Gerardo E. Sierra Martínez. *Introducción a los corpus lingüísticos*. México: Universidad Nacional Autónoma de México, Instituto de Ingeniería, 2017. 212 págs.

Julio Serrano
Universidad Autónoma Metropolitana-Iztapalapa,
Departamento de Filosofía

La comunidad académica mexicana y latinoamericana da la bienvenida a este volumen que sin duda resulta muy atractivo para estudiantes de licenciatura y posgrado, así como para todo investigador que quiera acercarse a la lingüística de corpus, a pesar de no contar con antecedentes en la materia. El volumen resulta, por lo tanto, muy didáctico para el lector porque presenta las facetas clave del trabajo con corpus lingüísticos: desde su clasificación y diseño, pasando por las decisiones analíticas en términos de herramientas y resaltado, hasta las múltiples aplicaciones del conocimiento generado por este amplio y cada vez más complejo abanico conceptual y técnico. En las líneas siguientes, además de hacer una breve descripción de la estructura y contenidos del libro, procuraré destacar los aspectos que, desde mi perspectiva, encontré más enriquecedores.

Introducción a los corpus lingüísticos es el primer manual que, sobre la materia, nos ofrece Gerardo Sierra, investigador del Instituto de Ingeniería de la UNAM, quien ha sido una de las figuras más importantes en el desarrollo y difusión de la lingüística computacional en México. El volumen se organiza en cinco partes principales, que compilan un total de 14 capítulos. Las partes son: I) Introducción a corpus, II) Compilación de corpus, III) Anotación de corpus, IV) Herramientas y técnicas de análisis y V) Aplicaciones.

Lo primero que puede destacarse del libro es que el autor hace hincapié en que la lingüística de corpus nos dota de herramientas conceptuales, matemáticas y computacionales para estudiar el lenguaje en *todos* sus niveles estructurales. Los ejemplos de investigaciones pasadas y en curso que nos presenta Sierra incluyen trabajos

en fonética, fonología, morfología, sintaxis, semántica, pragmática, discurso y —no podían faltar— sobre aspectos sociolingüísticos. Los antecedentes de investigación en lingüística de corpus y los proyectos más recientes se compilan en el capítulo 2, "Descripción de corpus existentes", donde se mencionan la historia y características generales de importantes —y en algunos casos, legendarios— corpus mexicanos como el *Corpus del Español Mexicano Contemporáneo* (CEMC) de El Colegio de México —pionero en la lexicografía basada en corpus computarizado a nivel mundial—, el reciente *Corpus Electrónico del Español Colonial Mexicano* (Coreecom) de la UNAM, coordinado por Beatriz Arias en el Instituto de Investigaciones Filológicas, así como el *Corpus Diacrónico y Diatópico del Español de América* (Cordiam) de la Academia Mexicana de la Lengua. Gerardo Sierra también nos reseña brevemente los corpus más importantes en lingüística en el mundo, como el *Corpus del Español del Siglo xxi* (Corpes XXI) y el *Corpus Diacrónico del Español* (Corde) de la Real Academia Española, el *Child Language Data Exchange System* (Childes) y el *Corpus del Español* de Mark Davies. Por supuesto, el Grupo de Ingeniería Lingüística (GIL), del cual Sierra es creador, ha diseñado y construido varios corpus, además de haber apoyado otros proyectos en la materia. Entre estos, puedo destacar el *Corpus de las Sexualidades en México* (CSMX), el *Corpus Histórico del Español de México* (CHEM), el *Rethorical Structure Theory* (RST) *Spanish Treebank* o el corpus paralelo náhuatl-español *Axolotl*; finalmente, el *Corpus de Ad-hocracia*, sobre el español parlamentario mexicano, coordinado por Margarita Palacios en la Facultad de Filosofía y Letras de la UNAM, también se destaca en esta revisión.

En el capítulo 3, "Clasificación de corpus", se enumeran y describen los distintos parámetros que deben tomarse en cuenta para dicha clasificación, que resultan muy útiles porque permiten empatar los objetivos de la investigación con las características finales del corpus que se habrá de utilizar. En este sentido, el autor nos menciona que deben considerarse criterios como el origen de los datos (orales o textuales) y el tipo de codificación y anotación.

También debe tomarse en cuenta el propósito del estudio (específico o multipropósito), la distribución del tipo de texto (desequilibrado, equilibrado o piramidal) y su representatividad (oportunista vs. representativo), por mencionar algunos parámetros. Por otra parte, el capítulo 4, "Internet como corpus", está dedicado específicamente al aprovechamiento de este medio electrónico como corpus y ofrece algunos recursos en línea para generar nuestros propios cuerpos de datos (buscadores como Google y metabuscadores como WebCrawler son algunos de los varios ejemplos citados).

