Guillermo Rojo

Introducción a la lingüística de corpus en español

London-New York, Routledge 2021, 380 p.

Beatriz Gómez-Pablos [gomezpablos@fedu.uniba.sk]
Univerzita Komenského, Eslovaquia

HTTPS://DOI.ORG/10.5817/ERB2023-1-24

La obra de Guillermo Rojo no solo es pionera en el ámbito de la lengua española, sino que constituye un pilar de referencia para la lingüística de corpus. *Introducción a la lingüística de corpus en español* se divide en siete capítulos, que presentan un breve "resumen" al comienzo, un apartado de "lecturas complementarias" y otro sobre "cuestiones, problemas y temas de investigación" al final de cada capítulo, además de un apartado de "notas". Esto refleja en parte la marcada orientación didáctica que el autor ha querido conferir al libro, pues como apunta "está dirigido fundamentalmente a estudiantes de los últimos cursos de grados vinculados a la lingüística española, estudiantes de máster o doctorandos que desean adquirir formación en este terreno o necesitan profundizar en él" (p. XX). Obviamente, la lectura de esta obra confirma que cualquier investigador, especialista o estudioso sacará gran beneficio de ella.

El primer capítulo se encarga de cuestiones introductorias como qué es un corpus, para qué sirve (investigación sobre elementos léxicos, clases de palabras y otras categorías gramaticales, aspectos semánticos, cuestiones diacrónicas, aspectos sociolingüísticos, etc.), tipos de corpus (generales o dialectales, sincrónicos o diacrónicos, orales o escritos, generales o especializados, cerrados o abiertos, etc.) y cómo y cuándo surge la lingüística de corpus.

El segundo capítulo, "La lingüística de corpus y la metodología de investigación lingüística", comienza cuestionando la clásica división entre letras y ciencias, propone en su lugar la división entre ciencias formales y ciencias factuales o empíricas, y trata de definir el concepto de *cientificidad* aplicándolo a las ciencias empíricas, de las que la lingüística forma parte. A continuación, Rojo describe el carácter de los datos lingüísticos y el análisis que lleva a cabo el lingüista (básicamente descripción, clasificación y explicación), subrayando que en el trabajo con corpus no se realiza "una criba previa que pueda condicionar luego la consideración del fenómeno ni distorsione las estadísticas cuando son necesarias" (p. 44). El uso de los ordenadores ha supuesto en este sentido una revolución instrumental; si bien eso no implica que la lingüística de corpus sea una teoría. Se trata –como indica el autor– de "una forma de acercamiento al estudio de los fenómenos y elementos lingüísticos fundamentada en ciertos supuestos acerca de qué aspectos del análisis son realmente relevantes" (p. 48). Rojo señala que las características de la lingüística de corpus son la objetividad, fiabilidad, replicabilidad y relevancia; y se detiene a presentar las diferencias de la lingüística de corpus con respecto a la lingüística racionalista y a la lingüística descriptiva tradicional.

El tercer capítulo se ocupa del "Diseño, construcción y explotación del corpus". El autor retoma aquí algunas preguntas esbozadas en el primer capítulo, que pasa ahora a describir más detalladamente. Rojo, después de ofrecer una nueva definición de *corpus*, comenta algunas cuestiones que se plantean antes de diseñar un corpus y que conforman sus características; por ejemplo, la cantidad y extensión de los textos, el género de textos, el tipo de textos (orales o/y escritos), etc. Precisamente la congruencia entre el diseño y los objetivos es lo

que permite valorar su idoneidad (cfr. 67). Nuestro autor describe los tipos de corpus y, entre otras cosas, aprovecha para comentar las dificultades que presenta la trascripción de los textos orales. Dedica además un subapartado a los corpus referenciales, donde hace algunas observaciones sobre las ventajas y desventajas de algunos de ellos. En los puntos siguientes aborda temas como la codificación, la anotación, explotación de los corpus y cuestiones legales y éticas. Rojo no escatima en los ejemplos y facilita de este modo la comprensión de todas estas cuestiones.

