Introducción a la lingüística de corpus en español

Guillermo Rojo

Series Editor: Carol Klee
Spanish List Advisor: Javier Muñoz-Basols

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Índice general

Prólogo

Cualquier presentación general acerca de la evolución de los estudios lingüísticos en la segunda mitad del siglo xx situará en un lugar muy destacado la reconfiguración de objetivos y métodos derivada de la obra de Noam Chomsky. Ese efecto es innegable y puede detectarse en los ámbitos más diversos de nuestra disciplina, incluidos aquellos en los que la influencia de la lingüística de orientación racionalista se materializa más bien en la reacción contra su modo de entender la investigación. Sin que se pueda negar la enorme importancia de la gramática generativa, hay que reconocer, además, la existencia de algunas otras tendencias que han contribuido a diseñar el panorama general que presenta la lingüística en la segunda década del siglo xxi. La lista puede ser bastante larga si atendemos a la considerable variedad disciplinar existente, pero creo que hay algunas corrientes que destacan de modo especial por sus repercusiones en ámbitos muy distintos. La primera de ellas puede ser el conjunto de las aproximaciones de carácter funcionalista, surgidas en muchos casos como consecuencia de la necesidad de revisar los presupuestos básicos del estructuralismo tradicional, excesivamente dependiente de la fonología y poco adecuado para comprender y explicar otros componentes de las lenguas. La sociolingüística ha producido un cambio radical en nuestra forma de considerar las lenguas y el modo en que deben ser estudiadas. La variación, una faceta innegable pero incómoda, en las concepciones clásicas se ha convertido en un aspecto central, determinante de todo lo demás, de nuestra comprensión de la estructura y funcionamiento de las lenguas. Además, ha contribuido decisivamente a remodelar los estudios de orientación diacrónica, considerablemente beneficiados también por la existencia de los corpus. La lingüística de corpus (LC) ha sido, en mi opinión, el tercer gran elemento renovador de los estudios lingüísticos. Surge como consecuencia de las posibilidades brindadas por el uso de computadoras en lingüística y pasa, en muy pocos años, de ser un recurso que permite realizar con ventaja tareas como la identificación y reunión de datos a convertirse en el motor de un cambio metodológico general cuyas consecuencias podemos observar ahora en los más diversos ámbitos de nuestra disciplina.

La lingüística hispánica llegó con cierto retraso a la LC. A comienzos de los años noventa, época en la que se diseñó y completó el British National Corpus, constituido por cien millones de formas y modelo para todos los corpus de referencia posteriores, los corpus de español tenían un tamaño mucho menor y se enmarcaban habitualmente en proyectos de investigación de ámbito europeo, como CRATER y PAROLE, o eran construidos como recurso auxiliar de proyectos lexicográficos, como el corpus CUMBRE, el Vox-Biblograf o el Corpus del Español Mexicano Contemporáneo, pero es de destacar que algunos otros presentan un carácter innovador, como sucede con los construidos en torno a las celebraciones del V Centenario. En 1995, la Real Academia Española tomó la decisión de abandonar su sistema tradicional de recogida de datos y comenzar la construcción del CREA primero y el CORDE

pocos meses después. Las primeras versiones de ambos corpus fueron publicadas en 1998 y, a partir de ese momento, la LC experimentó un crecimiento muy notable en el ámbito hispánico hasta llegar a la situación actual. Podemos hoy reunir datos procedentes de corpus constituidos por miles de millones de formas como el Es-Ten-Ten o el Corpus del Español (web/dialectos), trabajar con corpus de referencia como el CREA o el CORPES, usar corpus diacrónicos como CORDE o CORDIAM, con corpus especializados del estilo de los proyectos Biblia Medieval o CHARTA, con corpus orales como PRESEEA, ESLORA o COSER, corpus de aprendices de español L2 como CAES y un largo etcétera. En definitiva, la lingüística hispánica presenta en este punto un panorama todavía bastante alejado del que tiene el inglés, pero semejante al que se puede observar en muchas otras lenguas.

