Lingüística de corpus y lingüística histórica iberorrománica

Beihefte zur Zeitschrift für romanische Philologie

Herausgegeben von
Claudia Polzin-Haumann und Wolfgang Schweickard

Band 405

Lingüística de corpus y lingüística histórica iberorrománica

Editado por
Johannes Kabatek

Con la colaboración de
Carlota de Benito Moreno

DE GRUYTER

ISBN 978-3-11-046022-3
e-ISBN (PDF) 978-3-11-046235-7
e-ISBN (EPUB) 978-3-11-046050-6
ISSN 0084-5396

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Satz: jürgen ullrich typosatz, Nördlingen
Druck und Bindung: CPI books GmbH, Leck
♾ Gedruckt auf säurefreiem Papier
Printed in Germany

www.degruyter.com

Índice

III. Corpus y análisis cuantitativos

IV. Cuestiones lingüísticas diacrónicas iberorrománicas y lingüística de corpus

Johannes Kabatek

Un nuevo capítulo en la lingüística histórica iberorrománica: el trabajo crítico con los corpus. Introducción a este volumen

1 Introducción a la introducción

Mucho se ha escrito en los últimos años sobre la lingüística histórica y los corpus y mucho se ha trabajado en el ámbito de las lenguas iberorrománicas para mejorar tanto los corpus como los trabajos que se basan en ellos. El volumen que presentamos a continuación se enmarca en una nueva etapa de la lingüística de corpus, una etapa caracterizada por una visión crítica, tal vez menos entusiasta que hace veinte o treinta años, pero también más refinada y más adecuada a su objeto de estudio. Mientras que hace unos años la aparición de los primeros grandes corpus históricos de las lenguas iberorrománicas (sobre todo del español) fue recibida con general entusiasmo ante las posibilidades casi interminables de observar muy fácilmente fenómenos cuyo estudio antes exigía un arduo y dificultoso trabajo a mano, en la actualidad la disponibilidad masiva de datos y el fácil acceso a ellos se considera algo ya completamente normal y forma parte del día a día de investigadores y estudiantes. Al mismo tiempo, como es habitual en el avance de una disciplina, junto con las nuevas posibilidades aparecieron también nuevos problemas y surgieron nuevas tareas. Una mirada menos entusiasta, más sobria y más crítica ha creado nuevas exigencias, basadas en (a) el cuestionamiento de la relación entre datos primarios e historia de la lengua; (b) el cuestionamiento de los mismos datos primarios; (c) la crítica hacia el diseño de los corpus; (d) la crítica hacia las posibilidades ofrecidas por los corpus y los bancos de datos. Pero, como es natural, los investigadores no solo se han dedicado a la crítica, sino también al remedio. Gracias a ello, hoy en día ya disponemos de más y mejores corpus, de más y mejores herramientas para el tratamiento de los datos y, finalmente, de una serie de nuevos estándares más o menos establecidos en la comunidad, algunos de ellos presentados a lo largo de las páginas de este libro.

Johannes Kabatek: Universität Zürich

2 Lenguas iberorrománicas

Antes de entrar en el debate acerca de los cuatro puntos mencionados, me parece oportuno decir algo sobre el enfoque *iberorrománico* de este libro. Por un lado, las limitaciones areales y tipológicas son en cierta medida arbitrarias y se deben a circunstancias a veces casuales. Por otro lado, si frente a las visiones monolingües existe, con amplia tradición, una filología románica bien establecida que se justifica por el origen común de las lenguas neolatinas, no podemos decir lo mismo de las lenguas romances de la Península Ibérica. Hay, sin embargo, una serie importante de antecedentes, sobre todo en geografía lingüística (si pensamos por ejemplo en el ALPI) o en obras que relacionan la historia de la lengua con la historia de los espacios, como el famoso manual de Baldinger (1971). Aun así, es más común que los espacios investigados tomen como punto de partida las configuraciones políticas actuales y no las geográficas. Pero no hay que olvidar que la consideración de un determinado espacio histórico a partir de espacios nacionales actuales corresponde a la tan citada «teleología invertida» (Oesterreicher 2007), difícilmente justificable desde la perspectiva de la emergencia histórica y debida en gran parte a limitaciones derivadas de fronteras actuales y de posibilidades de financiación de proyectos, marcadas por un pensamiento territorial y político. Es fuera de los ámbitos políticos peninsulares —muchas veces por falta de recursos y de posibilidades de crear departamentos separados, pero también por una tradición que privilegia lo plural frente a lo monolítico y que es consciente del provecho de la comparación— donde la *iberorromanística* es ya un hecho establecido, y su tarea consiste precisamente en propagar el mensaje de que la comparación de lo semejante puede ser altamente provechosa.[1]

Los corpus históricos no escapan al efecto de esta «teleología invertida»: suelen estar hechos por instituciones nacionales o de un ámbito lingüístico actual determinado y construyen el pasado a partir del presente, ignorando por tanto el hecho de que los límites claros se van borrando conforme retrocedemos en el tiempo. Así pues, juntar aquí trabajos sobre diferentes lenguas iberorrománicas tiene una doble finalidad: por un lado, las razones están en los mismos objetos de

1 La base de la mayoría de las contribuciones a este volumen fueron los trabajos presentados en el marco del Tercer Coloquio Internacional sobre Corpus diacrónicos en lenguas iberorrománicas (CODILI III), celebrado en la Universidad de Zúrich en verano del 2014 (www.codili.ch). Algunas de las reflexiones aquí presentadas también se deben a las discusiones llevadas al cabo durante el curso de invierno ALPES (Abriendo Líneas en el Pasado del Español) en Kandersteg, Berna, en enero de 2016. Agradezco a los participantes de ambos encuentros (en parte coincidentes) sus valiosas contribuciones, y al Fondo Nacional Suizo y a la Confederación de las Universidades Suizas su generoso apoyo.

estudio, en los datos y fenómenos relacionados, y, por otro lado, 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.

