Gerardo E. Sierra Martínez, *Introducción a los Corpus Lingüísticos*, México, Universidad Nacional Autónoma de México, 2017, 214 pp. ISBN: 978-607-029-898-1.

Beatriz Arias Álvarez
Centro de Lingüística Hispánica "Juan M. Lope Blanch"
Instituto de Investigaciones Filológicas, UNAM

El libro *Introducción a los Corpus Lingüísticos* de Gerardo Sierra tiene un doble propósito: por un lado, ofrecer un panorama general, pero no por ello menos profundo, de las diferentes características y elementos que conforman los diversos corpus informatizados; por el otro, brindar una bibliografía pertinente sobre los trabajos que otros investigadores han realizado sobre dicho tema.

Atendiendo al primer propósito, el libro se encuentra dividido en cinco partes: "Introducción a corpus", "Compilación de corpus", "Anotación de corpus", "Herramientas y técnicas de análisis" y, por último, "Aplicaciones". Estas cinco partes se subdividen además en diferentes apartados; tomando en cuenta el segundo propósito, cada uno de los capítulos contiene lecturas sugeridas, además de ofrecer, al final de la obra, una bibliografía general.

En cuanto a las partes y capítulos en los que se divide el libro podemos señalar lo siguiente: en la primera de las cinco partes, "Introducción a corpus", se da una definición del término. Para el autor un corpus lingüístico "debe ser un modelo de realidad lingüística que se quiere observar" (p. 3), "un conjunto de textos de materiales escritos y/o hablados, debidamente recopilados para realizar ciertos análisis lingüísticos" (p. 4). Ahora bien, para poder construir un corpus, y para que este sea un

modelo de realidad lingüística deben seguirse determinados lineamientos: hay que llevar a cabo una recopilación adecuada de textos, ya orales, ya escritos; señalar además los objetivos que se pretenden alcanzar, de ahí se derivan las características que debe tener el corpus en cuestión; hay que prestar también atención al tipo de datos según el estudio a realizar, además de considerar la representatividad, la variedad, el equilibrio, la selectividad y el tamaño. Ahora bien, estos parámetros que deben ser considerados para cualquier tipo de corpus, deben ser tomados también en cuenta para los digitalizados. Si bien antes el investigador debía trabajar largas horas para recabar información y plasmarla en ficheros y archivos, y, posteriormente, ya obtenidos los datos, debía organizarlos y reorganizarlos para encontrar la información que requería; con los corpus digitalizados se ahorra tiempo y se puede cruzar la información en instantes y de manera fiable. En un corpus digital, el dato o datos se pueden manipular de manera más sencilla y rápida, su precisión es exacta; además, la transferencia resulta más cómoda. Y a todo esto hay que sumar que los corpus de este tipo se actualizan exitosamente, ya que pueden incorporarse nuevos textos y datos de manera confiable. Los diferentes corpus existentes, según el origen, la autoría, la lengua, el propósito, la cantidad o la distribución, pueden mejorar la calidad de la investigación. Sin embargo, hay que considerar dos aspectos importantes: no toda recopilación de textos en internet es un corpus, en palabras del autor conviene recordar que "en todo corpus se debe tener una clasificación de las fichas o de los textos, de manera que sea recuperable únicamente la información que interesa" (p. 14); además existen desventajas, en un primer momento la incorporación de material implica soportes electrónicos, especialistas, inversión económica, en ocasiones la obtención de derechos de autor y, por supuesto, la posibilidad de que el corpus presente fallas técnicas en el momento de uso.

