Signos Lingüísticos, vol. IV, núm. 7, enero-junio, 2008, 95-100

CORPUS LINGÜÍSTICOS AUTOMATIZADOS, UN MÉTODO PARA OBSERVAR EL COMPORTAMIENTO DE LA LENGUA*

WILLELMIRA CASTILLEJOS LÓPEZ**
Universidad Autónoma Chapingo

Resumen: El campo de la investigación lingüística se ha abierto recientemente al uso de la tecnología informática. La lengua puede concebirse ahora como un sistema de signos *cuantificables* y útiles para la comunicación humana. El énfasis en lo cuantificable reside en el hecho de que la observación de su estructura a través de un corpus automatizado manipulable con herramientas para medir la frecuencia de las palabras, las compañías preferidas de unas palabras con otras, la sinonimia, las palabras claves, entre otros aspectos, permite concluir que la lengua es una entidad más sujeta a normas fijas que al libre arbitrio. La lengua se parece mucho a las matemáticas. La investigación en lingüística de corpus parece así demostrarlo.

PALABRAS CLAVE: CORPUS, FRECUENCIA, HERRAMIENTAS, REPRESENTATIVIDAD, TEXTOS

Abstract: *In recent times, linguistic research has been especially linked to computer science technology. A language can be now perceived as a system of* quantitative *signs useful for human communication. Emphasis on quantity is focused on its structure analysis by means of computerized corpora, equipped with a series of tools to measure the frequency of words, preferred companies of one and another word, synonyms, key words, etcetera, aspects which reflect the language as an entity mainly subject to fixes*

* Este ensayo es el resumen de una ponencia dada en el *VI Simposio: Enseñanza-Aprendizaje de la Lengua y la Literatura*, que se celebró en la Universidad Autónoma Chapingo los días 1 y 2 de agosto de 2008.

** williecastillejos@hotmail.com

norms, and not to free choices as normally thought. Language is very similar to mathematics. Corpus linguistics seems to prove this assertion.

KEY WORDS: CORPUS, FREQUENCY, TOOLS, REPRESENTATIVENESS, TEXTS

La lengua es una entidad de dos caras, por un lado es un sistema ordenado, con una estructura y un léxico definidos, pero por otro lado tiene la cualidad de arbitrariedad e infinitud. Tenemos entonces la no relación entre la imagen acústica y la imagen conceptual (arbitrariedad) y las múltiples formas que tenemos de emitir un mensaje (infinitud), y al mismo tiempo tenemos la sujeción a las normas lingüísticas. La observación de estas normas, traducidas en comportamientos lingüísticos es una tarea que ha ocupado la investigación de la lengua desde los albores de esta ciencia. Sin embargo, recientemente ciertos métodos de investigación han revolucionado la idea que teníamos del comportamiento lingüístico. Dicha revolución ha ocurrido a la par del desarrollo de las tecnologías computacionales y de la información. Un sitio muy importante lo ocupa los corpus lingüísticos automatizados, principal instrumento metodológico de la lingüística de corpus.

¿Qué es un corpus lingüístico automatizado? A continuación la definición de algunos autores:

AUTOR	DEFINICIÓN
Meyer (2002: XI)	[…] a collection of texts or parts of texts upon which some general linguistic analysis can be conducted.
McEnery y Wilson (2001: 29)	[connotations of the term 'corpus'] may be considered under four main headings: sampling and representativeness, finite size, machine-readable form, and standard reference.
Hunston (2002: 3 y 32)	[a corpus is] nothing other than a store of used language [but a] basic distinction that needs to be made is between a corpus as a collection of texts and a corpus as a collection of samples of language (also cited in Scott 2000).
Kennedy (1998: 4)	A corpus constitutes an empirical basis not only for identifying the elements and structural patterns which make up the systems we use in a language, but also for mapping out our use of these systems.

