Explorando el uso del lenguaje
con corpus lingüísticos

Por Ana Abigahil Flores Hernández y Pauline Moore

Ilustración: Jisel Flores

La lingüística de corpus es un área de ciencia de frontera en la lingüística aplicada actual, donde se reúnen grandes cantidades de lenguaje hablado y escrito con la finalidad de detectar patrones de uso y tendencias de cambio en el uso de la lengua. La lingüística ha utilizado desde hace mucho tiempo grandes y pequeñas colecciones de textos para explorar el uso de la lengua literaria y cotidiana. Sin embargo, no es hasta la invención de equipos de grabación en audio y computadores de gran potencia que se han podido explotar estas tecnologías para ofrecer una mirada a una muestra significativa y representativa del lenguaje y llevar a cabo su análisis de una manera eficiente y sistemática. Investigar el uso del lenguaje de esta manera ofrece una fuente confiable en que basar nuestras observaciones, pues la intuición del investigador de la lengua es parcial y se limita a su experiencia personal, mientras que los datos que ofrece un corpus grande facilitan un panorama más extenso.

Actualmente, varias universidades e instituciones nacionales y extranjeras colaboran en la creación de corpus de lenguas como español, inglés, francés, italiano, chino, entre muchas otras. Hay varias formas de clasificar, pero para dar una idea de su utilidad los podemos categorizar inicialmente en contemporáneos e históricos. Los corpus contemporáneos o sincrónicos reúnen grandes cantidades de textos orales y escritos de una variedad de géneros (conversaciones, clases, ponencias, artículos científicos, novelas, artículos periodísticos, blogs, etc.) producidos por hablantes de todas las edades, nacionalidades y condiciones sociales con distintos niveles de formalidad. En español, por ejemplo, existe un gran corpus sincrónico elaborado por la Real Academia

Española, el CREA (Corpus de Referencia del Español Actual: http://corpus.rae.es/creanet.html) que incluye textos recopilados en 22 países de habla española, desde Argentina hasta Venezuela, y sobre temas tan diversos como la biología y la tauromaquia. El CREA incluye datos de los últimos 25 años de uso del español, por lo que está constantemente alimentado por sus curadores y cambia todos los años; por esta característica es difícil especificar en un momento dado cuantas palabras hay en el corpus, pero en 2008 contaba con aproximadamente 160 millones de formas.

Los corpus históricos o diacrónicos reúnen textos del pasado, por ejemplo, la Real Academia Española también ofrece un corpus diacrónico que incorpora textos históricos del español: el CORDE (Corpus Diacrónico del Español: http://corpus.rae.es/cordenet.html). En este se reúnen 125 millones de formas desde los primeros textos en español producidos en la Edad Media, a través de los siglos de oro, hasta textos relativamente contemporáneos de 1975. Este corpus nos permite rastrear los usos de la lengua a través del tiempo, así como observar cómo han evolucionado naturalmente de acuerdo con las necesidades e intereses de sus usuarios.

El CREA y el CORDE son corpus generales, pero también podemos reunir un corpus para satisfacer una pregunta de investigación, por ejemplo, cuáles expresiones lingüísticas son propias de los toluqueños. La intuición del investigador le puede sugerir algunas frases típicas, pero si contara con un corpus de producción lingüística de los toluqueños se podría responder con mayor seguridad. Los corpus específicos se pueden analizar de manera aislada o pueden compararse con los corpus generales para identificar patrones específicos de esta población.

LOS CORPUS CONTEMPORÁNEOS O SINCRÓNICOS REÚNEN GRANDES CANTIDADES DE TEXTOS ORALES Y ESCRITOS (CONVERSACIONES, CLASES, PONENCIAS, ARTÍCULOS CIENTÍFICOS, NOVELAS, ARTÍCULOS PERIODÍSTICOS, BLOGS, ETC.) PRODUCIDOS POR HABLANTES DE TODO EL MUNDO

En el proyecto de estancia posdoctoral financiado por el Conacyt "Corpus mexicano de aprendientes de una segunda lengua", que realiza la doctora Ana Abigahil Flores Hernández en la Maestría en Lingüística Aplicada de la Facultad de Lenguas, bajo la supervisión de la doctora Pauline Moore, se levanta un corpus longitudinal oral de aprendientes del inglés como lengua extranjera mediante entrevistas semiestructuradas. Este corpus servirá para identificar las fortalezas de nuestros estudiantes en el aprendizaje del inglés, lo que es un gran apoyo para el diseño de materiales didácticos más adecuados y será el primer corpus oral de aprendientes en México. El proceso de recopilación de la primera etapa del corpus requiere un año para aplicar las entrevistas, transcribirlas, cargarlas en el *software* de análisis y etiquetar información gramatical para facilitar su análisis posterior.

Ana Abigahil Flores Hernández es becaria Conacyt posdoctorante en la Facultad de Lenguas de la UAEM. Investiga aspectos de la adquisición de una segunda lengua a través de la evaluación y el análisis de corpus lingüístico.

Pauline Moore es doctora en Lingüística y profesora de tiempo completo de la Facultad de Lenguas de la UAEM desde 1993, donde imparte unidades de aprendizaje relacionadas a la lingüística aplicada.