Gerardo Sierra Martínez. *Introducción a los Corpus Lingüísticos.* Universidad Nacional Autónoma de México-Instituto de Ingeniería, México, 2017; 212 pp.

Rebeca Barriga Villanueva
El Colegio de México
rbarriga@colmex.mx

Entre la ingeniería y el saber lingüístico

Indiscutiblemente la lingüística de corpus ha representado un paso fundamental en la historia de la lingüística moderna. El tener acceso fácil y rápido a una cantidad exhaustiva de datos de la lengua en uso –ya oral, ya escrita– ha permitido una nueva forma de reflexión sobre sus estructuras y procesos, de ahí que surjan teorías innovadoras y metodologías de vanguardia que busquen dar cuenta del dato lingüístico con confiabilidad y versatilidad, además de explicar los complejos comportamientos de sistemas dinámicos en continua variación y cambio. Y es que unir los avances teóricos y metodológicos de las ciencias del lenguaje con los hallazgos de la ingeniería computacional ha potenciado la investigación empírica y la posibilidad de analizar los datos desde diferentes ópticas y criterios lingüísticos. Se trata de un verdadero juego de métodos computacionales que atrapa en el rigor matemático y estadístico fórmulas del comportamiento de pequeños elementos lingüísticos o complicadas combinatorias que permiten explorar en las entretelas de todos los niveles de la lengua en uso, en el tiempo y en el espacio históricos en que son producidas.

Gerardo Sierra Martínez incursiona en este complejo universo de los materiales de registro, muestras, procesos automatizados y de hechos observables con alcances múltiples en *Introducción a los Corpus Lingüísticos*, un libro valioso y útil desde muchos puntos de vista y para muchos tipos de lectores, legos o especialistas. Descansa este libro en una estructura de cinco apartados que siguen un ordenamiento lógico que va de la "Introducción al corpus" a sus "Aplicaciones", pasando por la "Compilación", la "Anotación" y las "Herramientas de análisis". Cada uno de estos apartados, formados a su vez de incisos o artículos –un total de catorce–, van construyendo una fuerte y variada riqueza descriptiva y explicativa para usuarios con diversos focos de atención, siempre acompañados de un inciso final dedicado a las referencias bibliográficas de los temas ahí tratados. Huelga hablar del valor adicional de una bibliografía especializada que cobija todas las partes. Por todo ello, éste es un libro difícil de reseñar, pues cada parte es importante y se vincula con otra igualmente relevante. Por el momento, bástenos con transitar por el índice para darnos cuenta de ello. En una suerte de estrategia lúdica que nos permite itinerar libremente, escojo de aquí y de allá algunos temas que considero sobresalientes porque invitan a la reflexión, ilustran el porqué de la importancia

ISSN: 2448-8194

de los corpus y son una invitación abierta e insistente para que los lectores construyan sus propios significados a partir de una exploración minuciosa de las partes, autónomas entre sí, pero estrechamente entretejidas.

Así, en el primer apartado, "Introducción al corpus", Sierra abre, centrándose en la naturaleza del corpus, parte de su definición, de aparente sencillez, «conjunto de textos de materiales escritos y o hablados debidamente recopilados para hacer análisis lingüísticos […] y en soporte informático» (4-6), pero, como él mismo señala, con una dificultad intrínseca que emana de la necesidad de precisar cada una de las partes que lo constituyen y de situarlas dentro del contexto del avance tecnológico del momento, de suyo vertiginoso, y de las exigencias de precisión y confiabilidad que se requieren. Se detiene también en las características que hacen a un corpus: contención de datos, representatividad, variedad, equilibrio, selectividad y, finalmente, el discutible tamaño finito (un millón de palabras como unidad de medida). Con el curso de los avances, hay también la necesidad de penetrar en los rasgos distintivos de los corpus que han devenido en informatizados, cobijados en soportes electrónicos que recuperan la información, merced a una clasificación precisa, emanada de un conocimiento profundo de la lingüística y la ingeniería computacional. Sierra aborda al final de este apartado el espinoso tema de los derechos de autor y la propiedad intelectual, que son parte de los múltiples requisitos de una sociedad global que lucha contra el plagio o los usos indebidos. Los otros tres artículos que forman este apartado recorren temas de sumo interés; el mero título es elocuente: "Descripción de corpus existentes". Sierra da cuenta de una considerable cantidad de corpus en español e inglés que son ampliamente conocidos y manejados, menciono algunos de ellos: *Corpus del español mexicano contemporáneo* (CEMC). *Corpus diacrónico del español* (CORDE) y CHILDES. Añadiría otros que, por su valor y su especificidad, podrían sumarse al recuento de nuestro autor: el *Corpus sociolingüístico de la Ciudad de México*, coordinado por Pedro Martín Butragueño y Yolanda Lastra en El Colegio de México, y los *Estudios de etapas tempranas de adquisición* (ETAL), de Cecilia Rojas, desarrollado en la Universidad Nacional Autónoma de México (UNAM). En "Clasificación de corpus", Sierra hace una tipología de corpus: según el origen de los datos, la lengua y la documentación, por mencionar algunos, a los que se suma el Internet como corpus.

Con respecto del segundo apartado, "Compilación de corpus textuales", aparecen los artículos que se refieren a la importancia de la identificación del objetivo y de la selección de textos, ya que de su estrecha relación depende el éxito de la empresa. Habrá que poner límites geográficos y temporales, así como determinar y decidir el alcance y la pertinencia de la información almacenada, como títulos, subtítulos, portadas y otros datos que pueden añadir valor al cuerpo esencial. No podría dejar de mencionar la importancia de la estandarización de los formatos, pues pone en juego algunas restricciones emanadas de la procedencia de los datos de naturaleza oral o escrita, bien que provengan de Internet o de registros en audio, pues requieren de una sistematización especial. Por último, Sierra trata el tema de la transcripción, crucial, pues le subyacen decisiones involucradas con teorías y métodos comprometidos con un análisis riguroso y confiable.

