estable (Pons); y da a conocer atractivos proyectos de investigación, como el CRELOC, *Clientela y redes locales en la castilla medieval. Estudio histórico y tecnologías documentales* (Jular).

ROBERTO CUADROS MUÑOZ

ANDRÉS ENRIQUE-ARIAS (ed.), *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus,* Madrid/Frankfurt: Iberoamericana/Vervuert, 2009.

En la actualidad, es innegable la importancia que han adquirido los corpus en las investigaciones de historia de la lengua y de gramática histórica. Por ejemplo, la *Sintaxis histórica de la lengua española* dirigida por Concepción Company (2006 y 2009) toma como base para su realización un corpus de datos textuales que sirve para realizar la descripción de los cambios sintácticos que se han producido en la historia del español. Como indica Andrés Enrique-Arias, en la introducción al volumen que aquí se reseña, "hoy en día la investigación de la diacronía de las lenguas iberorromances es prácticamente inconcebible sin recurrir a los grandes corpus diacrónicos, bien como fuente primaria de datos, o para validar hipótesis formuladas a partir de presupuestos teóricos" (p. 14). Esta necesidad de utilizar grandes bancos de datos en los estudios diacrónicos ha generado el nacimiento de una nueva disciplina: la lingüística de corpus. Su origen, según Rojo (2008), se remonta a la creación en 1964 del *Brown University Standard Corpus of Present-Day American English* (Brown Corpus), de Francis y Kučera, por tanto, "tiene algo menos de medio siglo" (véase Rojo 2008 y 2010 para un repaso de la historia y la situación de la disciplina en la actualidad, sobre todo en el ámbito hispánico).

El volumen *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus* es una valiosa contribución al progreso dentro de este ámbito, pues demuestra que, pese a ser un campo de investigación con una corta vida, los avances que se han realizado hasta el momento son extraordinarios, como confirma la exhaustiva recopilación bibliográfica efectuada por Joaquim Llisterri (<http://liceu.uab.es/~joaquim/language_resources/lang_res/biblio_corpus.html>). La obra que reseñamos reúne una cuidada selección de las comunicaciones presentadas en el *Coloquio Internacional de Corpus Diacrónicos en Lenguas Iberorromances* celebrado en Palma de Mallorca en el año 2007. Utilizamos el adjetivo "cuidada", pues el comité científico ha elegido los trabajos con un equilibrio casi

perfecto, ya que la teoría y la práctica, lo general y lo particular, se manifiestan en sus distintos capítulos. Por ello, se convierte en un referente no solo para los investigadores expertos en diacronía, sino para cualquier persona que quiera iniciarse en este terreno.

La estructura general ya da muestras de esta armonía. La publicación se divide en seis secciones: I: Cuestiones filológicas; II: Diseño de corpus para fines específicos; III: Cuestiones de arquitectura informática; IV: Corpus y variedades iberorromances; V: La Biblia como corpus; y VI: Aplicaciones del uso de corpus en la historia del español. Sin embargo, bajo esta clasificación subyace otra, de índole metodológica, que muestra la doble perspectiva con que se puede abordar la investigación en el seno de la lingüística de corpus. Por un lado, la del investigador que crea un corpus prácticamente desde la nada, con un objetivo específico (realizar un diccionario, una gramática, etc.) o sin marcarse más propósito que el de servir de instrumento para cualquier análisis; por otro, la del estudioso que quiere validar sus hipótesis respecto al comportamiento de un fenómeno lingüístico a partir del uso de esta herramienta. Ambos puntos de vista son, sin duda alguna, los que han hecho que este ámbito de la lingüística se halle en una mejora continua y en un perfeccionamiento progresivo.

Como ya se ha señalado antes, las contribuciones que figuran en este libro demuestran los avances que se han realizado y lo que se ha hecho durante la breve existencia de la lingüística de corpus, que es muchísimo y, sobre todo, de gran calidad. No obstante, los análisis de los autores no se limitan a la mera descripción, sino que señalan las dificultades, los problemas y los retos que esta disciplina debe abordar en el futuro. Por este motivo, mi propósito no va a ser detallar únicamente la temática de cada uno de los capítulos, sino dar cuenta de sus aportaciones más relevantes para el progreso en este ámbito.

