TEISEL
Tecnologías para la investigación en segundas lenguas
ISSN 2696-676X
Universitat de Barcelona
Facultat de Filologia i Comunicació

Reseña de / Review of / Ressenya de

Rojo, Guillermo (2021). *Introducción a la lingüística de corpus en español*. Routledge

ISBN: 9780367635848 (tapa blanda); 9780367635855 (tapa dura); 9781003119760 (libro electrónico); DOI: http://doi.org/10.4324/9781003119760

Rogelio NAZAR
Pontificia Universidad Católica de Valparaíso (Chile)
rogelio.nazar@pucv.cl
https://orcid.org/0000-0002-8853-1353

Palabras clave: metodología; lingüística de corpus en castellano; lexicología y gramática basadas en corpus; investigación lingüística

English:
***Keywords**: methodology; corpus linguistics in Spanish; corpus-based lexicology and gramar; linguistic research*

Català:
***Paraules clau**: metodologia; lingüística de corpus en castellà; lexicologia i gramàtica basada en corpus; recerca lingüística*

El libro de Guillermo Rojo que aquí se reseña se ubica en una tradición de volúmenes que introducen la metodología de trabajo con corpus, en la que podemos encontrar obras tales como las de Sinclair (1991), Stubbs (1996), McEnery y Hardie (2012) y Brezina (2018), entre otras, pero este tiene la particularidad de estar destinado a un lector hispanohablante. Viene a cubrir una necesidad, ya que no existía un manual para utilizar en cursos de grado o posgrado sobre estos temas, adaptado a las características de la lengua castellana. Será de ayuda en materias relacionadas con la lingüística de corpus pero también con la metodología de investigación lingüística. Sin duda, ejercerá influencia en investigadores de áreas diversas, debido a que las destrezas técnicas que expone son transversales, por ejemplo, a estudios vinculados con el léxico, la terminología y la gramática, ya sea que trabajen con corpus monolingüe o bilingüe, en primera (L1) o segundas lenguas (L2). Se trata de una obra de calidad, redactada con maestría y autoridad por quien posee la capacidad de procesar datos de primera mano. Al mismo tiempo, el libro es también testimonio de cómo ha cambiado la lingüística española en menos de 20 años ya que, por entonces, hablar de algunos de los temas que aquí se discuten podía atraer miradas de extrañeza. Con este libro, muchos conceptos nuevos y su terminología entran al *mainstream* de la lingüística en español.

En los primeros tres capítulos, Rojo presenta los aspectos básicos del trabajo con corpus. En el primero, titulado "La explotación básica de los corpus", cubre tanto temas técnicos como conceptos lingüísticos fundamentales.

Nazar, Rogelio (2022). Reseña de Rojo, Guillermo (2021). *Introducción a la lingüística de corpus en español*. Routledge. *TEISEL. Tecnologías para la investigación en segundas lenguas*, 1, Reseña e-rs01 (1-5). https://doi.org/10.1344/teisel.v1.36825

 | Recepción: 05/10/2021 | Revisión: 08/10/2021 | Aceptación: 22/11/2021 | Publicación: 03/01/2022

ISSN 2696-676X

Es aquí donde delimita el área de trabajo del libro y prepara al lector para los capítulos siguientes. Explica qué es la lingüística de corpus, qué es un corpus y los distintos tipos de técnicas y tecnologías disponibles en la actualidad para la investigación en este ámbito. Con el capítulo 2, "La lingüística de corpus y la metodología de la investigación lingüística", la introducción a la lingüística de corpus se articula con el tema más general de la metodología de la investigación científica. En este lugar se discute sobre cuestiones epistemológicas y se dialoga con otras aproximaciones metodológicas ya establecidas. Cabe destacar aquí, además, el enfoque histórico en el desarrollo de la metodología lingüística. En el capítulo 3, "Diseño, construcción y explotación de corpus", incluye aspectos ya un poco más operativos como la anotación, codificación y análisis de datos. El capítulo incluye también aspectos legales y éticos que, tal como ya habían señalado McEnery y Hardie (2012), conviene considerar. Se declara como un manual orientado a la docencia, y por ello en los capítulos iniciales incluye algunos temas ya transitados, como las dos culturas de Snow (1959), la enciclopedia china de Borges (1952) y la clasificación de las ciencias de Bunge (1972).

