Cuadernos
del Instituto Historia
de la Lengua

Una introducción a la Lingüística de Corpus*

Teubert, Wolfgang and Anna Cermáková (2007):
Corpus Linguistics. A short introduction, Continuum,
London, 153 pp., HB: 0-8264-9480-3.

Corpus Linguistics. A short introduction constituye una de las partes del libro *Lexicology and Corpus Linguistics*, de los autores Michael A. K. Halliday, Colin Yallop, Anna Cermáková y Wolfgang Teubert, obra publicada por la Editorial Continuum en 2004.

Corpus Linguistics. A short introduction está dividido en dos partes, «Language and corpus linguistics», del profesor Wolfgang Teubert y «Directions in corpus linguistics», escrita conjuntamente por Wolfgang Teubert y Anna Cermáková. El volumen incluye un glosario de términos de Lingüística de Corpus, una amplia sección de referencias bibliográficas y un índice en el que se recogen las páginas de aparición, a lo largo del texto, de determinados autores y conceptos lingüísticos.

La obra, como el título indica, pretende ser una introducción a la Lingüística de Corpus, ideal para estudiantes universitarios o docentes que se inician en el conocimiento de esta disciplina lingüística.

La primera parte comienza con un capítulo que contiene algunas reflexiones en torno a la teoría del lenguaje natural de Chomsky. Encabezado por la pregunta retórica «¿Son todas las lenguas lo mismo?», este capítulo realiza un breve recorrido por algunos postulados sobre los que la gramática y el lenguaje se han erigido a lo largo de la historia, especialmente aquellos surgidos con el Generativismo a partir de los años 60. El autor se plantea, entre otras cuestiones, si es adecuado concluir que todas las lenguas son similares simplemente porque podemos describirlas con términos iguales.

El siguiente capítulo, que trata varios aspectos de lingüística y significado, aborda algunas diferencias existentes entre gramática y vocabulario, notables sobre todo en el aprendizaje de segundas lenguas: si bien resulta relativamente fácil construir oraciones correctas desde el punto

* Este trabajo se integra dentro del proyecto HUM2007-6070/FILO, financiado por el Ministerio de Ciencia e Innovación.

de vista gramatical, no lo es tanto desde el punto de vista semántico, puesto que para traducir una palabra, los diccionarios bilingües ofrecen múltiples opciones y pocas instrucciones. De acuerdo con el autor, esta dificultad de elección estriba en que la mayoría de las palabras, por sí solas, no constituyen unidades de significado y, por este motivo, en los diccionarios aparecen registradas como palabras polisémicas.

El tercer capítulo ahonda en la importancia de la Lingüística de Corpus para la detección de las colocaciones y las frases idiomáticas de una lengua, así como en las ventajas de su aplicación en el ámbito de la lexicografía, ya que los datos obtenidos en las búsquedas en corpus informatizados nos muestran que el lenguaje es mucho más idiomático de lo que pensamos. Posteriormente, en el capítulo titulado «Lingüística de Corpus: una mirada diferente al lenguaje», el profesor Teubert nos explica que, para la Lingüística de Corpus, «el significado es, como el lenguaje, un fenómeno social» (pág. 37). Desde este punto de vista, la Lingüística de Corpus, a diferencia del Generativismo y de la Lingüística Cognitiva, estudia el *significado* en el discurso, de forma que el significado de una palabra o secuencia de palabras sería la suma de todo lo que se ha dicho sobre esa palabra o secuencia, tanto por los hablantes como en los diccionarios, es decir, que el *significado* «es algo que puede ser discutido por todos los miembros de la comunidad del discurso». Además, señala que con la Lingüística de Corpus, la *palabra*, tal y como la entendemos actualmente, deja de ser la unidad central del lenguaje para serlo la *palabra* con su significado original de *logos* o discurso.

La primera parte del libro termina con una breve historia de la Lingüística de Corpus, desde el corpus aún no informatizado *Survey of English Usage* de Randolph Quirk[1] (años 50), pasando por el corpus de Brown[2] (años 60), compilado por Nelson Francis y Henry Kucera, el LOB (Lancaster-Olso-Bergen), corpus de Inglés Británico (años 70), en el que colaboraron Leech, Hofland y Johansson y el *English Lexical Studies* (más conocido como OSTI report) de John Sinclair, proyecto iniciado en 1963 en la Universidad de Edimburgo y completado en la Universidad de Birmingham (Reino Unido) con el diseño del *Collins COBUILD English*

1 Materializado en lo que fue la gramática estándar del inglés durante muchas décadas: *A Comprehensive Grammar of the English Language* (Quirk *et al*., 1985).

