

Johannes Kabatek (ed.) (2016): *Lingüística de corpus y lingüística histórica iberorrománica*, con la colaboración de Carlota de Benito Moreno, Beihefte zur Zeitschrift für romanische Philologie, Edition Niemeyer, volumen 405, Berlín/Boston, Walter de Gruyter, 448 págs.

En la actualidad, la lingüística histórica basada en corpus se enmarca en una nueva etapa, caracterizada por una visión más crítica, pero también más refinada y más adecuada a su objeto de estudio. Junto a estos avances, también han surgido nuevos problemas y nuevas exigencias, basadas en el cuestionamiento de la relación entre datos primarios e historia de la lengua, el cuestionamiento de los mismos datos primarios, la crítica hacia el diseño de los corpus y la crítica hacia las posibilidades ofrecidas por los corpus y los bancos de datos. Ante este nuevo panorama, los investigadores también se han dedicado a remediar estos puntos críticos. Gracias a ello, hoy no solo disponemos de más y mejores corpus y herramientas para el tratamiento de los datos, sino también de una serie de nuevos estándares más o menos establecidos en la comunidad. Precisamente, esta nueva generación de la lingüística histórica que trabaja con corpus supone el eje central de la obra colectiva reseñada.

Es evidente que el mismo término «lingüística de corpus» hoy en día comprende disciplinas interrelacionadas y relativamente diferenciadas, entre las que se pueden distinguir, al menos, tres vertientes y finalidades: en primer lugar, encontramos la lingüística de corpus, disciplina que se ocupa de la creación de corpus, es decir, de los pasos que van desde la recolección de los datos primarios hasta su presentación en una plataforma consultable. Una segunda disciplina es la que está en estrecha relación con la informática y se ocupa tanto del tratamiento de los datos y de su etiquetaje como de los análisis cuantitativos y estadísticos. Por último, la tercera vertiente se encarga de realizar una «lingüística con corpus» ocupándose de fenómenos concretos de la historia de la lengua y basando su análisis en datos de corpus. Esta vertiente final es la más presente en este libro, aunque se debe mencionar que, en varios de los trabajos, se constata que la creación del corpus y la investigación de los fenómenos no están separados, sino en las mismas manos.

El enfoque adoptado en este libro es *iberorrománico*, en tanto que agrupa trabajos sobre diferentes lenguas iberorrománicas, salvando la conocida «teleología invertida» que también afecta a los corpus. La finalidad de optar por este enfoque es doble: por un lado, las razones están en los mismos objetos de estudio, en los datos y fenómenos relacionados, y, por otro, en el hecho de que el intercambio y la comparación son útiles en sí mismos, especialmente en una disciplina no muy antigua y a la vez muy dinámica como es la lingüística de corpus.

Esta obra colectiva sobre la lingüística histórica y de corpus, editada por Johannes Kabatek, en colaboración de Carlota de Benito, reúne un total de diecio-

cho trabajos. Estas contribuciones han sido fruto, por un lado, del *Tercer Coloquio Internacional sobre Corpus diacrónicos en lenguas iberorrománicas* (CODILI III), celebrado en la Universidad de Zúrich en verano del 2014 y, por otro, del curso de invierno ALPES (Abriendo Líneas en el Pasado del Español), celebrado en Kandersteg (Berna) en enero de 2016.

El libro se abre con un capítulo introductorio, a cargo del editor, en el que se plantea el tema central de la obra y los puntos principales de debate, desde una perspectiva crítica. Los estudios que componen el volumen se dividen en cuatro apartados; aunque la repartición no es una separación tajante, sí se sigue una línea temática: la primera sección agrupa trabajos cuyo objetivo principal es reflexionar acerca de problemas generales de la lingüística de corpus, partiendo de casos y cuestiones empíricas concretas; la segunda sección se dedica a la presentación de corpus; la tercera sección, a los análisis cuantitativos; y la cuarta sección reúne análisis de diversa índole (cualitativo, variacional e idiolectal) basados en trabajos con corpus.