La influencia de la LC en la lingüística hispánica ha sido intensa, comparable a la que ha tenido en otras lenguas, quizá con un factor de repercusión especial en una tradición que con cierta frecuencia trabajaba con pocos datos, procedentes casi siempre de los mismos textos. La investigación sobre el español, en todas sus variedades y perspectivas, puede practicarse hoy con una solidez y un bagaje empírico que resultaban inimaginables hace tan solo treinta años. Sin embargo, la importancia de este proceso no ha tenido efectos visibles en la configuración general de la LC ni en la presentación de sus características generales o su historia. Hay que señalar que, en este punto, no se trata únicamente de la marginación del español. En realidad, son todas las lenguas distintas del inglés y todas las tradiciones investigadoras que no son la anglosajona las que están ocultas, incluso para quienes nos movemos habitualmente en otros contextos. No es algo específico de la LC; por citar solo casos muy claros, las historias de la lexicografía no reflejan la importancia (ni la existencia) del *Diccionario de autoridades*, Andrés Bello no aparece en las referidas a la gramática, las obras de Keniston o Fernández Ramírez no son conocidas ni mencionadas fuera de nuestro ámbito específico . . . En la misma línea, fuera de la lingüística hispánica, son muy escasas las referencias a los corpus de español y las que aparecen se refieren casi exclusivamente a los construidos en el mundo anglosajón.

Es, sin duda, la proyección en la lingüística del fenómeno más general de la escasa atención que se presta en el mundo científico a la investigación producida en español. En el caso de la LC, a estos factores generales se suma otro de carácter específico: no existen introducciones a esta corriente escritas en español que utilicen corpus de español y muestren cómo se pueden tratar e intentar resolver con datos de corpus problemas de lingüística española. Ese es, precisamente, el vacío que me he propuesto llenar con este libro. Lo he concebido como una introducción general a la LC planteada desde la tradición hispánica. Por tanto, los ejemplos, las ilustraciones, los problemas y los corpus manejados se refieren al español, aunque, como es lógico, sin ocultar los vínculos pertinentes con otras lenguas.

El libro tiene una marcada orientación didáctica. Está dirigido fundamentalmente a estudiantes de los últimos cursos de grados vinculados a lingüística española, estudiantes de máster y doctorandos que desean adquirir formación en este terreno o necesitan profundizar en él. Esta orientación, producto de una larga experiencia en la impartición de cursos de maestría y especialización sobre LC, explica la organización general del libro. Su punto de partida (capítulo 1) consiste en una descripción rápida y superficial de qué es un corpus textual y cuáles pueden ser las formas y ámbitos en los que puede ser utilizado para la investigación. Los demás capítulos van desarrollando, con la extensión y profundidad adecuadas a un texto introductorio, los temas esbozados en el primer capítulo. Esta estructura hace inevitables y también aconsejables algunas repeticiones: los mismos aspectos son tratados en cada ocasión a un nivel diferente.

La obra es una introducción práctica a la LC del español. Este carácter implica que debe ocuparse de todos los aspectos generales de esta orientación, pero también —quizá incluso, sobre todo— de la forma concreta en que puede acometerse la recogida de datos sobre los fenómenos acerca de los que se pretende realizar la investigación. En otras palabras, se atiende tanto al planteamiento de los problemas sobre los que se puede trabajar como a la forma en que hay que manejar las aplicaciones de consulta de diferentes corpus de español. En este sentido, los problemas analizados tienen una justificación intrínseca, pero su aparición en un punto determinado está casi siempre determinada por la aplicación de un procedimiento concreto que permite reunir los datos pertinentes.

