C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages.

Entidad: Unión Europea.

Grupos participantes: Università di Firenze, Université de Provence, Fundação da Universidae de Lisboa, Universidad Autónoma de Madrid, Instituto Cervantes, European Language Resources Association Agency SARL, Pitch Instruments France SARL, Editions Honoré CHAMPION, Istituto Trentino di Cultura.

Responsable: Dr. Antonio Moreno Sandoval. Profesor titular de la Universidad Autónoma de Madrid. (sandoval@maria.lllf.uam.es, 913975250).

Resumen del proyecto:

C-ORAL-ROM es un corpus multilingüe de habla espontánea. Está formado por grabaciones en cuatro lenguas romances (francés, español, italiano y portugués) que suman más de 1.200.000 palabras y su objetivo principal es el de proporcionar cuatro corpus comparables.

El proyecto está financiado por la Unión Europea y está siendo desarrollado por un consorcio europeo coordinado por la Universidad de Florencia. El resultado final se presentará en formato DVD, con un formato homogéneo e incluirá herramientas informáticas para el análisis del texto y de la señal acústica así como estudios de lingüística comparativa.

A diferencia de la mayoría de los corpus orales existentes, C-ORAL-ROM no se centra en grabaciones con contextos planificados (como, por ejemplo, los teléfonos de información), sino que busca la mayor varibilidad y espontaneidad posibles. A pesar de ello, se intenta mantener una calidad acústica elevada para facilitar su utilización. La varibilidad intenta reflejar las características reales del habla, donde las conversaciones no se restringen a unos dominios semánticos preestablecidos y la prosodia incluye un gran rango de posibilidades.

La varibilidad es, precisamente, la mayor dificultad a la hora de conseguir cuatro corpus comparables. Para lograr dicha comparabilidad, se han establecido cantidades de palabras respetando cinco parámetros:

1) Estructura dialógica (concretamente: monólogos, diálogos y conversaciones con más de dos hablantes).

2) Dominio social (familiar-privado, público, medios de comunicación).

3) Variación de géneros (entrevistas, telediarios, etc.).

4) Distinción entre textos formales e informales.

5) Caracterización de los hablantes (edad, sexo, educación, ocupación y procedencia).

Las proporciones han sido exactamente las mismas en los cuatro corpora. Es evidente que la comparabilidad absoluta en cuanto a dominios semánticos será muy difícil, pero sí las probabilidades de que aparezcan fenómenos de cualquier nivel lingüístico (fonético, sintáctico, pragmático, etc.).

Debido al tamaño reducido de los corpora y aunque se registra la procedencia de los hablantes, C-ORAL-ROM no pretende mostrar aspectos diatópicos, sino acercarse a un estándar en cada una de las lenguas. De este modo, se han registrado mayoritariamente las siguientes variedades: francés del sur, español de Castilla, portugués continental e italiano toscano. No se esperan, por lo tanto, grandes variaciones fonéticas dentro de cada corpus.

Los corpora han sido transcritos textualmente con anotaciones sobre la prosodia. Cada texto está formado por una cabecera con la información sobre la transcripción, su grabación y los participantes, y por el texto transcrito dividido en turnos dialógicos y, dentro de ellos, por enunciados. Además, aparecen comentarios vinculados a los turnos con información contextual.

Los elementos prosódicos que han sido etiquetados son los enunciados (que se corresponden con actos de habla), las variaciones en la curva entonativa, las palabras incompletas y los reinicios, y los solapamientos.

Este corpus multilingüe anotado está alineado con el sonido para facilitar su estudio. La unidad que se ha utilizado para el alineamiento ha sido el enunciado. Se ha preferido dicha medida al uso de las palabras como unidades por las características propias del habla espontánea, que incluyen la coarticulación de las palabras en unidades prosódicas. El alineamiento de una sola palabra habría podido parecer artificial. La idea fundamental en la que se basa C-ORAL-ROM es que, mientras los textos escritos se estructuran sintácticamente, los textos de habla espontánea se caracterizan fundamentalmente por los enunciados, entendidos estos como actos de habla pragmáticos.

El formato acústico utilizado está determinado por el objetivo del proyecto de reutilizar grabaciones existentes además de aquellas que se han grabado específicamente para la elaboración de este corpus. Las cuatro universidades implicadas poseían corpus orales anteriores. Las características del formato son: mono/estéreo, con una frecuencia de 22050 Hz y 16 bits. En el caso del corpus español, se decidió realizar todas las grabaciones para mantener una calidad acústica homogénea y digital.

El alineamiento del texto y el sonido así como el análisis de la onda acústica se realiza a través del software WinPitchCorpus, implimentado dentro del presente proyecto. La herramienta está diseñada para el tratamiento de grandes corpora y elabora bases de datos en formato XML.

Una última característica importante que resaltar en el proyecto son las exigencias legales que se han establecido, fundamentada en la autorización por escrito de todos los participantes para la grabación y transcripción de su voz, así como la publicación de ésta.

C-ORAL-ROM pretende ser un instrumento de gran utilidad para toda persona interesada en el estudio de las lenguas romances y facilita la descripción de cada lengua en todos sus niveles.