En el primer trabajo, Enrique-Arias reflexiona sobre lo que él mismo denomina «parámetro perspectiva», es decir, la manera en que los estudiosos acceden a los datos lingüísticos. Además, tomando como modelo el corpus *Biblia Medieval*, presenta una serie de ventajas de los corpus paralelos, los cuales permiten un acercamiento diferente a la materia, respecto a los corpus convencionales, al pasar la traducción por dos fases, una fase semasiológica y otra onomasiológica y encontrar así soluciones inesperadas para un mismo contenido o un contenido semejante. Siguiendo esta misma línea, del Rey Quesada también destaca la utilidad de los corpus paralelos y defiende la necesidad de disponer de más corpus de esta tipología con textos traducidos para poder medir el impacto de la traducción en comparación con las producciones originales de una lengua en diferentes épocas. Por su parte, Octavio de Toledo y Huerta intenta, a través de una serie de ejemplos, demostrar el valor del CORDE como herramienta fiable para el estudio filológico del *primer español moderno*, español hablado desde finales del siglo XVII hasta principios del XIX). Precisamente este período es fundamental para los estudios históricos y diacrónicos y está escasamente estudiado. Cierra la primera sección, Torruella, quien reflexiona acerca de una mejora del Corpus Informatizat del Català Antic (CICA), especialmente en cuestiones relativas a la representatividad del corpus y el equilibrio de los datos contenidos en él, con tres propuestas concretas: dividir los periodos cronológicos en cortes de 50 años, evitando así cortes de tipo histórico o sociocultural; mantener el equilibrio textual mediante un sistema de correspondencia «equivalente» para los ejes diacrónico y tipológico; y, por último, representar la mayor riqueza léxica posible cuantificándola mediante un índice NOMC %.

La segunda sección se abre con el trabajo de Álvarez Blanco y González Seoane, quienes presentan las características del corpus gallego Gondomar, el cual recoge todo tipo de documentación desde la época de los *Séculos Oscuros* hasta el

siglo XVIII. El segundo trabajo, a cargo de María Francisca Xavier, está dedicado a la presentación de la historia y las posibilidades del Corpus Informatizado do Português Medieval (CIPM). Este corpus, diseñado ya en los años noventa, ha ido ampliándose y complementándose hasta servir como base para el *Dicionário do Português Medieval* gracias al vínculo directo existente entre corpus y lexicografía. En el siguiente capítulo, Marcet Rodríguez y Sánchez González de Herrero presentan otro corpus medieval, en esta ocasión de ámbito regional y de reciente creación. Este corpus, restringido a la zona de Miranda de Ebro, está formado por un total de 203 documentos, editados según los criterios de la Red CHARTA. En este trabajo se presentan dichos documentos y se analizan teniendo en cuenta variables gráficas y morfosintácticas. El tercer capítulo de la sección, elaborado conjuntamente por Carvalheiro, Costa, Marquilhas, Pinto, Pratas y Vaamonde, presenta el corpus Post Scriptum, de la Universidade de Lisboa. Este corpus recoge cartas privadas tanto en portugués como en español moderno. Los autores demuestran la utilidad de este corpus para estudiar la geografía y la diacronía de ciertos fenómenos lingüísticos mediante tres ejemplos concretos: el marcador discursivo *pois* y el relativo *cujo* en portugués y la cuestión del leísmo, laísmo y loísmo en español. Cierra esta segunda sección el trabajo de Rojo sobre los dos corpus del español actual de la RAE: el CREA y el CORPES XXI. El autor no se limita a presentar el CORPES XXI, sino que trata también la representatividad o el salto cuantitativo y cualitativo que hay entre la versión original del CREA y la nueva plataforma del CORPES XXI. Finalmente, plantea nuevos objetivos del CORPES XXI para el futuro.

