

Recensió: *Korpuslinguistik* de Carmen Scherer
Ferran Robles i Sabater, Universitat de València (España)

Entre las formas de aproximación al estudio científico de la lengua que han experimentado un desarrollo más acelerado en los últimos años ocupa un lugar muy destacado el empleo de los corpus de textos, en cualquiera de sus variantes. La lingüística de corpus, que en otros tiempos fuera descartada por algunos autores, que no la consideraban capaz de proporcionar datos válidos para el análisis sistemático los hechos de la lengua, ha generado en la última década una cantidad relevante de investigaciones que demuestran su aplicabilidad a ámbitos tan diversos como la sociolingüística, la caracterización de los géneros discursivos, la teoría y la práctica de la traducción, la elaboración de diccionarios electrónicos o los aspectos cuantitativos y cualitativos de la adquisición y la enseñanza de los idiomas modernos. Por ello mismo resulta sorprendente que, a pesar de su gran implantación y del creciente número de seminarios universitarios y reuniones científicas dedicadas a su estudio, todavía son pocos los manuales de introducción al manejo y aprovechamiento de los corpus lingüísticos. La obra que aquí reseñamos sólo es, en realidad, la tercera de sus características publicada en Alemania y, junto a la recientemente aparecida *Korpuslinguistik: Eine Einführung* de Lemnitzer y Zinsmeister¹, la única dedicada de manera exclusiva al alemán.

El volumen de Carmen Scherer forma parte de la serie «Kurze Einführungen in die germanistische Linguistik», que, bajo la dirección de Jörg Meibauer y Markus Steinbach, produce desde hace tres años materiales didácticos de iniciación para alumnos universitarios alemanes de Filología y Traducción. Como los demás títulos de esta colección, en *Korpuslinguistik* no encontramos más que una breve y sucinta exposición, en aproximadamente 100 páginas, de los principales rasgos de esta rama de la lingüística aplicada. Su objetivo es doble: por una parte, pretende familiarizar al estudiante con toda una serie de conceptos y utilidades nuevos en el uso de las recopilaciones de textos de distinta naturaleza y, por otra, quiere estimular la creación de corpus propios y el análisis de los datos reunidos en éstos.

El libro consta de cinco capítulos y presenta una estructura lineal. El primer apartado es una introducción a la noción de la lingüística de corpus y a su valor y aplicabilidad para el estudio metódico de la lengua. Partiendo de una supuesta incapacidad de las teorías lingüísticas para ir más allá del establecimiento de unas reglas que determinen los usos gramaticales o agramaticales de la un idioma («Die Sprachwissenschaft gibt auf die Frage nach richtig oder falsch keine Antwort. Sie bewertet Sprache nicht, vielmehr unterscheidet sie zwischen Formen, die bestimmten sprachlichen Regeln entsprechen, und solchen, die dies nicht tun», p. 1), Scherer reclama que a una multiplicidad de fenómenos relacionados con la adecuación o no de determinadas realizaciones lingüísticas a contextos comunicativos concretos o de la frecuencia con que se manifiestan rasgos muy concretos en el habla de una comunidad dialectal o un grupo generacional o geosocial, solamente se les podrá dar una respuesta apropiada mediante la observación directa de las producciones de los usuarios del idioma: «Um eine Antwort auf diese Frage zu finden, muss man vielmehr den tatsächlichen Sprachgebrauch untersuchen und die Auftretenshäufigkeit eines bestimmten Phänomens ermitteln. Das bedeutet, man muss empirisch arbeiten» (p. 1).

La mayor parte de este capítulo está dedicado a la definición del concepto de corpus —«eine Sammlung von Texten oder Textteilen, die bewusst nach bestimmten sprachwissenschaftlichen Kriterien ausgewählt und geordnet werden» (p. 3)— y a los criterios que es necesario seguir en su construcción. La validez de un corpus para los fines de este modo de acercamiento al estudio del lenguaje está basada en su representatividad, la cual sólo podrá alcanzarse mediante la fijación previa de un

¹ Lemnitzer, Lothar/Zinsmeister, Heike: *Korpuslinguistik: Eine Einführung*. Tübingen, Narr, 2006.

objeto de análisis concreto, a partir del cual será posible determinar qué clases de textos hay que seleccionar y qué fenómenos característicos del uso de la lengua deben ser considerados. La nómina de ámbitos a los que puede ser aplicado el análisis de corpus es extensa: «die Erforschung von sprachlichen Strukturen und Varietäten, die Erstellung von Wörterbüchern, die Erstellung von Grammatiken, der Fremdsprachunterricht, die Übersetzung, die Computerlinguistik» (p. 10). El apartado incluye, además, algunos ejemplos de empleo de los corpus en campos como la gramática o la lexicografía.

El capítulo segundo describe y comenta los diferentes tipos de corpus lingüísticos. Para ello parte de una clasificación de éstos basada en los fines específicos para los que cada corpus es concebido: «Da ein Korpus immer im Hinblick auf einen bestimmten Verwendungszweck erstellt wird, hat jedes Korpus eigene Charakteristika» (p. 16). Scherer no establece una taxonomía de corpus, sino más bien unos criterios formales que permiten su descripción por oposición. Entre las distinciones más importantes que identifica la autora encontramos: corpus informatizados vs. no informatizados, corpus completos vs. parciales, corpus anotados vs. no anotados, corpus escritos vs. hablados, corpus de la lengua actual vs. históricos o corpus monolingües vs. multilingües. Los diferentes subapartados de este capítulo revisa cada uno de estos criterios. En la sección referida al contraste de los corpus monolingües y multilingües resultan de especial interés los comentarios de la autora sobre las posibilidades que ofrecen algunas de estas recopilaciones de textos, como el corpus paralelo bilingüe de la lengua escrita *English/German Translation Corpus* de la Universidad de Chemnitz (con textos procedentes de los ámbitos de la política, la ciencia y el turismo) y del repertorio multilingüe OPUS Corpus diseñado por Jörg Tiedemann (Universidad de Uppsala), mediante los cuales son posibles los estudios contrastivos y de traducción entre dos o más idiomas.