En los capítulos siguientes, el libro comienza a profundizar en los temas centrales de la investigación basada en corpus. En el capítulo 4, con el título "Recuperación de información contenida en corpus textuales: el léxico", se trabaja primero el concepto de palabra y se destaca la importancia del estudio de la frecuencia léxica, un aspecto de la lexicología que ha atraído cada vez más atención desde que la lingüística de corpus se consolidó como metodología de trabajo. Un campo que se desarrolla ampliamente en este capítulo es el estudio de la variación del vocabulario, desde los puntos de vista diatópico, diacrónico, diastrático y diafásico. Posiblemente, algunos lectores encontrarán que un aspecto menos desarrollado es el análisis semántico basado en corpus. Menciona el tema por ejemplo al presentar el estudio de la coaparición en el análisis de la polisemia. Explica el caso de *bombilla*, que aparecerá junto a distintas palabras dependiendo de que se use con el sentido de foco o como la bombilla del mate, etc. Sin embargo, se trata de un tema complejo que podría haberse desarrollado de manera más extensiva.

El capítulo 5, "Recuperación de información contenida en corpus textuales: fenómenos gramaticales", es probablemente el más atractivo del libro. Resultará útil como guía para investigadores interesados en aplicar la lingüística de corpus al estudio de los fenómenos gramaticales, pero sin duda será iluminador también para los gramáticos que aún basan su metodología de trabajo en la introspección. El interés radica en que, en lingüística de corpus, el análisis gramatical es generalmente más complejo que el léxico, ya que muchos de los fenómenos gramaticales no son directamente observables o presentan mayor dificultad para su medición con medios mecánicos. En este capítulo se estudia, por ejemplo, el comportamiento verbal, los adverbios en *-mente*, los fenómenos de concordancia, la adaptación de préstamos, las relaciones sintácticas y algunos aspectos de la variación diacrónica de la gramática española, entre otra diversidad de temas que pueden ser muy estimulantes particularmente para investigadores que se encuentren en las primeras etapas de su carrera y estén aún definiendo su línea de trabajo o tema de investigación.

El capítulo también dedica atención al trabajo con el español como L2. Describe técnicas que pueden ser de utilidad para la docencia, ya que pueden ser fuente de inspiración para proponer ejercicios a los estudiantes y de esa

Nazar, Rogelio (2022). Reseña de Rojo, Guillermo (2021). *Introducción a la lingüística de corpus en español*. Routledge. *TEISEL. Tecnologías para la investigación en segundas lenguas*, 1, Reseña e-r01 (1-5). https://doi.org/10.1344/teisel.v1.36825

 | Recepción: 05/10/2021 | Revisión: 08/10/2021 | Aceptación: 22/11/2021 | Publicación: 03/01/2022

ISSN 2696-676X

forma llevar el corpus al aula. Interesará particularmente a los lectores de la revista *TEISEL* porque formula preguntas y expone una amplia variedad de posibilidades de investigación en L2. Presenta los usos de corpus formados por producciones orales o escritas de aprendientes de español como L2 y explica, por ejemplo, ya sea la relación entre las variables nivel de conocimiento de la L2 y la riqueza de vocabulario o la frecuencia de aparición de secuencias gramaticalmente incorrectas, como artículo + demostrativo + sustantivo, etc., o bien la relación entre las variables L1 y L2 en la interferencia lingüística. En este último caso, se observan regularidades en el desempeño de los aprendientes de castellano en función de su L1, en cuanto a características y tipos de error, tal como la selección incorrecta de preposiciones en combinación con verbos.

