Diseño de un corpus textual para el estudio de la lengua antigua

Prof. Dr. Joan Torruella

El esquema que propuso el conferenciante para su exposición es el que sigue a continuación:

1.- Presentación
    1.1. Humanidades digitales; 1.2. Corpus
1.3. Lingüística de corpus
2.- Corpus
    2.1. ¿Qué es un corpus?
    2.2. Aportaciones de los corpus
    2.3. Tipos de corpus
    2.4. Corpus de lectura
3.- Características del corpus
    3.1. Finalidad
    3.2. Medios
    3.3. Representatividad y equilibrio
    3.4. Codificación
    3.5. Estructura
4.- Selección del corpus
    4.1. Recopilación de las obras
    4.2. Recopilación de los documentos
    4.3. Recopilación de las muestras
    4.4. Filiación de los documentos
5.- Edición de los documentos
    5.1. Tipos de edición
    5.2. Normas de edición
6. Conclusiones

La presentación de los estudios de Lingüística de corpus ocupó la primera parte de la ponencia, pues la irrupción de las nuevas tecnologías ha ejercido una modificación de los estudios lingüísticos. Sobre todo, ha sido la difusión del ordenador personal la que ha modificado las herramientas y los estudios.

En los últimos años hemos pasado de ordenadores pioneros que no podían usar programas a los que actualmente funcionan en pequeñas computadoras de sobremesa o portátiles. No solo no disponíamos de ordenadores sino que tampoco contábamos con software. Por supuesto, no se disponía de la red de internet. De este modo, con computadoras y la red de redes (www) se puede trabajar de forma muy distinta a la "filología de sillón". Se pueden presentar los resultados de modo digital. Internet ha sido una revolución parecida a la Imprenta, pero tal vez haya que decir que han sido dos revoluciones:

1) se expande el saber a muchas personas, bibliotecas, etc.; y

2) ha cambiado el soporte, que se parece más al cambio del rollo de pergamino por el códex (sistema libre de páginas) en el siglo VI.

Con internet, hemos entrado en otra dimensión o en otro paradigma de la filología: la Lingüística de corpus. Por tanto, los corpus han pasado al primer plano de la investigación y se trabaja en la construcción y análisis de corpus.

Los corpus ya fueron usados por Samuel Johnson (siglo XVII), quien se puso a elaborar un diccionario (con 40.000 lemas que estaban en los libros). El jesuita Alexander Cruden (18 monjes, 18 horas al día y dos años para elaborar la concordancia de la Biblia). Roberto Busa (1949) preparó las concordancias con ordenadores para todas las obras de Sto. Tomás.

Ciertamente, entre los expertos hay distintas definiciones de corpus en esta materia, por lo que conviene revisar las siguientes:

a) Eagles, 1996: "a corpus is a collection of pieces of language that are selected and ordered according to explicit linguistica criteria in orden to be used as a sample of the language". Aquí falta añadir la importancia de la informática.

b) Aquilino Sánchez ha añadido que la LC precisa la Informática: "Un corpus lingüístico es un conjunto de datos lingüísticos (pertenecientes al uso oral o escrito de la lengua, o ambos), sistematizados según determinados criterios, suficientemente extensos en amplitud y profundidad de manera que sean representativos del total del uso lingüístico o de alguno de sus ámbitos, y dispuestos de tal modo que puedan ser procesados mediante ordenador con el fin de obtener resultados varios y útiles para la descripción y el análisis."

En cambio, no es un corpus en la LC la colección o recopilación de documentos informatizados. Eso es un "Archivo informatizado". Tampoco es un corpus la Biblioteca Virtual Miguel de Cervantes, que debemos denominar Biblioteca de Textos Electrónicos. Mientras, un *Computer corpus* o *Corpus informatizado* es el CORDE, donde hay unos criterios previamente definidos que se respetan.

El ponente trató con detenimiento y prolijamente las aportaciones de los corpus informatizados, cuya trascendencia radica en

a) La fuente de recuperación de datos

b) El banco de pruebas de hipótesis

c) Su carácter de metodología para crear sistemas robustos de procesamiento de la lengua natural.

Es obvio que la multiplicidad de parámetros que se han aplicado en los distintos tipos de corpus tuvo que ser examinada por el Prof. Torruella en la exposición:

a) Modalidad (lengua oral / lengua escrita / modalidad mixta)

b) Temática: General / Especializado (genérico / canónico)

c) Época: Histórico / Contemporáneo

d) Temporalidad: Diacrónico / sincrónico

e) Magnitud: Grande / Restringido; Cerrado / Abierto / Monitor

f) Distribución: Equilibrado / Proporcional / Piramidal

g) Número de ediciones: Monoedición / Pluriedición (Relación entre lenguas: Comparable / Paralelo / Alineado

h) Número de lenguas (mono o plurilingüe)

i) Tipo de ediciones (facsimilar, paleográfica, normalizada, crítica): el proyecto internacional de CHARTA muestra tres ediciones.

j) Muestras (textual, referencia, léxico)

k) Información extralingüística

l) Marcaje o etiquetado (codificación)

De esta manera, el investigador va construyendo el Corpus provisional frente a los Corpus definitivos y a los llamados Corpus de lectura. Indudablemente, hay pros y contras a la hora de establecer los criterios de confección de corpus. El Corpus de lectura está constituido por el conjunto de los textos que nos permiten interrogar el corpus.

En el Corpus de estudio de la lengua antigua se ofrecen los textos para que puedan familiarizarse los investigadores con lo que pueden investigar. Los criterios de edición utilizados se han definido teniendo en cuenta que los resultados de las consultas del corpus debían servir, principalmente, para estudios filológicos sobre la lengua desde los puntos de vista más variados: gráfico/fonético, lexicográfico y morfosintáctico, por ejemplo.

En conclusión, el Prof. Torruella ofreció una erudita y densa exposición sobre la nueva metodología de la investigación en Historia de la Lengua Española, con las nuevas herramientas que ofrece la Lingüística de corpus y la informatización o digitalización de la nueva era en que nos hallamos. Aquella filología o lingüística que se realizaba en los archivos, con la bibliografía impresa y sin apoyos técnicos se ha transformado hoy en una filología digital, que ha generado el área de las Humanidades Digitales. Tanto las bibliotecas digitales en marcha, como los bancos de datos y los proyectos de investigación que ofrecen al estudiante y a los investigadores un Corpus de documentos codificados con criterios lingüísticos y filológicos, se convierten en los pilares de una nueva metodología para la investigación y la docencia universitaria. Creemos que la ponencia del Prof. Dr. Torruela sirvió para afianzar en los oyentes (investigadores, estudiantes, profesores, etc.) la imperiosa necesidad de colaborar en la elaboración de Corpus lingüísticos y de partir de tales corpus en cualquier proyecto universitario, sea de carácter investigador, didáctico, pedagógico, histórico-lingüístico, sociolingüístico o lexicográfico, por citar solo algunas perspectivas que se nos ofrecen en el Horizonte del Humanismo Digital.