Son varios los trabajos que se presentan desde la primera de las perspectivas señalada unas líneas más arriba, es decir, la del diseñador de corpus que debe dar cuenta no solo de cuáles van a ser los criterios de selección de los textos, sino también qué clase de explotaciones se podrán hacer de estos. La preocupación por la elección de los documentos que deben componer un corpus es una de las cuestiones que ponen de manifiesto los capítulos elaborados por José Antonio Pascual y Carlos Domínguez, "Un corpus para un *Nuevo diccionario histórico del español*" (pp. 79-93), y Joan Torruella, "Bases científicas en la investigación a partir de corpus: el caso del *Corpus Informatitzat del Català Antic*" (pp. 95-115). Los tres autores llegan a la conclusión de que, para la contrucción de un corpus, el investigador debe tener en cuenta dos parámetros a la hora de seleccionar los textos. En primer lugar, la representatividad de esos documentos y, en segundo lugar, su equilibrio desde el punto de vista diacrónico, diatópico y diastrático-diafásico. Evidentemente, no es tarea fácil

poner en práctica ambos aspectos. Tal y como señala Enrique-Arias en este mismo volumen, "la disponibilidad de materiales no es la que quiere el investigador, sino que está condicionada por la disponibilidad de textos de cada época" (p. 275). Sin embargo, su estricta aplicación puede observarse tanto en el *Corpus del Nuevo Diccionario Histórico* (*CNDH*) como en el *Corpus Informatitzat del Català Antic* (*CICA*) –elaborados por los anteriores investigadores y descritos en sus respectivos capítulos– y es una práctica que define la calidad de cualquier banco de datos. Es decir, el mejor corpus no es el que ofrece un mayor número de palabras sino el que ha hecho una mejor selección de los textos y ha llegado a establecer ese equilibrio entre las variantes que los afectan.

Por otra parte, las posibilidades que debe ofrecer al investigador esta herramienta una vez diseñada es también otro de los aspectos que figura en las páginas de muchas de las contribuciones de la obra. Este es el caso del capítulo dedicado al *Corpus de documentos españoles anteriores a 1700 (CODEA)* elaborado por Pedro Sánchez-Prieto Borja *et al.* (pp. 25-38). Este corpus, que reúne textos de tipo cancilleresco anteriores al siglo XVIII, no limita su empleo como recurso lingüístico, sino que se ha concebido con la voluntad de que sirva en otros ámbitos, como el de la paleografía, la historia general, la archivística o la diplomática, es decir, se trata de un corpus interdisciplinar. Uno de los grandes aciertos del *CODEA*, que lo distingue de otros anteriores, es su triple presentación, ya que se puede consultar el facsímil, una transcripción paleográfica y una presentación crítica. Esto muestra, claramente, la intención expresa de estar siempre al servicio del investigador, sea cual sea el objetivo de su estudio.

La línea iniciada por el *CODEA* se ha continuado en otras iniciativas como la que se presenta en el capítulo "Los documentos americanos (o indianos) en la red CHARTA" elaborado por Micaela Carrera de la Red y Miguel Gutiérrez (pp. 39-53). Esta contribución plantea también una cuestión importantísima que afecta a la construcción de corpus, sobre todo, los de carácter diacrónico: la edición de los textos antiguos. En este sentido, el problema se encuentra en la sistematización y homogeneización de las grafías, ya que los documentos están editados siguiendo distintas metodologías. Esto se manifiesta en el caso, por ejemplo, de la ortografía del español de América, lengua de la documentación que atesora el corpus que se presenta en este capítulo, pero también se puede observar en aquellos que reúnen textos de otras variedades lingüísticas como el judeoespañol, sistema lingüístico a caballo entre el hebreo y el español, como se expone en el capítulo "Pensando en corpus en modo colaborativo: hacia el prototipo del corpus judeoespañol digital" elaborado por Ana Stulic-Etchevers y Soufiane Rouissi (pp. 117-134).

Asimismo, uno de los avances más importantes que se ha conseguido dentro de la lingüística de corpus es la posibilidad de estudiar la multiva-

riación. Es decir, su diseño debe prever los distintos tipos de variación a los que se ve sometida la lengua y que afectan a su evolución. En definitiva, los bancos de datos deben permitir reflejar la influencia que tienen los factores extralingüísticos en el devenir histórico de una lengua. Así, la estructura de cualquier corpus que pretenda poder dar cuenta de la multivariación no solo debe atender al origen geográfico del autor, sino que debe conocer su procedencia social, su sexo y su nivel cultural, y no solo debe establecer la fecha de composición, sino el tipo de texto en el que se inserta o el grado de formalidad que en ella se emplea. Esta preocupación se observa en muchos de los bancos de datos descritos en los capítulos de esta obra como, por ejemplo, los ya citados de Sánchez-Prieto y Carrera de la Red y Gutiérrez, pero también en los de Beatriz Arias, "Confección de un corpus para conocer el origen, la evolución y la consolidación del español en la Nueva España" (pp. 55-75); Ricardo Pichel y Xavier Varela Barreiro, "O corpus do galego medieval *Tesouro Medieval Informatizado da Lingua Galega* (TMILG)" (pp. 195-215); Miguel Calderón y María Teresa García Godoy, "El Corpus diacrónico del español del reino de Granada (CORDEREGRA)" (pp. 229-249); y Mercedes Abad Merino, "Uso de corpus documental en sociolingüística histórica y retos para su elaboración en el sureste peninsular (siglos XV-XVII)" (pp. 251-266). Por ejemplo, en el corpus presentado por Carrera de la Red y Gutiérrez se distinguen diferentes tipologías textuales como consultas, memoriales, probanzas, capitulaciones, así como distintos registros (petición, queja, etc.). Del mismo modo, en el descrito por Beatriz Arias se marca la procedencia social del amanuense según fuera indígena, criollo, mestizo, negro o español, así como el grado de formalidad del texto.