En el cuarto capítulo, el más extenso de la obra, el lector se adentra en el trabajo práctico con los corpus. El autor muestra las diversas formas en las que se puede extraer y analizar aspectos relacionados con el componente léxico del español: la frecuencia de formas ortográficas, de lemas, de expresiones complejas, la variación diatópica, diacrónica, diastrática y diafásica en el léxico, las coapariciones, y el análisis del significado. Aprovecha para explicar conceptos como *token* (cada una de las formas ortográficas que aparecen en un texto) y *type* (cada una de las formas diferentes contenidas en un texto), comenta que en un corpus "el aumento del volumen total no tiene paralelo con el aumento de las formas distintas" (p. 134), y apunta que siempre queda sin resolver el problema de las formas homógrafas. En este contexto surgen además cuestiones como: si se debe partir del infinitivo en los verbos o si se deben considerar todas las formas del paradigma verbal; cómo integrar las "unidades multipalabras" y las unidades fraseológicas; si incluir las "entidades nombradas" (nombres de personas, lugares, entidades sociales, políticas etc.), entre otros interrogantes. Rojo presenta numerosos ejemplos para cada caso y tablas con los resultados obtenidos. Es más, con gran sentido didáctico, muestra los pasos para llegar a dichos resultados y el corpus concreto que ha empleado. A su vez, a lo largo del capítulo subraya que no basta con constatar la frecuencia de un fenómeno léxico, sino que la interpretación correcta requiere una investigación más profunda para desentrañar lo que realmente sucede. Esto afecta sobre todo al análisis del significado.

En paralelo al anterior, el quinto capítulo se ocupa de la recuperación de información gramatical contenida en los corpus textuales. Las cuestiones que Rojo analiza aquí son: la frecuencia de clases de palabras, la frecuencia de categorías y subcategorías gramaticales (frecuencia de uso y frecuencia de inventario de las tres conjugaciones), la frecuencia de uso de los modos y tiempos verbales, la frecuencia de las perífrasis verbales, y otras cuestiones más concretas como: los adverbios en *-mente*; las expresiones *detrás de mí/detrás mío/mía*; la adaptación de préstamos (en singular y plural); las construcciones del tipo *se los dije* o *informar que/informar de que*; fenómenos gramaticales abordados desde la perspectiva diacrónica (como las formas en *-ra* y en *-se* o los superlativos en *-ísimo*), etc. Cierra el capítulo un conjunto de consideraciones sobre la posible utilidad del trabajo con corpus y las posibles aplicaciones en la enseñanza y aprendizaje de segundas lenguas. Todos estos ejemplos, permiten hacerse una idea de las grandes posibilidades que ofrecen los corpus, tanto para las investigaciones léxicas como para las gramáticas.

El penúltimo capítulo, "Otras cuestiones centrales en lingüística de corpus", versa sobre los antecedentes y evolución de la lingüística de corpus, sus ventajas e inconvenientes, la estructura estadística de los corpus, el tamaño, representatividad y equilibrio de los corpus, y termina con unas breves pinceladas sobre las perspectivas de futuro. Rojo menciona aquí cinco inconvenientes actuales que se pueden entender como retos de futuro: a) el hecho de que la ausencia de un elemento o estructura no permite deducir que tal elemento o estructura sea imposible en una lengua; b) la dificultad de trabajar con la ausencia de elementos; c) los corpus no pueden contener todo lo que es posible en una lengua; d) el lenguaje se presenta fuera de su contexto comunicativo (no aparecen los gestos, miradas, etc.); y e) en un corpus general puede haber textos mal seleccionados, ediciones poco adecuadas, codifica-

ciones insuficientes, etc. A estos inconvenientes se suman otros que nuestro autor ha ido anotando a lo largo de la obra.

Por último, el objetivo del séptimo capítulo, "Herramienta de recuperación de datos: resumen y ampliación", consiste en analizar algunas herramientas que permiten recuperar y procesar información obtenida directamente de textos o corpus textuales, sin la intermediación de las aplicaciones de consulta descritas anteriormente. Un "Glosario de términos" muy útil completa la obra.

Introducción a la lingüística de corpus en español merece el elogio de todo trabajo realizado con rigurosidad. Se trata de una obra de obligada consulta para todos los interesados en la materia; que, sin duda, agradecerán la visión panorámica, la explicación pormenorizada sobre cómo utilizar los diferentes corpus y la riqueza de ejemplos, tablas y gráficos que aclaran no solo de forma diáfana diversas cuestiones teóricas, sino que reflejan también el gran sentido didáctico de su autor.