Cada capítulo va precedido por un resumen y finaliza con un apartado de lecturas complementarias recomendadas y otro en el que se plantean diferentes cuestiones, problemas y temas de investigación. Dado el carácter introductorio de la obra, se ha pretendido que las lecturas complementarias sean adecuadas al nivel de conocimientos que se supone en los lectores, aunque no siempre ha sido posible conseguirlo. Por la misma razón, en el último capítulo se proponen tareas relativamente sencillas —o que, al menos, pueden ser tratadas con facilidad— incluidas siempre no por su relevancia teórica, sino para provocar la aplicación de las técnicas de recuperación y análisis descritas en los apartados precedentes. Dadas las características especiales del capítulo 7, me ha parecido conveniente en este caso incluir bloques de tareas prácticas que están distribuidos a lo largo de todo el texto, con la intención de ir comprobando la comprensión progresiva del funcionamiento de las utilidades analizadas.

El capítulo 1 pretende, como se ha dicho ya, proporcionar una introducción general a los corpus, la lingüística de corpus y los diversos ámbitos en los que puede ser aplicada. En el segundo se tratan algunas cuestiones metodológicas generales que considero de gran interés para la investigación en lingüística. El tercero retoma las cuestiones generales sobre corpus y lingüística de corpus descritos superficialmente en el capítulo 1, pero ahora a un nivel considerablemente más alto, aunque sin alejarse del carácter básico de toda la obra. Consiste en la descripción de todas las tareas que hay que realizar para construir un corpus y los aspectos más importantes del trabajo con estos recursos. Algunas de las cuestiones tratadas están más relacionadas con el diseño y construcción de los corpus que con su explotación, pero es imprescindible, en mi opinión, que exista una comprensión adecuada de todo lo que está implicado en un corpus para lograr una explotación adecuada de los datos que pueden obtenerse mediante su consulta.

Los capítulos 4 y 5 constituyen la aplicación de las técnicas habituales en la LC a fenómenos léxicos y gramaticales, respectivamente. Como he indicado ya, las cuestiones tratadas han sido seleccionadas no por su interés propio, sino sobre todo como ejemplos reales con los que se puede aprender a recuperar y tratar datos de los corpus textuales. Cuando es necesario se hace una breve descripción del problema que se va a tratar, para que todos los lectores de la obra tengan el conocimiento necesario y puedan entender adecuadamente la naturaleza de lo que se trata en cada caso. He pretendido poner ejemplos correspondientes a diferentes enfoques: sincrónico, diatópico, diacrónico, diastrático y atender también a otras posibles utilizaciones de los corpus, como el estudio de la lengua hablada, del español de los aprendices de esta lengua como L2 o el español rural.

En el capítulo 6 se desarrollan algunas cuestiones generales a las que se alude en los capítulos anteriores, pero sin la extensión que merecen para poder ser entendidas en todas sus dimensiones. Son, en cierto modo, grandes temas de la LC a los que aquí se dedica atención, aunque se tratan al nivel básico que corresponde a una obra de este tipo. Finalmente, en el

capítulo 7 se exponen algunas posibilidades avanzadas de unas cuantas aplicaciones de consulta y también una serie de herramientas informáticas de utilidad para quienes necesiten obtener la información contenida en textos que no están integrados en corpus y que, en consecuencia, no disponen de las aplicaciones de consulta a las que estamos acostumbrados. Estas últimas son herramientas propias de algunos sistemas operativos que pueden facilitar considerablemente el trabajo con los textos, las listas de ejemplos, listas de frecuencias y, en general, tanto con los textos como con los resultados de su análisis. Su desarrollo se hace a un nivel elemental y sin requerir conocimientos especiales de programación, pero puede ser manejado también como una introducción muy elemental a la llamada ciencia de los datos (*data science*) aplicada a la lingüística. La obra se cierra con un breve glosario que puede servir para la consulta rápida del significado de algunos conceptos fundamentales en LC, un índice temático y las referencias bibliográficas de los recursos y obras citadas en el texto.