McEnery, Xiao y Tono (2006: 4)	[…] a corpus can be defined as a body of naturally occurring language, though strictly speaking: It should be added that computer corpora are rarely haphazard collections of textual material: They are generally assembled with particular purposes in mind, and are often assembled to be (informally speaking) *representative* of some language or text type. (Leech, 1992: 116)

De estas definiciones se recogen las siguientes características importantes de un corpus:

- Tiene un propósito estrictamente ligado al estudio de la lengua. No es una recopilación de textos electrónicos tipo biblioteca virtual o archivo electrónico.
- La selección de textos no es arbitraria sino que corresponde a una variedad particular de una lengua, de la cual es una muestra representativa.
- La automatización de los textos permite su manipulación mediante instrumentos informáticos que permiten, por ejemplo, la localización rápida de una palabra, su contexto inmediato anterior y posterior, su frecuencia y su categoría gramatical.

Pero quizás el aspecto más sobresaliente de un corpus es que permite el estudio de la lengua real, la que los hablantes usan cuando escriben o hablan de manera natural. La enseñanza de una lengua no se encuentra ya sujeta a las intuiciones de un autor o a los ejemplos forzados. Varios son ya los libros cuya enseñanza de la lengua proviene de los datos proporcionados por un corpus. Este método es, sobre todo, utilizado en la enseñanza del inglés. Asimismo, el trabajo lexicográfico se vincula ya con el análisis de datos contenidos en un corpus. La Real Academia Española cuenta con un Corpus de referencia del español actual: CREA, en el que se pueden consultar las frecuencias y usos de vocablos en diferentes campos del conocimiento.

La posibilidad de que actualmente podamos aprender acerca de la lengua de este modo, no es fortuita. Desde siempre, su estudio ha supuesto el análisis de ciertas cantidades de texto a fin de obtener regularidades, en otras épocas el análisis de corpus era totalmente manual, tomaba mucho tiempo y era más propenso a errores en los resultados. El avance de la

informática ha cambiado este panorama y lo que antes podía obtenerse en días, ahora se obtiene en segundos, con mayor precisión y con base en conclusiones que provienen de un gran número de datos. Estadísticamente hablando, si la lengua constituye una población numerosa de datos léxicos y gramaticales y de formulaciones *cuasi* infinitas, el estudio de los datos de un corpus automatizado es una representación real de esa población a partir de la cual se pueden obtener conclusiones acertadas. En este sentido, el estudio de los corpus contribuye al estudio científico de la lengua.

Sin demeritar las posibilidades infinitas de las expresiones lingüísticas, a través de un corpus automatizado, podemos observar que la lengua sigue un patrón constante. La idea de infinitud se ve limitada por la regularidad con la que ciertas palabras se ven constantemente acompañadas de otras palabras, de esas y no de otras. Lo mismo ocurre con las expresiones gramaticales, a ello me refiero cuando hablo de un *comportamiento* de la lengua. De pronto pareciera que lo subjetivo se vuelve objetivo, que este aspecto humano tan variado en dialectos y más variado aun en idiolectos, se puede capturar y describir con la certeza de la información de una fuente confiable: los hablantes y el uso que dan a la lengua, plasmados en una especie de fotografía lingüística.

Un corpus lingüístico automatizado tiene básicamente dos funciones: la investigación y la enseñanza de las lenguas. En cuanto a la primera, existen los especializados que permiten estudiar aspectos particulares de la lengua, por ejemplo, los que etiquetan los errores lingüísticos de los hablantes, los de personas que aprenden una segunda lengua y los orales, entre otros. Lo que se pretende descubrir en este tipo de corpus son los tipos de errores más comunes entre los hablantes de una lengua nativa o extranjera, las estrategias que utilizan para comunicarse en una segunda lengua, el grado de interferencia entre lengua nativa y lengua extranjera, y las características de la lengua oral, entre tantos otros propósitos de investigación. En lo que se refiere a la enseñanza, los corpus reflejan el uso real de una lengua determinada a través de la manipulación de diversos instrumentos como las listas de frecuencias, las colocaciones, las categorías gramaticales, las concordancias, etcétera. De este modo, a un estudioso de la lengua le resulta útil saber qué palabras acompañan a otras normalmente, cuáles son las posibilidades de combinación de los vocablos de acuerdo con su categoría gramatical, si se trata de una lengua SVO o VSO, por ejemplo.