3 Lingüística de corpus y lingüística con corpus

Resulta claro que el término «lingüística de corpus» hoy en día comprende disciplinas relativamente diferenciadas (véase Parodi 2010 y Torruella, en este volumen), en las que deberíamos distinguir al menos tres vertientes y finalidades: por un lado, la lingüística de corpus 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 (ver, p. ej. Rojo, con el caso del CORPES, o Xavier, con el del CIPM, ambos en este volumen). Una segunda disciplina es la que está en estrecha relación con la informática y se ocupa, por un lado, del tratamiento de los datos y de su etiquetaje, y por otro lado, de los análisis cuantitativos y estadísticos a los que invitan los propios datos del corpus. Finalmente, la tercera vertiente es la más presente en este volumen: la que hace una lingüística «con corpus», ocupándose de fenómenos concretos de la historia de la lengua y basando su análisis en datos de corpus. Entre las tres vertientes hay, obviamente, una estrecha relación y, en tanto que un corpus no es un fin en sí, sino que se crea *para algo*, se necesita un intercambio continuo del creador del corpus con los usuarios que lo utilizan para un estudio concreto. Entre ambos puede haber discrepancias, ya que sus objetivos y condicionamientos son diferentes: el usuario pide el mayor número de datos posible, a poder ser libremente disponibles y fielmente editados, etiquetados y con acceso a los textos plenos, mientras que el que configura el corpus lucha con limitaciones técnicas, recursos de tiempo y de personal, derechos de autores y, a veces, limitaciones institucionales. Afortunadamente, en el mundo de las lenguas iberorrománicas, los que crean corpus y los que trabajan con ellos generalmente no están separados ni viven en mundos diferentes; en varios de los trabajos de este volumen se ve que la creación del corpus y la investigación de los fenómenos está en las mismas manos.

4 Corpus, lengua, representatividad

Tal vez el tema más discutido en los últimos años —y también presente a lo largo de los trabajos de este volumen— sea el de la representatividad de los datos y la cuestión de la relación entre los corpus y la historia de la lengua. Mientras que en

los albores de la lingüística de corpus moderna todavía era frecuente encontrar un postulado abstracto de representatividad absoluta de un corpus para la historia de una lengua, hoy en día ya pocos creen que algo así pueda existir y se habla más bien de una representatividad relativa, es decir, una representatividad *con respecto a algo*. Aquí hay que introducir una serie de precisiones: en primer lugar, hay que tener cuidado con la frecuente equiparación entre corpus y lengua y hay que recordar siempre que la lengua no es un fenómeno comparable a los fenómenos de la naturaleza que se limitan, en general, a la evolución material de lo físicamente medible.

Los corpus son colecciones de *textos* que nos permiten tener una visión indirecta de la lengua, ya que la producción de textos a partir de la competencia lingüística de los individuos está condicionada por una serie de factores que el corpus no permite ver (factores pragmáticos, sociales, individuales). Por ello, los datos de los corpus no nos ofrecen la historia de la lengua como tal, sino que son datos que hay que interpretar con respecto a todos los factores de su producción, en el sentido de una *recontextualización* (Oesterreicher 2001). La recontextualización es una tarea interminable, no limitable a dos o tres factores; es una tarea *hermenéutica* y, por lo tanto, siempre abierta. El corpus contiene lengua, naturalmente, pero el corpus no contiene *la lengua*, ni como objeto abstracto, ni como objeto concreto y mental. El corpus se limita a una colección de producciones casuales de lengua: nos ofrece una ventana que permite acceder a una parte de esta, pero no al todo, y deja, por tanto, abierta la especulación acerca de lo que no se puede ver. Aun así, incluso lo invisible tiene que suponerse como existente y los principios de actualidad y de empatía[2] nos llevan a identificar a partir del corpus factores necesariamente existentes pero no aparentemente presentes: sabemos que la lengua no es un sistema homogéneo y unitario y que los textos no son, pese a lo que se haya creído en algún momento, producto de una competencia lingüística generadora de textos que se puede reconstruir directamente sobre la base de estos. Sabemos también que una lengua histórica no es un solo sistema, sino un diasistema complejo, y que cada hablar se determina diatópica, diastrática y diafásicamente. También sabemos que el hablar no solo corresponde a una sintopía, sinestratía y sinfasía, sino que también está inserto en tradiciones discursivas, en moldes repetitivos anclados en configuraciones pragmáticas identificables y semióticamente relevantes. Y sabemos, por último, que el hablar presenta rasgos

2 El llamado «principio de la actualidad» suele atribuirse hoy en día a Labov (1974) aunque fue ya muy claramente formulado por Osthoff y Brugmann (1878, IX–X) en su manifiesto neogramático. Es este un principio que deriva de la empatía que tenemos como hablantes con cualquier otro hablante y, a partir de ahí, con cualquier situación lingüística, presente o pasada (cf. Kabatek 2015).

individuales, tradiciones que un mismo individuo crea y cultiva y que lo distinguen frente a los demás: su «estilo» personal.[3]