Los aspectos cuantitativos son un elemento fundamental en la LC y, en ese sentido, son frecuentes las alusiones a que los grados en lingüística deberían incluir cursos de estadística, lo mismo que se ha hecho desde hace ya tiempo, en algunas otras especialidades adscribibles a las llamadas Humanidades. Aunque estoy de acuerdo con esa idea, he decidido no pasar en el libro de algunos conceptos muy básicos. El déficit de formación matemática y estadística en la generalidad de los estudiantes de lingüística no se puede resolver con un capítulo en una obra introductoria, sino que requiere un planteamiento distinto, difícilmente compatible con la orientación y tamaño de este libro.

En la parte más práctica de la obra, he tratado de poner ejemplos de trabajo con una amplia gama de corpus, pero no he pretendido incluirlos todos. He buscado siempre los de carácter más general, los más fácilmente accesibles y los que resultan más adecuados para la técnica que se pretende ilustrar en cada caso o el problema que se desea resolver. De forma inevitable, he utilizado sobre todo los que conozco mejor, aunque primando siempre la adecuación a los propósitos generales del libro.

El capítulo de agradecimientos es casi tan largo como el tiempo que he necesitado para terminar este libro. El primer lugar lo ocupan, sin duda, los numerosos estudiantes con los que he tenido oportunidad de trabajar sobre temas relacionados con la LC en muy diversas universidades. Como me ha ocurrido en muchas otras ocasiones, ha sido el esfuerzo necesario para explicar a otras personas los muy diversos temas tratados aquí lo que me ha permitido comprenderlos con la profundidad necesaria. El nutrido grupo de colegas que me ha acompañado en la USC a lo largo de mi carrera profesional ha sido un factor decisivo en mi interés por la LC, que arrancó cuando, hace ya muchos años, decidimos construir la Base de Datos Sintácticos del Español (BDS), que sigue siendo tan útil ahora como lo fue en el momento en que la hicimos pública. El agradecimiento que siento hacia todos ellos tiene una intensidad muy especial en el caso de Victoria Vázquez Rozas, de quien he aprendido continuamente desde sus tiempos de estudiante y con quien tengo la fortuna de seguir colaborando en ESLORA y otros proyectos de investigación. Al otro lado de una frontera borrosa debo citar a las personas que me han acompañado en el largo proceso de diseño y construcción del Corpus do Galego Actual (CORGA): Marisol López, Eva Domínguez y Mario Barcala, y también a Ignacio Palacios, con quien he podido poner a punto el Corpus de Aprendices de Español como L2 (CAES). Una buena parte de lo que hay en este libro es consecuencia directa de la decisión que tomó la Real Academia Española en 1995, cuando decidió acometer la confección del Corpus de Referencia del Español Actual (CREA). Sus responsables en aquel momento (Fernando Lázaro Carreter, Ángel Martín Municio y Víctor García de la Concha) creyeron en las posibilidades y la importancia de un proyecto que se situaba muy

lejos de la actividad realizada por la RAE hasta ese momento, lo hicieron posible y marcaron el rumbo que han seguido quienes los han sucedido. En el CREA y todos los corpus académicos que han venido después han colaborado cientos de personas, tanto en la RAE como en equipos situados en universidades de todo el mundo y algunas otras academias de Asociación de Academias de la Lengua Española. Mercedes Sánchez ha sido una ayuda constante y leal en el trabajo que hemos realizado durante estos veinticinco años. Finalmente, debo expresar mi reconocimiento a la editorial Routledge, que ha acogido la publicación de esta obra con un interés que todavía no ha dejado de sorprenderme. Los revisores anónimos del original, que me han transmitido valiosísimas sugerencias, Samantha Vale Noya y Rosie McEwan, editoras de la obra, y Carol Klee, responsable de la colección, han hecho sencillos, motivadores y agradables los tramos finales de un proceso que nunca se habría materializado sin el entusiasmo, el celo profesional y la generosidad de Javier Muñoz-Basols.

Framán, agosto de 2020.