En el capítulo 6, "Otras cuestiones centrales en lingüística de corpus", el autor aborda una serie de reflexiones generales sobre la lingüística de corpus, entre las que cabe destacar, nuevamente, los antecedentes históricos de la disciplina. Es también el lugar en el que presenta la discusión de algunos aspectos de lingüística cuantitativa y la estructura estadística de la lengua. Aquí establece, además, la relación con conceptos generales de la estadística como el muestreo, ya que explica los problemas de la determinación del tamaño de la muestra, su representatividad y equilibrio. Presenta también una reflexión en profundidad sobre el presente y el futuro de la lingüística de corpus, así como su diálogo con otras ramas lingüísticas y otras disciplinas.

El capítulo 7, por último, se titula "Herramientas de recuperación de datos: resumen y ampliación" y está dedicado a cuestiones más avanzadas de informática, como el uso de expresiones regulares y algunos comandos de Linux. El capítulo no incluye un desarrollo en profundidad de los aspectos más técnicos como el tratamiento estadístico de datos, medidas de evaluación, aplicación de etiquetadores morfosintácticos ni conceptos básicos de programación. Esto se debe seguramente al tipo de público al que va dirigida la obra, ya que la idea es que pueda comprenderse sin necesidad de ser experto en lingüística computacional. Sin embargo, tal vez hubiese sido deseable aprovechar la oportunidad de presentar un capítulo algo más desafiante para este tipo de lector.

Posiblemente, un mayor desarrollo de conceptos de programación, más allá de lo que explica sobre AWK (Aho, Kernighan y Weinberger, 1988), hubiese sido beneficioso, ya que esta habilidad representa una forma de alfabetización y de acceso al pensamiento lógico. Más allá de la cuestión meramente operativa, resulta conveniente tener la capacidad de pensar con símbolos y lenguajes formales. En este ámbito, además, la cuestión del sistema operativo no es solo un aspecto técnico: un sistema operativo es prácticamente una ideología, y como tal puede inhibir la libertad y la creatividad. Cierto tipo de herramientas, más allá de sus interfaces gráficas amigables, no ofrecen al investigador ni la autonomía ni la potencia de procesamiento de lenguajes como Perl o Python que, a la vez, son simples en comparación con otros. Rojo no ignora este hecho, y lo explicó en una charla en Barcelona allá por 2010, en la que intentó convencer a la audiencia asegurando que "nadie se ha muerto por aprender un poquito de Perl". La gente se reía pero la cuestión es bastante seria. Las habilidades de programación ubican al investigador en una situación distinta a la

Nazar, Rogelio (2022). Reseña de Rojo, Guillermo (2021). *Introducción a la lingüística de corpus en español*. Routledge. *TEISEL. Tecnologías para la investigación en segundas lenguas*, 1, Reseña e-r01 (1-5). https://doi.org/10.1344/teisel.v1.36825

 | Recepción: 05/10/2021 | Revisión: 08/10/2021 | Aceptación: 22/11/2021 | Publicación: 03/01/2022

ISSN 2696-676X

de quien utiliza interfaces que solo permiten realizar las operaciones decididas por quien diseñó el programa y que, además, tienden a orientarse hacia el trabajo cualitativo.

Otra cuestión que podría echarse a faltar en el libro ya en general es un mayor énfasis en el contraste entre métodos cuantitativos y cualitativos. En lingüística de corpus es muy frecuente el enfoque cualitativo, y el corpus suele utilizarse más como medio de comprobación de teorías que como medio de descubrimiento, aunque el tema no se discute mayormente. Por ejemplo, cuando en el libro se estudia la alternancia *cocodrilo/crocodilo*, resulta necesario saber, de manera previa, que tal alternancia existe, para buscar cada una de las formas y observar su frecuencia de aparición en el corpus. Desde el punto de vista cuantitativo, en cambio, cabe otro tipo de preguntas y métodos. Es posible, por ejemplo, encontrar estas alternancias aplicando medidas de similitud ortográfica o semántica. Otro ejemplo similar se da en la descripción que ofrece de la interfaz de búsqueda de enegramas de Google Books. La interfaz permite buscar palabras para obtener curvas de distribución de frecuencias en la serie temporal. Sin embargo, lo verdaderamente interesante del recurso es descargar la totalidad del corpus, ya que entonces las posibilidades se multiplican. Con métodos cuantitativos y la utilización de un lenguaje de *scripting* es posible, por ejemplo, proceder al revés, y así introducir curvas para obtener las palabras, claramente un resultado que no puede obtenerse con el método cualitativo.