¿Y con todas estas precisiones queremos hacer lingüística histórica? ¿Puede haber un corpus que nos permita distinguir tanta variación? ¿O es la lingüística histórica basada en corpus simplemente una ilusión, una reducción a pocos factores que nunca llegará a descubrir las dimensiones totales de su objeto? Me parece que la respuesta debe ser la de todo trabajo científico: no llegaremos nunca a una ciencia «total» o perfecta: nunca llegaremos a describir el objeto de nuestro estudio de manera completa, pero la utopía debe ser la de un paulatino acercamiento al objeto y una continua distinción entre lo que se aproxima más a él y lo que está más distante. En este sentido, llegamos también a lo que se ha venido a llamar «la paradoja de Enrique» («Enrique's paradox», cf. Enrique-Arias 2012, 96): «Una paradoja de la composición de los corpus diacrónicos es que, por una lado, deben ser heterogéneos (tienen que incluir textos de diferentes autores, épocas, géneros, registros, dialectos) y a la vez deben ser homogéneos (es decir, los diferentes cortes sincrónicos representados en el corpus tienen que ser comparables entre sí)».[4] La paradoja es solo aparente: para llegar a una descripción válida, es imprescindible que identifiquemos los factores de heterogeneidad. Solo a partir de esa identificación será posible garantizar que los factores heterogéneos estén lo suficientemente representados y que no estemos comparando peras con manzanas. Por ejemplo, solo sabiendo cuál es el papel de las tradiciones textuales en un caso concreto podemos averiguar si un cambio observado es un cambio de la lengua o solo una particularidad de una tradición textual particular. Tenemos, pues, que vivir con lo que parece ser una paradoja: en ella reside, en realidad, la tensión de nuestro trabajo de reconstrucción histórica de los fenómenos.

---

3 Para dar cuenta de los hechos individuales, Mario Barra ha insistido últimamente en lo que ha denominado el «método idiolectal» (Barra 2015; ver también Iglesias, en este volumen), consistente en el estudio de la historia de la lengua basado en «gramáticas individuales». Aunque me parece problemática la noción de «gramática individual», medir el espectro de posibilidades gramaticales de las que dispone un individuo sí resulta un acercamiento muy interesante que habría, en todo caso, que relacionar con su interacción con variedades y tradiciones discursivas para la reconstrucción de lo que al final llamaremos diacronía.

4 Se desarrolla este principio en Rosemeyer/Enrique-Arias (en prensa): «Longitudinal analyses of syntactic change, however, need language examples that differ with regard to the state of development of the language rather than their usage contexts. This methodological challenge has been formulated in terms of a *comparability paradox* in historical corpus design (Enrique-Arias 2012, 97): a historical corpus has to be *diverse* because it must contain texts that represent different periods, genres or dialects. At the same time this corpus must be *uniform* (that is, the distribution of content type, genres or dialects along the different chronological sections in the corpus must be as similar as possible so they can be compared).»

5 Nuevos estándares

5.1 La base: los documentos y las ediciones

Mientras que la primera fase de la lingüística moderna de corpus históricos estaba basada en una tradición que venía de las ediciones tradicionales en papel, en la actualidad en muchos casos los documentos se preparan ya con vistas a su presentación en un corpus digitalizado. Esto cambia radicalmente la concepción del corpus y abre toda una serie de posibilidades nuevas. Especialmente en el caso de los textos medievales, la edición en papel suponía siempre una decisión por parte del editor entre fidelidad paleográfica, enmienda e intervención y los corpus diacrónicos se basaban en ediciones de diversa índole.[5] Al introducir los textos en el corpus, hubo generalmente que prescindir del aparato de notas y de las variantes: así, lo que aparecía en la versión digitalizada solía ser el resultado del escaneo de ediciones publicadas que habían pasado por un proceso OCR y una corrección manual más o menos meticulosa, dependiendo del caso. Esta técnica sigue contribuyendo a la ampliación de la cantidad de datos históricos disponibles, aunque, obviamente, ha mejorado sustancialmente y ahora nos encontramos a leguas de los problemas que presentaba el reconocimiento automatizado de texto hace no tantos años. Hoy en día, un sencillo programa de reconocimiento que podamos manejar en nuestro ordenador da mejores resultados que las técnicas más sofisticadas de hace no muchos años, y la microtomografía está empezando a permitir incluso la lectura de documentos sin abrirlos. Aun así, casos debidos a errores de reconocimiento, como el muy citado de *mafia* en CORDE, siguen estando presentes en algunos corpus.[6]

Sin embargo, existen también otras posibilidades y, en el mundo de la lingüística iberorrománica histórica se puede decir que el estándar que encuentra

5 En el caso del español, un caso excepcional es el ya histórico ADMYTE, nacido en circunstancias particularmente afortunadas y que desde el inicio trabajó con ediciones hechas para su integración en el corpus, véase Marcos Marín 1993.

6 El italianismo *mafia* parece a primera vista ser muy temprano en español según el CORDE (ya en el s. XVI hay ejemplos como «con mafia y trato de algunos de sus contrarios», de 1579), pero su existencia se debe a malas lecturas del escaneo de *maña*. Aunque el caso es conocido sigue presente en CORDE. También hay una serie de casos desde el siglo XVI en el *Corpus del español* de Mark Davies. Sin embargo, en el *Corpus del Nuevo Diccionario Histórico* de la RAE, que incorpora los textos del CORDE, el ejemplo citado está corregido por *maña*. Se ve aquí que la nueva generación de los corpus académicos no solo da un salto con respecto a las herramientas técnicas sino también con respecto a la calidad de los datos.

