Tratamiento de los datos: transcripción y etiquetado

La transliteración de las entrevistas es una tarea ineludible que facilita la posterior labor de codificación de las variables lingüísticas. Con todo, el trabajo de transcripción se ve dificultado por el hecho evidente de que las normas ortográficas empleadas en la escritura no son adecuadas ni suficientes para dar cuenta, a través del canal visual-gráfico, de la mayor parte de las interacciones lingüísticas que se producen a través del canal auditivo-oral.

Como es sabido, las convenciones ortográficas se adaptan de manera más o menos cómoda a los estilos de habla formales, especialmente si se trata de emisiones producidas a través del código escrito. Sin embargo, los registros orales más espontáneos presentan una serie de características que hacen difícil su acomodo al código escrito: desviaciones de los usos normativos de la lengua utilizada, vacilaciones, presencia de pausas y silencios sin justificación sintáctica, aparición abundante de palabras cortadas y autocorrecciones, suspensiones y abandonos voluntarios de turno. Todos estos fenómenos, así como la presencia de interrupciones y solapamientos entre las intervenciones de los hablantes son acontecimientos de muy frecuente aparición en las entrevistas propias de los estudios sociolingüísticos como el que aquí se presenta y obligan a los investigadores a valerse de herramientas de marcación, más o menos exhaustivas, que permitan reflejar en las transposiciones al código escrito la aparición de elementos propios del código oral.

Los textos usados en la investigación PRESEEA-MA siguen las directrices metodológicas del proyecto PRESEEA que culminaron con la publicación en formato digital de las "Marcas y etiquetas mínimas obligatorias. Vers. 1.2. 17 - 02 - 2008"[1]. En ellas se recomienda el empleo de unas convenciones de transliteración elementales que incluyen el etiquetado de los textos mediante el uso de las normas internacionales de marcación textual de la Text Encoding Initiative (TEI)[2], ya que se trata de un sistema implantado en los medios industriales y de investigación de un gran número de países. Se trata de un sistema de transcripción estándar que asume los fundamentos del *Standard Generalized Markup Language* (SGML), según los cuales, aquellos fragmentos de texto que se desean marcar deben ir precedidos y seguidos por una etiqueta doble, es decir, un membrete que aparece entre paréntesis angulares y que resume el fenómeno acaecido o el rasgo textual en el que se insiste. La etiqueta de cierre contiene, además, una barra oblicua que evita la confusión con una posible marca de apertura con la misma leyenda.

Los textos que ahora se ofrecen muestran las diversas etapas por las que ha pasado el proceso de marcación textual del Proyecto PRESEEA. De este modo, en los córpora de los niveles de instrucción bajo y medio el sistema de etiquetado es ligeramente diferente al que se emplea en el nivel de instrucción superior que, como decimos, supone la culminación del proceso[3]. En cualquier caso, todos los textos han superado diversas fases de ensayo del sistema de etiquetado propuesto y han pasado por varios filtros de

[1] http://preseea.linguas.net/Portals/0/Metodologia/Marcas_etiquetas_minimas_obligatorias_1_2.pdf [última consulta 2 de febrero de 2014]

[2] Cf.: Sperberg-McQueen - Burnards (eds.) (2002).

[3] En la introducción de las ediciones de los corpus correspondientes a los diferentes niveles de instrucción se encuentra la explicación pormenorizada de las normas de etiquetado empleadas en cada caso. Vida Castro (ed.), 2007: 38-66; Ávila Muñoz, Lasarte Cervantes y Villena Ponsoda (eds.), 2008: 29-84; Lasarte Cervantes, Sánchez Sáez, Ávila Muñoz y Villena Ponsoda (eds.): 29-100.

corrección. Con todo, la tarea de marcación de los materiales podría alargarse eternamente si siguiéramos sometiéndolos a nuevas y profundas revisiones. El corpus editado, en cualquier caso, sigue unas convenciones de etiquetado cuyo objetivo principal es el de facilitar el análisis de los datos lingüísticos.