Aspectos prácticos, tecnológicos y legales en la construcción de corpus de habla espontánea: la experiencia del proyecto C-ORAL-ROM

Antonio Moreno Sandoval
Manuel Alcántara Pla

Universidad Autónoma de Madrid

Entre enero de 2001 y abril de 2004 se ha compilado el mayor corpus oral multilingüe hasta la fecha. Está formado por cuatro lenguas romances (español, francés, italiano y portugués) y en total recoge más de 1.200.000 palabras, 120 horas de grabación y 770 textos. El proyecto ha sido financiado por la Unión Europea dentro del V Programa Marco (Tecnologías de la Sociedad de Información) con un presupuesto total de más de un millón de euros. Han participado nueve equipos, entre los que destacan los cuatro que han recogido los corpus y son propietarios de los derechos de autor: la Universidad de Florencia (coordinadora), la Universidad de Aix-en-Provence, el Centro de Lingüística de la Universidad de Lisboa y la Universidad Autónoma de Madrid. También han participado la empresa WinPitch (desarrolladora de software para la trascripción y alineamiento), el Instituto Cervantes (como evaluador de los recursos en su aplicación a la enseñanza de lenguas) y ELDA (organismo europeo de distribución de recursos lingüísticos en formato electrónico). La editorial John Benjamins participa en la publicación del corpus en libro y DVD.

Realizar un proyecto complejo como éste ha supuesto enfrentarse a problemas de todo tipo, desde metodológicos (cómo conseguir un formato común y una estructura común en los cuatro subcorpus, partiendo de tradiciones diferentes) hasta posibles problemas legales con respecto a la autoría y la privacidad de los participantes. El objeto de esta comunicación es presentar un inventario de problemas y posibles soluciones en la elaboración de corpus de habla espontánea, basado en nuestra experiencia.

En primer lugar se tratarán los aspectos de diseño: si el objetivo central es elaborar un corpus comparable entre cuatro lenguas, de manera que pueda utilizarse para hacer estudios contrastivos, el primer requisito es establecer unos patrones en común. En concreto, la estructura del corpus, el formato y normas de transcripción, el formato y normas de etiquetado.

En segundo lugar se abordará el aspecto tecnológico. Por ejemplo, qué características acústicas y situacionales deben darse para considerar una grabación adecuada. Por otra parte, se trata de un corpus completamente digitalizado, donde se accede por ordenador a la información, tanto a la fuente acústica como a la transcripción, como a los estudios de frecuencias y datos sociolingüisticos de los participantes. De hecho, el corpus se comercializará en DVD, en una versión multimedia. Esto ha obligado a enfrentarse a problemas tecnológicos como el de alinear el sonido con la transcripción, especialmente complicado cuando se da el solapamiento de varios participantes.

Por último, veremos las cuestiones legales que afectan a corpus de este tipo. Dado que C-ORAL-ROM se diseñó para ser utilizado como corpus de referencia, de manera que cualquier investigador o tecnólogo lo pueda usar o citar, los participantes tenían que dar su consentimiento por escrito, dado que de otro modo se podría incurrir en una violación al derecho de privacidad (es el caso de las grabaciones en contexto familiar o privado) o al derecho de propiedad intelectual (en el caso de las grabaciones en entorno público o medios de comunicación). Se explicará cómo se gestionaron los permisos y se dará el modelo de carta de consentimiento para la grabación, transcripción y comercialización de habla espontánea.