cada vez más aplicación hoy en día es el establecido por la red CHARTA,[7] según el cual el corpus no se limita a una edición cualquiera, sino que presenta una «edición múltiple», con la versión paleográfica al lado de una edición crítica y con acceso a la imagen de manuscrito, que permite comprobar la fiabilidad de ambas versiones. Varios de los trabajos aquí presentes trabajan con datos de CHARTA o de otros corpus relacionados con esta red, como CODEA (cf. Marcet Rodríguez & Sánchez González de Herrero; Moral del Hoyo, en este volumen).[8]

5.2 La mirada crítica de los corpus existentes: el «CORDEmáforo»

Como indicábamos más arriba, otro aspecto esencial de la nueva lingüística de corpus es la mirada crítica hacia herramientas establecidas. En el mundo hispánico, no cabe duda de que el corpus histórico más establecido es el CORDE de la RAE,[9] plataforma imprescindible para los estudios de la historia del español. Es innegable que el CORDE permitió una enorme ampliación de la base de datos históricos disponibles y, pese a cualquier posible crítica de detalles, es una empresa que ha dado un enorme rendimiento. En los últimos años se ha observado que el CORDE, además de las limitaciones técnicas del banco de datos, presenta una serie de fuentes de posibles errores, las cuales, sin embargo, no son

7 Véanse los criterios de edición en http://www.charta.es/criterios-de-edicion-/ Cf. también Sánchez-Prieto Borja/Torrens Álvarez (2012).

8 Otra de las innovaciones de los últimos años consiste en la llamada «edición social», en la que varias personas colaboran según el principio *wiki* (Price 2016).

9 Desconozco si también es el más utilizado, ya que carecemos de datos sobre la utilización de los corpus. Probablemente el corpus histórico español más usado sea el *Corpus del español* de Mark Davies. Se trata de un corpus que, sobre todo en sus inicios, tuvo un impacto importante, dada la enorme rapidez de su sistema de búsqueda. En una segunda fase, llamó la atención por la presentación parcialmente etiquetada de los datos y por la integración de un módulo muy útil de visualización y es usado bastante hasta la actualidad (véase Nieuwenhuijsen, en este volumen). Sin embargo, es también un corpus muy controvertido, en primer lugar por la falta de fiabilidad filológica de parte de los datos y los criterios algo arbitrarios de su configuración. Además, desde hace algún tiempo el corpus se presenta también con fines comerciales. En la actualidad, Mark Davies anuncia en su página una nueva versión tanto del *Corpus del español* como del *Corpus do português* (elaborado en colaboración con Michael Ferreira), modernizada y ampliada: el corpus del español tendrá 2.000 millones de palabras y el del portugués, 1.000 millones. Ambos se pondrán a disposición del público en 2016. Estos corpus tendrán información POS y anotación sintáctica y permitirán también el acceso a los textos planos. El aumento de la cantidad de textos se refiere sobre todo a la época moderna. Véase la información en http://corpus.byu.edu/neh2015.asp.

razón para el rechazo del CORDE como fuente, sino que exigen una utilización crítica del corpus. Por ejemplo, como acabamos de señalar, existen casos de erratas que se deben a errores de pasaje de los datos primarios y sería de agradecer que hubiese un mecanismo de corrección continua del corpus.

Otra cuestión que se ha señalado (cf. Octavio de Toledo, en este volumen) es la del desequilibrio textual: la cantidad de los textos varía considerablemente en las diferentes épocas y también varía, lógicamente, la gama de tradiciones discursivas disponible de cada época. No obstante, el mayor problema del CORDE tal vez sea, por lo menos para las épocas remotas, el de las fechas de los documentos, que es en realidad un problema no exclusivo del CORDE, sino de la lingüística histórica como tal. Una práctica bastante general en la tradición de la disciplina (y no solo en el mundo iberorrománico) solía ser suponer que la fecha de supuesta o comprobada composición de una obra era la relevante, proporcionándose solo esta, sin importar que el texto manejado procediera de copias o de ediciones posteriores. La RAE, poniendo a disposición del público el Corpus del Nuevo Diccionario Histórico del Español (CDH / CNDHE), ha puesto remedio a ese defecto, indicando entre corchetes la información sobre la fecha del «testimonio base», es decir, del manuscrito, frente a la supuesta fecha de composición del texto «original». Sin embargo, sigue siendo frecuente encontrar en trabajos de historia de la lengua un texto como el *Calila e Dimna*, por poner un ejemplo, como representante del siglo XIII, aunque sabemos que los dos manuscritos en los que se basan las ediciones son del siglo XV y que el lapso de dos siglos no se produjo sin dejar huellas en el texto. Hay suficientes estudios de originales y copias (cf. p. ej. Morala 2002; Santiago 2004; Díez de Revenga 2012; Miguel Franco 2012) en diferentes ámbitos textuales como para poder afirmar que la idea tradicional de que en el acto de copia del texto se preserva lo fundamental (o que, como mucho, se cambian algunas grafías) carece de fundamento empírico. Para poner remedio a ese problema, Octavio de Toledo / Rodríguez Molina (en prensa) han preparado una lista de los documentos contenidos en el CORDE en la que se evalúa la calidad de estos para los estudios diacrónicos, llegando a establecer una tripartición entre documentos perfectamente válidos y bien fechados (luz verde), documentos algo problemáticos (ámbar) y documentos muy problemáticos (rojo) —en los que la fecha de composición no coincide con la fecha del documento utilizada en el corpus—. Tal «Cordemáforo» permitirá, pues, limitar los estudios a los documentos fiables o, incluso, comparar un estudio que no aplique el filtro de calidad de documento con otro que sí lo tome en consideración, lo que seguramente ofrecerá resultados sorprendentes en algunos casos. Evidentemente, la diferencia entre las tres categorías no es tajante, sino relativa, pero permite en todo caso establecer «jerarquías de fiabilidad» de los textos: un original siempre es más fiable que una copia, un fenómeno basado en varios testimonios es siempre más fiable que un hápax, etc.