Para entender más en qué consisten estos corpus lingüísticos automatizados y cuáles son los aportes que pueden dar al aprendizaje de una lengua, explico dos ejemplos:

1. *Corpus del español (100 millones de palabras)*

Se trata de un corpus diseñado por Mark Davies de la Brigham Young University. Contiene más de 20 000 textos del español de los siglos XIII al XX y permite búsquedas de una manera rápida y sencilla; se pueden buscar palabras exactas, frases, etiquetas, lemas, categorías gramaticales y cualquier combinación de estos datos. En cuanto a las colocaciones permite contextos con un máximo de diez palabras. En su página de Internet, el sistema explica que el corpus permite también hacer búsquedas basadas en la semántica. Por ejemplo, es posible comparar y contrastar las colocaciones de dos palabras diferentes y determinar la diferencia con respecto al significado de estas palabras. También es posible averiguar la frecuencia y la distribución de sinónimos —cerca de 30 000 palabras— y comparar su frecuencia de aparición en diferentes registros, así como en distintos periodos históricos y utilizarlas como parte de otras búsquedas. Finalmente, es posible crear de forma sencilla *listas propias* de palabras relacionadas semánticamente y utilizar dichas listas como parte de la búsqueda. [http://www.corpusdelespanol.org/x.asp]

A esto agréguese que el corpus permite también la búsqueda por frecuencia de uso y la comparación de la misma por registro (lengua oral, ficción, prensa, registro académico) y por periodo histórico (de los siglos XIII al XX).

2. *Wordsmith*

Su creador es Mike Scott. La última versión de este corpus es *Wordsmith Tools 5.0* y consiste en una serie de programas para observar el comportamiento de las palabras en los textos. Contiene tres herramientas principales: *Wordlist*, que proporciona un listado de palabras o grupos de palabras en orden alfabético o de frecuencia; *Concord*, que permite la observación de alguna palabra en contexto y *Keywords*, que identifica las palabras clave de un texto, es decir, aquellas cuya frecuencia es relativa-

mente alta en comparación con otras. A diferencia del corpus de Mike Davies, *Wordsmith* es un programa listo para alimentarse con los textos que se requieran para el objetivo de investigación mientras que *El corpus del español* ya está alimentado y listo para utilizarse en la red.

Pareciera que esta forma de aproximarse a la lengua es muy fría y numérica, pero ¿acaso no es la lengua una entidad que permite su contemplación desde distintos ángulos? Pues bien, este es un ángulo científico que puede aportar interesantes conclusiones acerca de lo que es esta capacidad exclusivamente humana, puede dar explicaciones de cómo opera la lengua en el hombre aunque no pueda lograr lo que el hombre, porque indudablemente la complejidad de esta capacidad es tal que con todo y los avances tecnológicos, la lengua y los pensamientos sean lo único que se le quede al hombre para él solo.

BIBLIOGRAFÍA

Davies, Mark (2002), *El corpus del español (100 millones de palabras, siglo XIII-siglo XX)* [http://www.corpusdelespanol.org.] consultado el 15 de julio de 2008.

Hunston, Susan (2002), *Corpora in Applied Linguistics*, Cambridge, Estados Unidos, Cambridge University Press.

Kennedy, Graeme (1998), *An Introduction to Corpus Linguistics*, Londres, Inglaterra, Longman.

McEnery, Tony y Andrew Wilson (2001), *Corpus Linguistics*, Edimburgh, Reino Unido, Edimburgh University Press.

McEnery Tony, Richard Xiao y Yukio Tono (2006), *Corpus-Based Language Studies*, Londres/Nueva York, Inglaterra/Estados Unidos, Routledge.

Meyer, Charles F. (2002), *English Corpus Linguistics*, Cambridge, Estados Unidos, Cambridge University Press.

Scott, Mike (1996), *Wordsmith Version 5.0,* Oxford University Press [http://www.lexically.net/wordsmith/] consultado el 3 de Julio de 2008.