Diseño y elaboración de una base de conocimiento para una lengua histórica. Aplicación a la recopilación de un corpus paralelo

Javier Martín Arista

Esta charla presenta una base de conocimiento elaborada a partir de la información estructurada contenida en varias bases de datos léxicas, tanto de tipo diccionario (base de datos de lexicología y morfología; base de datos de indexación de fuentes secundarias) como de tipo corpus (lematizador). En la primera parte se insiste en la naturaleza fragmentaria e inconsistente de los datos disponibles en las fuentes primarias y secundarias del inglés antiguo, lengua de análisis y discusión en esta presentación. En efecto, mientras que la escasez de datos excluye su tratamiento con procedimientos basados en la estadística, la falta de uniformidad en la ortografía de los textos impide la búsqueda y comparación directa de la información. En la primera parte también se explica que existen corpus, incluso anotados, pero no hay un corpus lematizado disponible, con lo que no es posible vincular la información de las formas atestiguadas a la información de los lemas correspondientes. En la segunda parte se describen los campos en los que se consigna la información en las distintas bases de datos, así las distintas presentaciones, las relaciones que se establecen entre los componentes de la base de conocimiento y las opciones de búsqueda. Por último, se muestra la aplicación de la base de conocimiento a la recopilación de un corpus paralelo, alineado y anotado inglés antiguo-inglés contemporáneo. Las conclusiones valoran los logros alcanzados y los aspectos pendientes de investigación futura, e insisten en la aplicabilidad del modelo de base de conocimiento a otras lenguas.