5.3 Nuevos corpus, nuevas herramientas

Más allá de los corpus grandes de generaciones anteriores, en el presente estamos asistiendo a tres tendencias en cuanto a la configuración de los corpus históricos: primero, hay una nueva generación de grandes corpus históricos que, desde el punto de vista técnico, superan ampliamente a los corpus anteriores; segundo, asistimos a una masificación de los datos disponibles en los corpus, sobre todo de la lengua actual, que permite la elaboración de estudios microdiacrónicos y la observación de las tendencias actuales en la evolución de la lengua,[10] y, por último, están apareciendo cada vez más corpus especializados, ya sean regionales o con finalidades particulares. Al mismo tiempo, los trabajos de corpus permiten, dada la gran amplitud de la base de datos, incluir más factores, que pueden tanto derivar de variables propiamente gramaticales como tener un carácter más bien «externo», como la importancia de una distinción clara de las distintas variedades o tradiciones discursivas, algo que, particularmente en los estudios de las lenguas iberorrománicas, se ha hecho prácticamente general. La mayor cantidad de datos disponibles abre también nuevas vías para los análisis cuantitativos: el «giro cuantitativo» se hace notar también en la lingüística histórica iberorrománica, sin que por ello se pierda de vista la base filológica de los textos.

5.4 Nuevos datos, nuevos factores, nuevas posibilidades

Si intentamos resumir las tendencias predominantes en la lingüística iberorrománica histórica basada en corpus tal como se presenta ahora en comparación con las épocas anteriores (cf. p. ej. Pusch/Kabatek/Raible 2005), vemos una particularidad interesante: si con la llegada de las nuevas técnicas de búsqueda de datos algunos pensaban que la lingüística histórica iba a ser más sencilla y más fácil, la realidad ha demostrado lo contrario. Así, los problemas tradicionales de reconstrucción siguen siendo los mismos y el acceso a más datos ha causado nuevos desafíos. Las cuestiones de la frecuencia, de la estadística y de la ponderación de datos se han planteado de forma nueva y, al mismo tiempo, nuevos factores se han añadido a la lista larga de posibles condicionantes del cambio lingüístico: la teoría del cambio lingüístico ha ido identificando, en las últimas décadas, un número creciente de factores sintácticos, semánticos, fónicos y pragmáticos que pueden condicionar los cambios y, dependiendo del fenómeno estudiado, la lista

10 Rojo (en este volumen) menciona, al lado del CORPES XXI, el Gigacorpus esTenTen, el mayor corpus del español disponible actualmente.

puede ser larga (cf. p. ej. Bouzouita, o Schulte/Blas Arroyo, en este volumen). A los factores lingüísticos se añaden factores extralingüísticos (históricos, sociales, culturales). Así, al considerar las dimensiones de variación arriba mencionadas, además de la tradicionalidad discursiva de los fenómenos y la posible individualidad de su uso concreto, puede parecer que los árboles son tan numerosos y diversos que ya no hay bosque visible. Esto no es así, sin embargo: una lingüística histórica con una base de datos fiables más amplia es precisamente la que produce los análisis más complejos y completos de las evoluciones y permite que nos acerquemos más a la reconstrucción adecuada del cambio.

Por otro lado, resulta evidente que no todos los factores tienen el mismo peso en cada cuestión empírica concreta y que la tarea del lingüista no consiste únicamente en la recolección de datos y la enumeración de factores, sino en su ponderación e interpretación. Nos hallamos, pues, en una fase de la lingüística histórica en la que hay más complejidad, más datos y más factores de lo que solía haber, pero también nuevas posibilidades de ordenar los datos y de presentarlos de forma que nos ofrezcan una imagen cada vez más acertada de lo ocurrido en la historia de las lenguas.

6 Los trabajos de este volumen

Los 18 trabajos reunidos en este volumen se inscriben en esta nueva generación de la lingüística histórica basada en corpus. Hemos dividido los trabajos en cuatro apartados, sin que la repartición corresponda a una separación tajante. La primera sección contiene trabajos cuyo objetivo consiste en reflexionar, a partir de casos y cuestiones empíricas concretas, acerca de problemas generales de la lingüística de corpus. La segunda sección se dedica a la presentación de corpus; la tercera, a los análisis cuantitativos y la cuarta se ocupa de análisis diversos (cualitativos, variacionales, idiolectales) basados en trabajos con corpus.

El libro se abre con las reflexiones de Andrés Enrique-Arias acerca de lo que el autor llama el «parámetro perspectiva». Se trata de «la perspectiva de sus usuarios [los del corpus], es decir, la manera en que los estudiosos acceden a los datos lingüísticos». Con el ejemplo del corpus *Biblia medieval*, Enrique-Arias muestra las ventajas de los corpus paralelos, enumerando toda una serie de factores que conforman el valor heurístico añadido de estos: mientras que habitualmente un corpus solo nos permite encontrar aquello que buscamos explícitamente (según el procedimiento semasiológico de la búsqueda), en un corpus paralelo de textos traducidos, dado que la traducción pasa por una fase semasiológica y otra onomasiológica, encontramos también soluciones inesperadas para el mismo contenido

o un contenido semejante. Claro está que los corpus paralelos no son de por sí mejores que otros corpus, pero permiten otro tipo de acercamiento a la materia y complementan en el estudio diacrónico a los corpus que el autor llama «convencionales».