En esta primera parte, se suma el recorrido que el autor hace por diferentes corpus electrónicos sobre el español, realizados dentro de la Universidad Nacional Autónoma de México, en instituciones mexicanas, además de los elaborados en otros

países, tanto del español como del inglés. Entre los primeros podemos hacer mención del Corpus Electrónico del Español Colonial Mexicano (COREECOM) del Instituto de Investigaciones Filológicas, además de los numerosos corpus que se realizan en el Instituto de Ingeniería o apoyados por este: Corpus Lingüístico de Ingeniería (CLI), Corpus de la Sexualidad en México (CSMX), Corpus Histórico del Español en México (CHEM), Axolotl, Corpus Electrónico para la Enseñanza de la Lengua Escrita (CEELE) o el Corpus Científico del Español de México (COCIEM); además se cuenta con el Corpus Diacrónico y Diatópico del Español de América (CORDIAM) realizado por la Academia Mexicana de la Lengua, el Corpus del Español Mexicano Contemporáneo (CEMC) trabajado en El Colegio de México, a los que deben sumarse el Corpus Diacrónico del Español (CORDE), El Corpus de Referencia del Español Actual (CREA) elaborados por la RAE y el Corpus del Español de Mark Davis, entre muchos otros. A lo anterior, el autor añade la explicación sobre los diferentes formatos electrónicos, los buscadores, los programas de análisis lingüístico, las ligas y los documentos disponibles en internet.

En la segunda parte, "Compilación de corpus", se ofrecen las particularidades de los corpus textuales y de los orales. Tanto en unos como en otros, es de suma importancia identificar el objetivo (parámetros, propósito, análisis lingüístico, enseñanza de idiomas, lexicografía) y los límites (temporales, diatópicos, temáticos, etc.); en los primeros, además, hay que seleccionar los textos escritos, digitalizarlos de acuerdo con el propósito y estandarizar los formatos. En los segundos, es necesario buscar a los hablantes, poner atención en las características de la grabación, utilizar ciertos programas para la transcripción del corpus (por ejemplo, *Praat* o *Speech tool*) y realizar diferentes tipos de transcripción según el caso (ortográfica o fónica, precisando, además, en el último caso, el tipo de alfabeto fonético utilizado).

En cuanto a la tercera parte, "Anotación de corpus", el autor nos señala que un corpus electrónico no es únicamente la recolección y organización de textos presentados en un formato textual determinado: "un mismo texto puede servir a distin-

tos tipos de análisis, de forma que resulta necesario, primero, identificar los elementos del texto que son de interés y, segundo, marcar los segmentos con las anotaciones que sean pertinentes" (p. 93). De ahí que haya elementos importantes a considerar para la elaboración de un corpus electrónico: por ejemplo, los metadatos, en los que se pueden especificar aquellos parámetros que caracterizan a los textos. Al mismo tiempo, es relevante realizar un etiquetado, en el cual se puede hacer énfasis en los aspectos lingüísticos que se quieran estudiar: nivel gráfico, fonológico, morfológico, léxico, semántico, etc. Para el autor los tipos de anotación que se pueden hacer sobre un corpus "están determinados por los niveles de análisis de la lengua y deben hacerse de acuerdo con ciertos principios" (p. 117), a saber: la anotación textual tiene tres tipos: por estructura textual, por tipología y ortográfica. La anotación fónica puede realizarse a través de la fonética, la fonología y la prosodia (acento, pausa, melodía, velocidad, cualidad de la voz). La morfológica se realiza a través de la 'lematización'. Para la morfosintáctica se puede utilizar el etiquetado de las partes de la oración POST. En la anotación sintáctica puede emplearse el Parsing parcial o total. Para la anotación semántica se puede utilizar el etiquetado de las características semánticas de una palabra, su anotación ontológica y la de relaciones semánticas. Para Gerardo Sierra cualquier tipo de anotación que se realice debe seguir determinados lineamientos: inteligibilidad, extracción, intercambio, documentación y estandarización. Además, "es necesario considerar los elementos que se desean etiquetar, el nombre que se les va a poner a cada una de las etiquetas y las características que tendrán los distintos elementos" (p. 99).

