Corpus Linguistics Across the World*. Amsterdam, Rodopi.

GUIRAO, J.M. y MORENO, A. (2004): "A Toolbox for tagging the Spanish C-ORAL-ROM corpus", en los Proceedings of the Worshop on Compiling and Processing Spoken Corpora, LREC-2004, Lisboa.

MORENO CABRERA, J.C. (2004): *Semántica y gramática: sucesos, papeles semanticos y relaciones sintácticas*, Madrid, A. Machado Libros.

MORENO SANDOVAL, A. (1991): *Un modelo computacional basado en la unificación para el análisis y generación de la morfología del español*. Tesis Doctoral, Universidad Autónoma de Madrid.

MORENO SANDOVAL, A. (2002): "La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM", en Actas de las Segundas Jornadas de Tecnologías del Habla, Univ. de Granada.

MORENO SANDOVAL , A. and GUIRAO, J.M. (2003). Tagging a spontaneous speech corpus of Spanish, en los Proceedings of RANLP 2003. Borovets, Bulgaria

PÉREZ MILANS, M. y MORENO SANDOVAL, A. (2004): "Diseño de corpus orales de lenguaje infantil: análisis comparado de CHILDES y CHIEDE", presentado en el simposio anual de la Sociedad Española de Lingüística, Madrid.

TORRE, D., CAMPOS, E., MORENO, A., COLÁS, J. y GARRIDO, J. (2004): "Resultados preliminares de la decodificación fonética sobre distintos tipos de habla espontánea", en *Actas de las Terceras Jornadas sobre Tecnologías del Habla*, Valencia.