En la misma línea, Santiago del Rey Quesada también subraya la utilidad de los corpus paralelos: su aportación es una apología de los estudios de traducción basados en corpus (*Corpus-based Translation Studies* o CTS) para el estudio de la historia de la lengua. Sus reflexiones se basan en un corpus paralelo de los *Colloquia* de Erasmo de Rotterdam y desarrollan temas como la relevancia de la lengua de origen, las tradiciones discursivas y el estilo personal en las traducciones. El autor postula la necesidad de disponer de más corpus paralelos con textos traducidos para poder medir el impacto de la traducción en comparación con las producciones originales de una lengua en diferentes épocas.

La contribución de Álvaro Octavio de Toledo y Huerta tiene, por un lado, una finalidad práctica y ejemplar, a saber, la de mostrar cómo se puede sacar provecho del CORDE como herramienta para el estudio del «primer español moderno» —definido por él como el español del periodo que comprende desde finales del siglo XVII hasta principios del XIX—. Así, el autor insiste en la importancia de esa época para el estudio de la historia del español, a pesar de la tendencia de la lingüística histórica de prestarle poca atención. Pero, más allá de la finalidad empírica (demostrada con una serie de ejemplos), el trabajo insiste en la necesidad de la *ponderación* de los datos y de la preparación equilibrada de lo que en el corpus se encuentra de una forma más bien desequilibrada, e identifica diferentes tipos de «difusión de los fenómenos y su dinámica variacional». Estas reflexiones van mucho más allá del periodo estudiado y muestran retos importantes para la lingüística diacrónica basada en trabajos con corpus.

También son de índole general las reflexiones de Joan Torruella, que, aunque versan acerca de una serie de propuestas concretas de mejora del *Corpus Informatitzat del Català Antic* (CICA), tratan también de la cuestión general de la representatividad del corpus y del equilibrio de los datos contenidos en él. Torruella se refiere a la cuestión de los cortes diacrónicos en un corpus (para lo que propone cortes de 50 años), el equilibrio textual (en una línea parecida a de Octavio de Toledo) y la comparabilidad de diferentes secciones de corpus, intentando ofrecer soluciones a la paradoja entre homogeneidad y heterogeneidad arriba mencionada. Además, el autor discute la pertinencia de diversos parámetros lexicométricos, diseñados con el fin de que el corpus represente, en la medida de lo posible, la mayor riqueza léxica posible de la lengua estudiada.

Abre la segunda sección, dedicada a la presentación de nuevos corpus o de proyectos de corpus, el trabajo de Rosario Álvarez Blanco y Ernesto González

Seoane, quienes presentan el corpus gallego *Gondomar*. Como es sabido, el gallego, después de una primera fase de producción escrita en la Edad Media (documentada en los corpus TMILG y COTAGAL), pasa a través de la época que se suele denominar los *séculos escuros* ('siglos oscuros'), con escasa producción escrita hasta el llamado *rexurdimento* del siglo XVIII. GONDOMAR recoge todo tipo de testimonios de esa época, clasificados según los parámetros que imponen los propios textos e incluyendo parodias del gallego y textos gallegos en el contexto del castellano, arrojando así algo de luz sobre esa época y permitiendo crear un eslabón entre la época medieval y la contemporánea.

Por su parte, Maria Francisca Xavier dedica su contribución a la presentación de la historia y las posibilidades del CIPM, *Corpus Informatizado do Português Medieval*, de la Universidade Nova de Lisboa. Es este un corpus diseñado ya en los años 1990 y ampliado y completado desde entonces. En este corpus, como en otros (por ejemplo, el CDH para el español), existe un vínculo directo entre corpus y lexicografía, ya que el CIPM sirve como base para el *Dicionário do Português Medieval*, un diccionario modular (con partes dedicadas a los verbos, los nombres propios y comunes y los términos) que viene publicándose desde 1999.

Otro corpus medieval, esta vez de ámbito regional, es presentado por Vicente Marcet Rodríguez y M.ª Nieves Sánchez González de Herrero. Se trata de un proyecto reciente, lanzado hace solo unos años, de recogida de la documentación de la zona —de gran importancia para la historia del castellano— de Miranda de Ebro, en el norte de Burgos. El corpus está formado por un total de 203 documentos de dos archivos, elaborados 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 corpus *Post Scriptum*, de la Universidade de Lisboa, es presentado por Catarina Carvalheiro, Ana Luísa Costa, Rita Marquilhas, Clara Pinto, Fernanda Pratas y Gael Vaamonde, y recoge datos tanto del portugués como del español moderno: cartas privadas de ambos lados del atlántico, conservadas en la documentación oficial de los tribunales españoles y portugueses. Los autores muestran la utilidad de este corpus para estudiar la geografía y la diacronía de ciertos fenómenos lingüísticos mediante tres ejemplos: el marcador discursivo *pois* y el relativo *cujo* en portugués así como la cuestión de *leísmo*, *laísmo* y *loísmo* en español.