También es importante destacar que hay diferentes lenguajes que se utilizan para el etiquetado, los más utilizados son el HTML y el XML. Mientras las etiquetas en HTML tienen como objetivo servir para la presentación de páginas web, señalar su estructura, su diseño y los hipertextos; en el segundo, XML, se deben identificar los elementos para darles formato, para poder procesarlos. Este último tipo de etiquetado es el más usado hoy en día, ya que ofrece distintas funcionalidades y ventajas,

por ejemplo, permite la organización de documentos, se puede emplear con otros lenguajes y el programador puede definir sus propias etiquetas según convengan a los objetivos; a lo que hay que sumar que con el tiempo se pueden incorporar nuevas etiquetas y que el sistema es fácilmente interpretable por las máquinas. El XML, por ejemplo, es el utilizado en el Corpus de Contextos Definitorios (CORCODE) del Instituto de Ingeniería.

Con respecto a la cuarta parte, "Herramientas y técnicas de análisis", el autor inicia con el conteo de palabras, cuyo punto de partida consiste en conocer la diferencia entre *token* y *type*. "El *token* o palabra es cada una de las formas que aparecen en el texto, sin importar cuántas veces ocurra cada una […]. Por su parte, el *type* se refiere a cada una de las formas o palabras diferentes que aparecen en un texto" (p. 137). Así, la riqueza léxica de un documento se encuentra en la relación entre los *token* y *type*. Tomando en cuenta lo anterior, Gerardo Sierra señala en este apartado las características de las listas de palabras (simples, canónicas, de lemas) y el orden que pueden seguir (alfabético, por frecuencias, por categoría gramatical, etc.). Hay que advertir que en la elaboración de listas de palabras el investigador se encuentra con varios problemas: el primero es la segmentación o identificación de los *token*; el segundo, la eliminación de signos gráficos, por ejemplo el apóstrofo para el inglés; el tercero, la eliminación de mayúsculas. Estos problemas no sólo los tiene el programador, también el usuario del corpus: por ejemplo, la presencia o no de un acento en una palabra puede distorsionar los resultados, lo mismo que el empleo de mayúsculas, no sólo en los nombres propios, también en los comunes a los que les precede un punto. En este apartado, además, se advierte la importancia de las 'concordancias', es decir, el contexto en el que se encuentra una palabra y que es fundamental para muchos estudios, no sólo semánticos, sino también morfosintácticos. En cuanto a las herramientas de análisis, el autor señala tres básicas: *Wordlis*t, *KeyWords* y *Concord*, además de varios instrumentos que completan dichas herramientas. A lo anterior se suman la descripción de varios software, entre los que destaca *Goldvarb*, "el cual fue diseñado para el estudio de la variación lingüística

mediante técnicas de análisis estadístico multivariable, ya que se ha mostrado que la variación lingüística no es producto del azar y está condicionado por múltiples factores" (pp. 152-153).

La última parte, "Aplicaciones", está dividida por el autor en tres subtemas: a) las aplicaciones en 'lingüística', en donde señala la utilidad de los corpus electrónicos y su empleo para los estudios fonológicos, morfológicos, sintácticos, semánticos, de análisis del discurso y pragmáticos. Los ejemplos son tomados de investigaciones que se han realizado en el Instituto de Ingeniería de la Universidad Nacional Autónoma de México y en otras instituciones mexicanas; b) el uso de las mismas en la 'lingüística aplicada', en donde se nombran los diccionarios que se han realizado, como ejemplo sobresale el de Luis Fernando Lara, llevado a cabo en El Colegio de México, además de las utilizadas en la obtención de términos, y dentro de la lingüística forense para la detección de plagios y de similitudes textuales; y c) se nos ofrecen las aplicaciones dentro de la 'tecnología del lenguaje', por ejemplo, en la extracción de contextos definitorios, en las traducciones automáticas o en la obtención de un léxico paralelo náhuatl-español.

Sin duda el libro de Gerardo Sierra no sólo es una introducción a la lingüística de corpus, en la cual se nos ofrecen las características, elementos, procesadores y etiquetados que deben considerarse en la elaboración de un buen corpus, es también una obra en la que se hace un recorrido por los principales corpus del español y se advierten los tipos de investigaciones que pueden obtenerse. Así, el libro *Introducción a los Corpus Lingüísticos* es imprescindible para todo aquel que realice o quiera realizar estudios sobre el tema desde diferentes áreas.