En esta línea, son destacables los esfuerzos que algunos investigadores han puesto en que los textos de los corpus se acerquen lo máximo posible al uso real mayoritario de la lengua de las distintas épocas, es decir, al lenguaje oral. Este hecho, que entronca directamente con la idea de representatividad que exponíamos al principio, se aprecia en el capítulo de Calderón y García Godoy. En sus propias palabras, "los textos que suscitan mayor interés en el proceso de selección son los susceptibles de adscribirse a hablantes de menor conciencia lingüística, verificable en la inseguridad escrituraria revelada y en la recurrencia a usos ortográficos que emulan la lengua oral" (p. 235). De nuevo, pues, los investigadores que se han preocupado por la elaboración de corpus no se limitan a ofrecer grandes cantidades de palabras, sino que pretenden que se puedan recuperar en su contexto extralingüístico y, a partir de ahí, poder crear o validar hipótesis respecto a su evolución.

Esta clasificación de lo extralingüístico no siempre es fácil de efectuar y depende en gran medida del tipo de texto que se esté manejando. Por ejemplo, de sobra son conocidos los problemas que plantean los textos

antiguos en cuanto a su datación, ya que pueden existir diversos manuscritos, copias del mismo texto pero de fechas diferentes, etc., y esto puede distorsionar, a posteriori, el análisis. Esto entronca, además, con un problema importante dentro de la historia de la lengua y la gramática histórica, que incide claramente en la construcción de un corpus: el establecimiento de periodos históricos, estudio conocido como periodización, tal y como se muestra en el citado capítulo de Torruella.

Además, es difícil clasificar un documento en función de la variedad diatópica que refleja en el caso de documentos antiguos, ya que la lengua presentaba un alto grado de variación y falta de fijación. Estas dificultades no deben suponer su marginación del corpus tal y como se expone en los capítulos de Xulio Viejo Fernández *et al.* "Un corpus lingüístico asturiano: *Eslema*" (pp. 183-194) y de Robert J. Blake y Gina Lee "*MOCA*: análisis de corpus en línea en un contexto histórico" (pp. 327-341). Ambos estudios muestran cómo textos híbridos entre el astur-leonés y el romance, en el primer caso, y entre el latín y el romance, en el segundo, pueden también ser muy útiles para describir el español de los orígenes.

Sin duda, otro de los aciertos de la publicación que aquí se reseña es que muestra también la tipología de corpus existentes en la actualidad. Podría pensarse que esta herramienta es solo un gran compendio de documentos, algo similar a lo que es Internet. Pero esta idea es errónea y este libro demuestra claramente no solo que esto no es así, sino que, además, estos pueden ser tan variados como diferentes son los objetivos que pueda tener un investigador. En algunos de los capítulos se exponen diversas propuestas de corpus con propósitos muy distintos, pero con un hilo conductor común: servir de herramienta para el análisis de la evolución lingüística.

Uno de los tipos de corpus que se presenta en este volumen es aquel que tiene por finalidad satisfacer las expectativas de cualquier investigador sea cual sea su objeto de estudio (léxico, morfológico, semántico, sintáctico, etc.). Este es el caso del *Corpus del español* creado por Mark Davies y cuyas ventajas se describen en el capítulo "Creating useful historical corpora: A comparison of *CORDE*" que presenta el capítulo, the *Corpus del español*, and the *Corpus do português*. Dejando aparte las cuestiones comparativas respecto al *CORDE*, cabe remarcar la dificultad que supone construir un corpus que abarque todas las posibilidades de análisis que un investigador pueda tener. Y, como muy detalladamente muestra Davies, el *Corpus del español* (y también el *Corpus do português*), según su opinión, se acercan a ese ideal porque "were designed with linguistic research in mind, rather tan just as an afterthought" (p. 165).