Frente a los corpus que se dedican a la documentación de épocas remotas, el trabajo de Guillermo Rojo traza la línea entre el CREA y el CORPES XXI, los dos corpus del español actual de la RAE. El autor, responsable de los proyectos de corpus en la Academia, no solo presenta el último de estos corpus, sino que trata también cuestiones generales de gran importancia, como la representatividad o el salto cuantitativo y cualitativo que hay entre la versión original del CREA (lan-

zada en 1998) y la nueva plataforma del CORPES XXI: la nueva generación de los corpus de la RAE permite búsquedas mucho más sofisticadas; visualizaciones de diferentes épocas y de diferencias regionales; búsquedas por formas, lemas y categorías gramaticales. Es fundamental resaltar que la interfaz del CORPES XXI no se ha creado únicamente para este corpus, sino que también se ha usado para la modernización de los corpus anteriores: así, la nueva versión de CREA presenta los textos hasta 2000 con la misma anotación que el CORPES XXI (que enlaza con el CREA a partir de 2000) y el CDH presenta los textos de CORDE con las nuevas herramientas de búsqueda. Por último, es de notar que el CORPES XXI, puesto que se presenta en diferentes secciones de cinco años cada una, ofrece también el acceso inmediato a la microdiacronía y el «change in progress».

En la tercera sección se discuten cuestiones cuantitativas y se presentan análisis frecuenciales de diferentes fenómenos. Dorien 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 a la luz de diversas variables (tipo de interrogación, modo verbal, tiempo, región), llegando así a dar una imagen de la evolución del fenómeno a ambos lados del atlántico.

El trabajo de Kim Schulte y José Luis Blas Arroyo se dedica a la evolución de cinco perífrasis modales. Los autores trabajan con un amplio corpus propio de textos de «inmediatez comunicativa» (Koch/Oesterreicher 2007) del siglo XVI al XX y efectúan un análisis multifactorial y frecuencial que les permite identificar correlaciones estadísticamente relevantes.

Miriam Bouzouita, por su parte, estudia un fenómeno concreto en el corpus *Biblia medieval*, interesándose por los futuros y condicionales sintéticos medievales seguidos de pronombre. 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.

Siguiendo con los análisis frecuenciales, pero en un ámbito muy distinto, el trabajo de María Jesús Torrens Álvarez y de Hiroto Ueda se ocupa de la grafía <j> cuando esta tiene valor consonántico. El análisis estadístico con el programa LETRAS (diseñado por el propio Ueda) permite trazar la línea del «nacimiento», en el corpus CORHEN, de dicha letra, que, a partir de ciertas variantes gráficas de la <i>, se especializa en la representación de la consonante. Este trabajo no muestra solo la importancia de la estadística, sino también de la paleografía como base de datos fiables, fundamento imprescindible para el análisis cuantitativo.

También analiza datos del CORHEN el trabajo de Carmen Moral del Hoyo: sobre la base de una selección de 278 documentos procedentes de este corpus y

elegidos según criterios diatópicos y diacrónicos, la autora muestra convincentemente cómo la variación y la evolución de las formas *–ié* / *–ía* de imperfecto evoluciona en una interacción entre factores espaciales y factores estructurales.

La productividad léxica es el tema del artículo de Inés Carrasco Cantos y de Livia Cristina García Aguiar, que está dedicado al análisis del total de los sufijos contenidos en el corpus DITECA (*Diccionario de textos concejiles de Andalucía*), un corpus de textos jurídicos del siglo XIII al siglo XVIII. El 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.

En la cuarta sección encontramos diferentes cuestiones de la lingüística diacrónica iberorrománica, planteadas a partir de los datos de corpus. Del español en el siglo XVI se ocupan tanto el trabajo de Beatriz Arias Álvarez y Juan Antonio Hernández Mendoza como el de Marta Fernández Alcaide: el primero, del español de Nueva España, y el segundo, de la comunicación entre el Nuevo Mundo y España. Así, Arias Álvarez y Hernández Mendoza presentan el *Corpus Electrónico del Español Colonial Mexicano* (COREECOM) y muestran cómo, a partir de los datos de este corpus, puede estudiarse la variación y evolución de ciertos fenómenos. Fernández Alcaide, en cambio, combina un pormenorizado análisis textual con algunas observaciones de índole general, que destacan la importancia del acceso a información sobre las primeras décadas de la colonia —infrarrepresentadas en los grandes corpus—, acceso solo recientemente alcanzable gracias a la recuperación de textos en proyectos como CHARTA o CORDIAM.

Por último, el citado método idiolectal es aplicado por Olivier Iglesias para investigar la evolución de la subida de clíticos —es decir, la variación entre *lo puedo decir* y *puedo decirlo*— 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, pero no lo que se había supuesto para la evolución posterior, dando la razón, por tanto, a lo dicho arriba (ver nota 3): el método idiolectal presenta nuevos retos y complementa los resultados obtenidos a partir de otros enfoques.

En suma, el panorama presentado en este libro es amplio y permite ver algunas de las principales áreas de los estudios que, con datos de corpus, intentan reconstruir la historia de las lenguas iberorrománicas. Estos trabajos dan muestra de una fase nueva de la lingüística histórica, una fase en la que se plantean nuevos retos, pero en la que, al mismo tiempo, se perfilan nuevas soluciones.