En la parte II, titulada "Compilación de corpus", se discuten varios puntos cruciales para dicha tarea, como la selección y obtención de textos, su digitalización y la estandarización de los mismos; el capítulo 6, "Compilación de corpus orales", en especial, lo dedica a estos corpus, entre los que destacan los distintos alfabetos computacionales estandarizados para la transcripción fonética de corpus.

Un logro de Gerardo Sierra es que explica, de manera ágil y sencilla, aspectos ciertamente complejos de la lingüística computacional y de corpus. Esto es muy evidente sobre todo en la parte III, sobre "Anotación de corpus". Por supuesto, es una introducción al tema, pero cualquier persona interesada en estos procesos de anotación y, sobre todo, de *etiquetado* de corpus, tiene bases firmes para acercarse a esta labor, como puede leerse en el capítulo 7, "Bases para la anotación de corpus". Al respecto, el lenguaje XML recibe una atención especial en el capítulo 8, "XML".

La parte IV, sobre "Herramientas y técnicas de análisis", es muy atractiva porque muestra algunos conceptos clave que, como lingüistas, nos hemos encontrado en nuestras tareas investigativas, como el clásico —y muchas veces incomprendido— concepto de *concordancia*, y que encuentran una definición y explicación muy clara por parte del autor: "lista de palabras localizadas en contexto". Encuentro especialmente útil el capítulo 10, "Técnicas de análisis", sobre concordancias, colocaciones y conteo de palabras, donde el autor describe e ilustra términos como *información*

mutua y *costo de colocaciones*, fundamentales para el análisis lexicográfico.

Es importante señalar que el autor aprovecha la vasta experiencia del GIL para explicar diferentes procedimientos habituales en la lingüística de corpus. Esta experiencia se destaca principalmente en la parte V, "Aplicaciones", tanto en lingüística (capítulo 12), como en lingüística aplicada (capítulo 13) y aplicaciones en tecnologías del lenguaje (capítulo 14). Las investigaciones de Valeria Benítez sobre sintaxis del español contemporáneo, las de Teresita Reyes sobre la fonología del español colonial mexicano y la tesis de Carlos Méndez sobre etiquetado automático de corpus (todos estos proyectos, acogidos por el GIL) permiten ilustrar los procedimientos, herramientas analíticas y resultados de una diversidad de problemas lingüísticos que solo pueden tener solución a través de recursos computacionales. De esta manera, a partir de la lectura de esta sección, el estudiante de licenciatura o posgrado cuenta con una útil guía sobre el proceso de investigación con corpus lingüísticos. La terminología, la lingüística forense, la enseñanza de segundas lenguas, el análisis del discurso y las tecnologías de reconocimiento de voz son solo algunas de las áreas de aplicación donde la lingüística de corpus tiene un papel fundamental.

Martin Haspelmath, en el Congreso de la Societas Linguistica Europaea de 2014,[1] señaló, entre otros puntos, el importante giro que está tomando la lingüística hacia acercamientos más empíricos, muy cuantitativos y menos conceptualistas y que el futuro se dirige hacia una visión no apriorística en el estudio del lenguaje. Estas tendencias científicas en nuestro campo, el cual debe ahora dialogar con las neurociencias, las disciplinas cognitivas y la investigación en evolución, implican, por supuesto, el uso de vastos corpus y el dominio de herramientas computacionales refinadas que nos permitan proponer generalizaciones. En este sentido, una obra como la de Sierra tiene la virtud de introducir de lleno al lec-

[1] Véase https://dlc.hypotheses.org/754

tor en este campo creciente y con el que, de manera inevitable, habremos de dialogar para asegurar el futuro de muchísimas ramas de nuestra disciplina.

Con un estilo de escritura sumamente ordenado y ágil, *Introducción a los corpus lingüísticos* de Gerardo Sierra es sin duda un referente obligado para cualquier investigador o estudiante interesado en dotar a sus investigaciones de mayor rigor empírico. Como ya lo señalamos, las tendencias actuales en lingüística recomiendan cada vez más acudir a la producción de los hablantes antes de proponer generalizaciones basadas solo en la intuición. Sabemos bien que el trabajo cuantitativo puro no existe: antes de cuantificar, debemos cualificar los elementos por analizar, y por esto mismo es que debemos acercarnos a esta manera de trabajar; libros como el que aquí se reseña nos señalan la gran diversidad de recursos con los que contamos y que debemos aprovechar en futuras investigaciones.