Otra de las propuestas que se describe es la expuesta por Aitor García Moreno en "El proyecto de 'Edición del fichero manuscrito de léxico judeoespañol de Cynthia Crews'" (pp. 217-228). En este trabajo, lo tradicional y lo innovador se dan la mano, ya que el objetivo principal de este proyec-

to es convertir en corpus un trabajo desarrollado de manera tradicional (a través de fichas), para poder realizar un mejor y más ágil aprovechamiento de la información.

Por otro lado, cabe destacar también el capítulo de Andrés Enrique-Arias titulado "Ventajas e inconvenientes del uso de *Biblia medieval* (un corpus paralelo y alineado de textos bíblicos) para la investigación en lingüística histórica del español" (pp. 269-283), pues nos descubre las utilidades de un corpus, constituido por las distintas ediciones de los textos bíblicos medievales, que hace posible la consulta simultánea de todos ellos. Esta modalidad representa un valor añadido, ya que permite comparar el mismo texto editado o traducido en épocas distintas, algo sumamente importante en cualquier estudio de índole diacrónica.

Finalmente, los capítulos de Kim Schulte "Using non-annotated diachronic corpora: benefits, methods and limitations" (pp. 167-180) y de Robert J. Blake y Gina Lee "*MOCA*: análisis de corpus en línea en un contexto histórico" (pp. 327-341) ponen de manifiesto la idea de que muchas veces el estudioso no satisface sus necesidades investigadoras a partir de los corpus existentes. Sin embargo, ello no debe impedir que el estudioso siga adelante porque existen herramientas informáticas que permiten poder diseñarlos "a medida", como las que se presentan en ambos trabajos (*regular expressions*, en el caso de Schulte; un programa de concordancias, en el de Blake y Lee). Estos trabajos prueban que los límites de la investigación no se hallan en los corpus, sino que los impone el propio investigador.

Hasta aquí se han descrito las aportaciones más relevantes que, a nuestro juicio, se ponen de manifiesto en la obra que aquí se reseña desde la perspectiva del diseñador de corpus. Pero, como se ha señalado al principio de esta recensión, el valor de este libro aumenta al incorporar también trabajos realizados desde otro punto de vista, es decir, el del investigador que los utiliza como herramientas de estudio en sus investigaciones.

Todos los capítulos que se abordan desde esta perspectiva coinciden en la utilidad de esta herramienta en los estudios diacrónicos, ya que estos permiten refutar o validar hipótesis ya planteadas o bien formular nuevas a partir de datos empíricos. Este hecho se pone de manifiesto en los capítulos de Valentina Vincis, "Describir textos a partir de estructuras de ilación oracional: un corpus de romanceamientos bíblicos medievales" (pp. 285-303); Lola Pons Rodríguez, "Notas sobre los nexos concesivos en los romanceamientos bíblicos medievales" (pp. 305-324); Josep Alba-Salas, "Las estructuras tipo *meter miedo* en diacronía: más detalles sobre la evolución histórica de las colocaciones causativas" (pp. 343-363); Dorien Nieuwenhuijsen, "El rastreo del desarrollo de algunos pronombres personales en español: (im)posibilidades de los corpus diacrónicos digitales" (pp. 365-384); Natalya I. Stolova, "Los corpus diacrónicos al servicio del estudio de los arcaísmos gramaticales" (pp. 385-400); y Enrique Pato,

"Notas aclaratorias sobre la historia del indefinido *alguien*: una aplicación directa del uso de corpus diacrónicos" (pp. 401-416). A pesar de ello, los análisis efectuados por estos autores (de índole mayoritariamente sintáctica) muestran claramente una visión crítica del empleo de los corpus en las investigaciones lingüísticas en aras de su mejora y perfeccionamiento.