Es evidente que un libro no puede desarrollar todos los temas con la misma profundidad y siempre habrá algunos que sacrificar. Así, debe valorarse por lo que es y no por lo que podría haber sido, y la valoración es muy positiva. El lector modelo es un individuo con formación en humanidades, sin un nivel avanzado de alfabetización tecnológica. Pero también podría ser útil para otra audiencia constituida por el grupo diametralmente opuesto, es decir, investigadores con formación en informática que tengan interés en dedicarse al procesamiento del lenguaje natural y necesiten fortalecer sus conocimientos de lingüística. Cualquiera sea el caso, se trata de un manual útil para la lingüística de corpus en español y también para la docencia en lingüística general. Dada su selección y organización de temas, puede adaptarse a distintas necesidades, pero podría ser especialmente útil para vertebrar el programa de un curso de lingüística de corpus de un semestre. Proporciona la base que permitirá luego a cada cual elegir un recorrido de lecturas. Para el desarrollo de destrezas técnicas, se requerirá inevitablemente de un periodo de ejercitación práctica. Para la formación avanzada en lingüística cuantitativa, el libro puede complementarse con otras lecturas como el manual de Manning y Schütze (1999) o el de Herdan (1964). Para finalizar, cabe señalar que esta obra es un recordatorio del privilegio de los lingüistas de la época actual, que disponen de tantas herramientas y posibilidades.

Referencias

Aho, Alfred; Kernighan, Brian; Weinberger, Peter (1988). *The AWK programming language*. Addison-Wesley Publishing Company.

Borges, Jorge Luis (1952). El idioma analítico de John Wilkins. En *Otras inquisiciones* (pp. 121-125). Sur.

Nazar, Rogelio (2022). Reseña de Rojo, Guillermo (2021). *Introducción a la lingüística de corpus en español*. Routledge. *TEISEL. Tecnologías para la investigación en segundas lenguas*, 1, Reseña e-r01 (1-5). https://doi.org/10.1344/teisel.v1.36825

 | Recepción: 05/10/2021 | Revisión: 08/10/2021 | Aceptación: 22/11/2021 | Publicación: 03/01/2022

ISSN 2696-676X

Brezina, Vaclav (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press. https://doi.org/10.1017/9781316410899

Bunge, Mario (1972). *La ciencia: su método y su filosofía*. Siglo veinte.

Herdan, Gustav (1964). *Quantitative linguistics*. Butterworths.

Manning, Christopher; Schütze, Hinrich (1999). *Foundations of statistical natural language processing*. MIT Press.

McEnery, Tony; Hardie, Andrew (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Sinclair, John (1991). *Corpus, concordance, collocation*. Oxford University Press.

Snow, Charles Percy (1959). *The two cultures*. Cambridge University Press.

Stubbs, Michael (1996). *Text and corpus analysis*. Blackwell.

Nazar, Rogelio (2022). Reseña de Rojo, Guillermo (2021). *Introducción a la lingüística de corpus en español*. Routledge. *TEISEL. Tecnologías para la investigación en segundas lenguas*, 1, Reseña e-r01 (1-5). https://doi.org/10.1344/teisel.v1.36825

 | Recepción: 05/10/2021 | Revisión: 08/10/2021 | Aceptación: 22/11/2021 | Publicación: 03/01/2022