Siguiendo la misma línea de las decisiones y de su estrecho compromiso con el análisis, en el apartado tercero, titulado "Anotación de corpus", Sierra aborda el etiquetado de los datos y las necesidades que de él emanan, una vez que ya están en formato electrónico, pero que aún no son procesados como corpus. Será imprescindible seleccionar y anotar los elementos a analizar: morfemas, tiempos verbales o tipos de oraciones, por ejemplo. Nos ofrece para lograrlo los principios sobre la anotación del corpus y los conceptos básicos del etiquetado: *inteligibilidad* –etiquetas bien diferenciadas y únicas–, *extracción* –suprimir la anotación de un corpus y convertirlo en texto simple–, *intercambio* –reutilización de textos ya etiquetados–, *documentación* –etiquetas con documentación disponible– y *estandarización* –los esquemas de anotación deben de estar basados en representaciones ampliamente definidas, aceptados por consenso y sin interpretación subjetiva–. Asimismo, presenta los conceptos básicos de etiquetado: *entidad de mensaje*, *elemento* (metadatos), *atributo de marcaje* (información adicional) y *referencias de identidad*.

De especial interés resulta el capítulo sobre los tipos de anotación –textual, fónica morfológica, sintáctica, semántica, discursiva y pragmática–, por su estrecha relación con los niveles de la lengua y sus subniveles. Por ejemplo, según la clasificación de Sierra, el nivel textual cuenta con los de estructura, tipología y ortografía, en tanto que el discursivo tendrá el de relaciones y el de anáfora y referencial, el cual, según su dicho, «es de las áreas en las que menos se ha incursionado por parte de quienes trabajan corpus lingüísticos informatizados» (131). Cabe preguntarse si la vía de explicación a esta escasez es la complejidad intrínseca del nivel de análisis mismo o se trata de la falta de herramientas que atrapen los vericuetos de la referencialidad. Concluye este apartado con la anotación de polaridad, la más novedosa entre la tipología de anotaciones, pues se hace cargo de la compleja detección objetiva de emociones (tristeza, miedo, ira) y opiniones (positiva, negativa, neutra).

Las "Herramientas y técnicas de análisis" son el foco de atención del cuarto apartado, dividido en dos: técnicas de análisis y herramientas de análisis textual. De las primeras, elijo para ilustrar las colocaciones por su intrincada naturaleza y los varios temas que pueden abarcar –expresiones terminológicas, expresiones discursivas, locuciones con sus diferentes tipos, expresiones idiomáticas y nombres de organismos–, muy presentes en las discusiones en el ámbito de la semántica léxica y de la sintaxis. En cuanto a las herramientas que están destinadas a analizar el comportamiento de las palabras de un corpus, menciono *WordSmith Tools*, que genera listados de palabras por orden alfabético o por frecuencia de aparición, y *Goldvarb*, programa de base estadística especializado en estudios de variacionismo que se caracteriza por su fuerza predictiva e inferencial.

Las "Aplicaciones", tema del quinto y último apartado son presentadas en la clásica y forzada división entre lingüística y lingüística aplicada. En cuanto a la primera, tratada en todos los niveles, del fonético-fonológico al semántico-pragmático, Sierra analiza las aplicaciones ya consolidadas o las que se encuentran aún en proceso. Todas éstas han sido realizadas por el Grupo de Ingeniería Lingüística (GIL) de la Universidad Nacional Autónoma de México, uno de cuyos frutos es el *Corpus Histórico del Español de México* (CHEM) y varias tesis de licenciatura, maestría y doctorado, que analizan varios procesos

y mecanismos del español diacrónico o sincrónico. Las aplicaciones en lingüística aplicada abarcan la lexicografía, la terminología, lingüística forense y el novedoso análisis estilométrico para la detección del plagio –detección de similitud textual, mediante criterios de discurso y semántica–. En la última sección de este apartado, nuestro autor pone sobre la mesa la necesidad imperiosa de un conocimiento del lenguaje natural que permita seguir construyendo conocimiento lingüístico y de abordarlo en nuevas áreas como la extracción de información, de contextos definitorios y de relaciones léxico-semánticas, a partir de textos de especialidad o de traducción automática –necesaria en un mundo multilingüe como el que habitamos.

Tras este apretado y vertiginoso trayecto por el índice, podríamos barajar rápidamente las bondades y la riqueza de la *Introducción al corpus lingüístico* que la hacen invaluable como instrumento guía, un manual de conocimiento acerca de las infinitas posibilidades del lenguaje natural, plasmadas en una formato atractivo y ordenado. Además, se agradece el orden de la presentación, la claridad y explicitud de los objetivos; el libro es fiel a la vocación natural por la docencia de su autor, no en vano la simiente fueron sus cursos impartidos tanto en licenciatura como en posgrado. Se reconoce también la forma accesible y completa de mostrar temas de gran envergadura teórica y metodológica de la lingüística y de la ingeniería computacional. La información reunida le confiere al texto una clara misión: construir conocimiento y abrir vetas de investigación. En este sentido, se hace fehaciente la solidez de la investigación mexicana que está a la vanguardia del avance tecnológico aplicado al español mexicano y a su funcionamiento. Sin embargo, de todos los valores, el más sobresaliente de este libro es que, a cada paso, motiva una reflexión en torno a la enorme potencialidad del lenguaje humano.