En la tercera sección se discuten cuestiones cuantitativas y se presentan análisis frecuenciales de diferentes fenómenos. Nieuwenhuijsen muestra la utilidad del trabajo estadístico al presentar un análisis cuantitativo de las oraciones subordinadas interrogativas indirectas y negadas dependientes del verbo *saber*, que se investigan teniendo en cuenta distintas variables (tipo de interrogación, modo verbal, tiempo, región). El trabajo de Schulte y Blas Arroyo se dedica al estudio de la evolución de cinco perífrasis modales en un amplio corpus formado por textos de «inmediatez comunicativa» del siglo XVI al XX. Tras efectuar un análisis multifactorial y frecuencial, identifican correlaciones estadísticamente relevantes. Por su parte, Bouzouita estudia los futuros y condicionales sintéticos medievales seguidos de pronombre en el corpus *Biblia Medieval*. La autora evalúa tres hipótesis que condicionan dicha construcción, aplicando un análisis estadístico que permite reconstruir la casuística que rige las posiciones pronominales y que incluye factores sintácticos, factores morfológicos y factores condicionados por las fuentes de la traducción. El trabajo frecuencial de Torrens Álvarez y Ueda se ocupa de la grafía <j> con valor consonántico en el corpus CORHEN. Gracias al análisis estadístico realizado con el programa LETRAS, se traza la línea del «nacimiento» de esta letra en el corpus.

Este trabajo no solo muestra la importancia de la estadística para el análisis cuantitativo, sino también de la paleografía. El siguiente capítulo también se ocupa del CORHEN; en este caso Moral del Hoyo, sobre 278 documentos procedentes de este corpus y elegidos según criterios diatópicos y diacrónicos, muestra cómo la variación y la evolución de las formas *-ié/ -ía* de imperfecto depende de factores tanto espaciales como estructurales. Cierra la tercera sección el artículo de Carrasco Cantos y García Aguiar, dedicado al análisis del total de los sufijos contenidos en el corpus DITECA (*Diccionario de textos concejiles de Andalucía*), conformado por textos jurídicos del siglo XIII al siglo XVIII. Este análisis frecuencial permite tanto identificar el grado de productividad de los diferentes sufijos analizados como observar cómo los diferentes sufijos se van especializando funcionalmente a lo largo de los siglos.

La cuarta sección la componen tres estudios que realizan distintos tipos de análisis (cualitativo, variacional e idiolectal, respectivamente) a partir del trabajo con corpus. Arias Álvarez y Hernández Mendoza presentan el Corpus Electrónico del Español Colonial Mexicano (COREECOM), conformado con documentos del siglo XVI, y muestran cómo, a partir de los datos de este corpus, puede estudiarse la variación y evolución de ciertos fenómenos. Por su parte, Fernández Alcaide, tomando como corpus cartas privadas del siglo XVI, combina un pormenorizado análisis textual con algunas observaciones de índole general, que destacan la importancia del acceso a información sobre las primeras décadas de las colonias americanas, hasta ahora posible gracias a la recuperación de textos en proyectos como CHARTA o CORDIAM. En el último trabajo de esta obra, Iglesias aplica el método idiolectal para investigar la evolución de la subida de clíticos en los últimos dos siglos. El autor investiga producciones textuales de seis individuos y confirma lo que con otros métodos se había observado con respecto de la evolución del siglo XIX al XX, demostrando así que este *método complementa los resultados obtenidos a partir de otros enfoques*.

En suma, esta obra colectiva nos presenta, a través de diversos trabajos, la situación actual de la lingüística histórica basada en corpus, desde un enfoque ibero-románico. Estas propuestas no solo demuestran que se ha iniciado una nueva etapa en la lingüística histórica, sino que plantean nuevos retos y abren la puerta a futuras investigaciones gracias al desarrollo de numerosas herramientas y de una acertada propuesta metodológica.

Paula Albitre Lamata  
Universidad Complutense de Madrid /  
Instituto Universitario Menéndez Pidal

Fecha de recepción: 2 de abril de 2020  
Fecha de aceptación: 1 de mayo de 2020