2 En la Universidad de Brown en Providence (Rhode Island), publicado en 1967 como *Computational Analysis of Present-Day American English*.

Language Dictionary[3] (1987), el primer diccionario hecho enteramente a partir de un corpus.

Un capítulo sobre lenguaje y representatividad, en el que se hacen explícitas las limitaciones de todo trabajo con un corpus, da comienzo a la segunda parte de esta obra, «Direcciones en Lingüística de Corpus». A pesar de que no existe un corpus ideal que pueda recoger todo lo escrito y dicho sobre algo, el texto insiste en la importancia de la elección de un corpus para caracterizar un estado o variedad de lengua.

El siguiente capítulo detalla las características de distintos tipos de corpora, como el *corpus de referencia*, que contiene el vocabulario estándar de una lengua, el *corpus monitor*, que es aquel que está abierto, continuamente actualizándose, los *corpora paralelos*, especialmente útiles para el estudio de los mecanismos de traducción e *Internet*, al fin y al cabo, el corpus virtual más consultado. Seguidamente, Cermáková y Teubert abordan el *significado* desde la Lingüística de Corpus, es decir, el *significado en el discurso* a través de dos aspectos fundamentales: el *uso* y la *paráfrasis*. El uso, por un lado, puede ser detectado por el ordenador, que resuelve mediante porcentajes si una palabra constituye una unidad de significado o es simplemente una parte de esa unidad. Las paráfrasis contenidas en los textos, sin embargo, requieren una interpretación de los datos, para aceptar o no su validez, que sólo el hombre puede realizar.

En un capítulo posterior, los autores analizan la palabra *globalization* en 200 concordancias del Bank of English. El programa informático obtiene la frecuencia de combinación de la palabra *globalization* con otras, frecuencia que sirve de guía para descubrir las posibles unidades de significado. Asimismo, a partir de estas combinaciones y de los aspectos semánticos que encubren, se puede llevar a cabo la interpretación de *globalization*, y por lo tanto, de su significado, en el propio discurso, sin pasar por alto que ese significado será siempre aproximado.

Seguidamente, se encuentran dos capítulos que albergan ejemplos prácticos de la aplicación de la Lingüística de Corpus en los ámbitos de la Lexicografía y la Traducción. En primer lugar, se estudian y solventan algunos problemas surgidos con la metodología lexicográfica tradicional en la inclusión de *friendly fire* en los diccionarios. A continuación, aparecen varios avances conseguidos gracias a la aplicación de las colocaciones

3 La historia de esta aventura se narra en la obra editada por Sinclair *Looking up: an account of the COBUILD project in lexical computing.*

obtenidas mediante la metodología de la Lingüística de Corpus en traducciones donde intervienen corpus paralelos.

Ya en la conclusión, los autores destacan algunos aspectos en torno al *significado* en Lingüística de Corpus, entre estos, que el significado «es un fenómeno social y no mental», que se puede obtener sólo en el discurso, que es algo negociable por los miembros de la comunidad del discurso y que siempre será parcial. Además, afirman que el significado de las unidades de significado es algo completamente distinto al entendimiento de estas unidades que cada ser humano puede hacer y verbalizar de manera personal.

En resumen, *Corpus Linguistics. A short introduction* es un libro claro y preciso, que contiene tanto una panorámica de la Lingüística de Corpus, desde su surgimiento hasta la actualidad, como casos prácticos en los que se exponen varias aplicaciones de la Lingüística de Corpus en Lexicografía y Traducción. La obra resulta especialmente interesante porque reflexiona sobre la concepción del lenguaje y del significado en algunas corrientes lingüísticas actuales y declara abiertamente los postulados de la Lingüística de Corpus al respecto. La singularidad de esta postura reside también en que ha sido el desarrollo de las herramientas informáticas lo que ha permitido esa nueva óptica desde la que observar el lenguaje y pensar en el *significado*.

Cristina Martín Herrero