Por ejemplo, los trabajos de Vincis y Pons Rodríguez ponen en evidencia que la tipología textual debe ser un índice a tener en cuenta no solo a la hora de diseñar un corpus (como han señalado otros autores de este libro), sino también en el momento de efectuar cualquier análisis. A su vez se lamentan del peso excesivo que tienen los textos literarios en algunos de ellos y la poca importancia que se concede a otros tipos textuales como los romanceamientos bíblicos, que, según su opinión, también pueden ser un fiel reflejo del uso lingüístico. Esta misma reivindicación, pero para la prosa cancilleresca, se manifiesta también en el capítulo de Sánchez-Prieto *et al.* citado antes. Por otro lado, los capítulos de Nieuwenhuijsen y Stolova dan cuenta de uno de los problemas que el investigador puede encontrarse cuando trabaja con corpus como es la imposibilidad de algunos de ellos de responder a búsquedas demasiado abiertas por la excesiva cantidad de datos que se obtiene, algo fácilmente superable si se parte de un corpus bien construido y lematizado. Finalmente, los estudios de Alba Salas y Pato muestran las posibilidades de esta herramienta para analizar el contraste entre unidades lingüísticas a lo largo de su historia. Como es bien sabido, cualquier proceso de cambio se caracteriza por la convivencia de dos formas, la no marcada y la marcada o innovadora. Los corpus bien diseñados permiten observar la lucha que se establece entre estas dos formas, tal y como se examina en el caso de los verbos *poner* y *meter* en las colocaciones causativas estudiadas por Alba Salas o en la evolución del indefinido *alguien* en competencia con *alguno* en el trabajo de Pato. Sin duda, cuanto mejor estructurado y más detallada sea la lematización de un corpus, mayor será el aprovechamiento que el investigador podrá conseguir con el objetivo de evidenciar los pasos seguidos por el cambio lingüístico.

En definitiva, como se ha tratado de mostrar en esta recensión, esta obra no es únicamente descriptiva sino que se caracteriza porque todos los trabajos que figuran en ella presentan una marcada visión crítica. Esto es así ya que solo desde esta perspectiva es posible mejorar y perfeccionar cualquier empresa. Por ello, estamos totalmente de acuerdo con Enrique-Arias cuando, en la introducción a esta publicación, afirma que "a pesar de los avances recientes en lingüística de corpus, estamos lejos de aprovechar en todo su potencial las oportunidades que ofrece el uso de las nuevas tecnologías para el estudio histórico de las lenguas iberorrománicas" (p. 15). De hecho, uno de los retos futuros de los investigadores en historia de la lengua y gramática histórica para sacar mucho más partido a esta

herramienta en sus investigaciones es profundizar en el terreno de la estadística, tal y como se expone en el capítulo de Torruella. Sabemos que este interés por seguir mejorando esta disciplina ya ha dado sus frutos, pues recientemente se ha celebrado el *Simposio Internacional de Corpus en Lenguas Iberorrománicas* (Barcelona, 10-12 de febrero de 2011), heredero del *Coloquio* cuyas contribuciones aquí se reseñan. Por tanto, la continuidad del diseño y la explotación de nuevos corpus parece estar garantizada, por lo que esta obra ha cumplido, con creces, sus objetivos.

REFERENCIAS BIBLIOGRÁFICAS

COMPANY COMPANY, CONCEPCIÓN (dir.) (2006): *Sintaxis histórica de la lengua española. Primera parte: la frase verbal*, México: UNAM, Fondo de Cultura Económica, 2 vols.

— (dir.) (2009): *Sintaxis histórica de la lengua española. Segunda parte: La frase nominal*, México: UNAM, Fondo de Cultura Económica, 2 vols.

ROJO, GUILLERMO (2008): "Lingüística de corpus y lingüística del español", en *Actas del XV Congreso de la ALFAL (Montevideo, 18-21 de agosto de 2008)*. Versión electrónica.

— (2010): "Aguja de navegar corpus", en V. M. Castel y L. Cubo de Severino (eds.), *La renovación de la palabra en el bicentenario de la Argentina. Los colores de la mirada lingüística*, Mendoza: Editorial FFyL, UNCuyo, 1151-1163.

CRISTINA BUENAFUENTES DE LA MATA
Universidad Autónoma de Barcelona

MARTA FERNÁNDEZ ALCAIDE/ARACELI LÓPEZ SERENA (eds.): *400 años de la lengua del Quijote. Estudios de historiografía e historia de la lengua española. Actas del V Congreso Nacional de la Asociación de Jóvenes Investigadores de Historiografía e Historia de la Lengua Española (Sevilla, 31 de marzo, 1 y 2 de abril de 2005)*, Sevilla: Universidad de Sevilla, 2008.

En algunas disciplinas filológicas europeas se ha convertido en tradición que los jóvenes investigadores no solo presenten los resultados de sus investigaciones sino que también los publiquen. En el caso de la Romanística en Alemania destaca el *Forum Junge Romanistik* y en España la prolífica *Asociación de Jóvenes Investigadores de Historiografía e Historia de la Lengua Española*. Esta última, en favor de las jóvenes generaciones de investigadores, se ha propuesto divulgar la historia de la lengua, lo que, hasta ahora, algunos lingüistas estructuralistas consideraban desfasado.

El amplio tomo que presenta los resultados del *V Congreso Nacional de la Asociación de Jóvenes Investigadores de Historiografía e Historia de la Lengua*