El tercer capítulo está dedicado a las formas de análisis de los materiales contenidos en un corpus. El empleo de corpus lingüísticos permite una aproximación tanto cuantitativa como cualitativa a los datos considerados y abre las puertas al estudio de un mismo hecho de la lengua desde una multiplicidad de perspectivas. El método a seguir en cada caso dependerá directamente del punto de vista adoptado por el investigador: «Der wesentliche Unterschied zwischen qualitativen und quantitativen Korpusanalysen besteht nicht darin, welche Fragestellungen untersucht werden, sondern wie diese untersucht werden» (p. 35). A los estudios realizados desde un punto de vista cuantitativo —centrados, especialmente, en la frecuencia de aparición de un elemento dentro del corpus (o de una parte de éste) y las posibles conclusiones que resultan del análisis de estos datos, como, por ejemplo, la determinación de la productividad de un patrón lingüístico, una combinación léxica o una variación sintáctica o gramatical—, se oponen los de carácter cualitativo —que pueden referirse a la constatación de un determinado fenómeno, su clasificación o su interpretación. Como ejemplos de empleo de corpus informatizados, Scherer cita el análisis de concordancias, que permite examinar una unidad léxica dentro de su contexto de uso, y el de las colocaciones, que proporciona material para un estudio tanto cuantitativo de la frecuencia de aparición conjunta del lexema con otras unidades léxicas como un análisis cualitativo que permita determinar los principales rasgos de los términos con los que tiende a combinarse (categoría gramatical, contenido semántico, adscripción a un campo léxico, etc.).

El capítulo cuarto describe en detalle el trabajo con corpus lingüísticos de creación propia y lo analiza en cada una de sus fases: 1) formulación de una problemática, 2) construcción de un corpus, 3) procesamiento del corpus, 4) recolección de los datos relevantes del corpus, 5) procesamiento de los datos y 6) análisis e interpretación de los datos. La autora defiende la conveniencia de elaborar una colección propia de textos cuando los corpus informatizados de referencia cuentan con poca información sobre el objeto de la investigación —como puede ser, por ejemplo, una cierta variedad de la lengua, un habla dialectal, una jerga profesional o la forma expresión de una

determinada capa de la societat. Esta recopilació debera ser capaç de oferir una quantitat rellevant de mostres i exemples representatius del fenomen estudiat: «Die wichtigsten Kriterien für den Aufbau eines Korpus sind dessen Größe, Inhalt, Beständigkeit und Repräsentativität» (p. 56). Otro aspecte destacat per la autora es el referit a la anotació de los textos previa al anàlisi de los ejemplos, que possibilita su codificació y clasificació.

El capítol quinto nos introdue en el treball con corpus lingüístics ya existents. La autora analiza en detall los corpus de referencia de uso más frecuente dentro de la Germanística (DWDS, Cosmas II y Tiger) y comenta las ventajas y los inconvenientes de emplear la red informática como fuente de datos si no se aplican los filtros de análisis necesarios sobre la procedencia de las muestras extraídas y el tipo de texto al que pertenecen, así como la edad, nivel cultural, extracción social, procedencia geográfica, etc., de sus autores. Entre tantas desventajas, Scherer apunta un par de elementos positivos del aprovechamiento de Internet como corpus lingüístico: la web nos permite realizar determinados estudios de carácter cualitativo — aunque no una valoración cuantitativa del fenómeno que analizamos—, y nos da acceso a una cantidad inmensa de realizaciones lingüísticas auténticas sobre fenómenos que son difíciles de corroborar mediante la información contenida en los corpus de uso habitual (p. 75).

El libro se cierra con una serie de informaciones sobre el acceso a los corpus lingüísticos informatizados y su empleo, así como otros enlaces a obras de análisis del trabajo con corpus.

Como corresponde a la naturaleza de este tipo de obras y su destinatario preferencial, *Korpuslinguistik* se caracteriza por su claridad y sencillez, no exenta de cierta erudición en apartados concretos, como el referido a las aplicaciones de la lingüística de corpus o el que versa sobre la recopilación y clasificación, así como el procesamiento de los datos lingüísticos. Incorpora numerosos ejemplos y uno o más ejercicios prácticos en cada subapartado, que permiten una mayor implicación del alumno con las cuestiones expuestas. Se echa de menos, no obstante, una mayor profundización en cuestiones que lógicamente debería incluir una obra dirigida a universitarios, como el análisis cuantitativo de datos, al que aquí sólo se alude, pero omitiendo numerosos detalles sobre su aplicación, o sobre las relaciones de lingüística de corpus con otras ramas de la lingüística aplicada, que se menciona de forma casi anecdótica al principio de la obra y que hubiera merecido un capítulo propio en un manual de iniciación de estas características.

Con todo, el libro de Scherer reúne de forma sintética todos los aspectos esenciales que deben tenerse en cuenta en una primera aproximación a la disciplina y resultará atractivo a quien aborde su lectura con escasos o nulos conocimientos previos de la materia, especialmente a los lectores no avezados a este tipo de herramienta que quieran familiarizarse con su empleo con finalidades docentes o investigadoras.

Bibliografía

Scherer, Carmen (2006): *Korpuslinguistik*. Heidelberg: Winter. (Kurze Einführungen in die Germanistische Linguistik, 2).