7 Lista de corpus y bancos de datos citados

ADMYTE – *Archivo Digital de Manuscritos y Textos Españoles*, http://www.admyte.com.
ALPI – *Atlas Lingüístico de la Península Ibérica*, http://westernlinguistics.ca/alpi/more_info.php?global_lang=sp.
Biblia medieval – Andrés Enrique-Arias, *Corpus Biblia medieval*, http://www.bibliamedieval.es.
CHARTA – *Corpus Hispánico y Americano en la Red: Textos Antiguos*, http://www.charta.es/.
CICA – Joan Torruella, *Corpus Informatitzat del Català Antic*, http://cica.cat.
CIPM – *Corpus Informatizado do Português Medieval*, http://cipm.fcsh.unl.pt.
CODEA – *Corpus de Documentos Españoles anteriores a 1700*, http://demos.bitext.com/codea/.
CODEA+2015 – *Corpus de Documentos Españoles Anteriores a 1800*, http://textoshispanicos.es.
CODEMA – *Corpus diacrónico de documentación malagueña*, http://www.corpuscharta.es/grupos.html.
CORDE – Real Academia Española, *Corpus Diacrónico del Español*, http://corpus.rae.es/cordenet.html.
CORDEREGRA – *Corpus diacrónico del español del reino de Granada (1492–1833)*, http://www.corpuscharta.es/grupos.html.
CORDIAM – Virginia Bertolotti / Concepción Company, *Corpus Diacrónico y Diatópico del Español de América*, http://www.cordiam.org.
COREECOM – *Corpus electrónico del español colonial mexicano*, http://www.iifl.unam.mx/coreecom/.
CORHEN – *Corpus Histórico del Español Norteño*, http://www.corpuscharta.es/grupos.html.
CORPES XXI – Real Academia Española, *Corpus del español del siglo XXI*, http://web.frl.es/CORPES/view/inicioExterno.view.
CORPUSDELESPANOL – Mark Davies, *Corpus del Español*, http://www.corpusdelespanol.org.
CORPUSDOPORTUGUES – Mark Davies/Michael Ferreira, *Corpus do português*, http://www.corpusdoportugues.org.
COSER – Inés Fernández-Ordóñez, *Corpus Oral y Sonoro del Español Rural*, http://www.lllf.uam.es/coser/index.php.
COTAGAL – *Corpus de Textos Antiguos de Galicia*, http://www.corpuscharta.es/grupos.html.
CDH – Real Academia Española, *Corpus del Nuevo diccionario histórico del español*, http://web.frl.es/CNDHE/view/inicioExterno.view.
CREA – Real Academia Española, *Corpus de referencia del español actual*, http://corpus.rae.es/creanet.htm.
CREA (anotado) – Real Academia Española, *Corpus de Referencia del Español Actual (CREA. Versión anotada)*, http://web.frl.es/CREA/view/inicioExterno.view.

DITECA – *Diccionario de Textos Concejiles de Andalucía*, http://www.arinta.uma.es.
esTenTen – Sketch Engine, *Spanish Gigacorpus*, https://www.sketchengine.co.uk
GONDOMAR – *Corpus dixital de textos galegos da Idade Moderna*, http://ilg.usc.es/gl/proxectos.
IMPACT-es – *Diachronic corpus of historical Spanish*, http://www.digitisation.eu/tools-resources/language-resources/impact-es/.
P.S. – Post Scriptum – *Arquivo digital da escrita quotidiana em Portugal e Espanha na época moderna*, http://www.clul.ul.pt/pt/recursos/462-post-scriptum-home.
TMILG – *Tesouro Medieval Informatizado da Lingua Galega*, https://ilg.usc.es/tmilg/.

8 Referencias bibliográficas

Baldinger, Kurt, *La formación de los dominios lingüísticos en la Península Ibérica*, trad. de E. Lledó y M. Macau, 2a. ed. corr. y aum., Madrid, Gredos, 1971.
Barra Jover, Mario, *Método y teoría del cambio lingüístico: argumentos en favor de un «método idiolectal»*, in: García Martín, José María (dir.), *Actas del IX Congreso Internacional de Historia de la Lengua Española (Cádiz 2012)*, Madrid, Iberoamericana/Vervuert, 2015, 263–292.
Díez de Revenga, Pilar, *La tradición textual en la Edad Media: una muestra de los siglos XIII y XIV*, in: Torrens Álvarez, María Jesús/Sánchez-Prieto Borja, Pedro (edd.), *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*, Bern et al., Lang, 2012, 47–58.
Enrique-Arias, Andrés, *Dos problemas en el uso de corpus diacrónicos del español: perspectiva y comparabilidad*, Scriptum Digital 1 (2012), 85–106.
Kabatek, Johannes, *¿Es posible una lingüística histórica basada en un corpus representativo?*, Iberoromania 77 (2013), 8–28.
Kabatek, Johannes, *Lingüística empática*, Rilce 30–3 (2014), 705–723.
Kabatek, Johannes, reseña de Torrens Álvarez/Sánchez-Prieto Borja, *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*, Bern et al., Lang, 2012, Romanische Forschungen 128 (2016), 243–248.
Koch, Peter/Oesterreicher, Wulf, *Lengua hablada en la Romania: francés, italiano, español*, trad. de Araceli López Serena, Madrid, Gredos, 2007.
Labov, William, *The use of the present to explain the past*, in: Heilmann, L. (ed.), *Proceedings of the 11th International Congress of Linguistics*, Bologna, il Mulino, 1975, 825–851.
Marcos Marín, Francisco, *La biblioteca electrónica en el Archivo Digital de Manuscritos y Textos Españoles*, Lexis XVII, (1993), 33–56.
Miguel Franco, Ruth, *Documentos originales y cartularios del archivo de la Catedral de Toledo: propuestas para un estudio comparativo*, in: Torrens Álvarez, María Jesús/Sánchez-Prieto Borja, Pedro (edd.), *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*, Bern et al., Lang, 2012, 197–218.
Morala, José Ramón, *«Originales y copias», El proceso de castellanización en el área leonesa,* in: María Teresa Echenique Elizondo/Juan Sánchez Méndez (edd.), *Actas del Quinto Congreso*

Internacional de Historia de la Lengua Española (Valencia, 31.1.–4.2. 2000), vol. 1, Madrid, Gredos, 2002, 1335–1345.

Octavio de Toledo y Huerta, Álvaro/Rodríguez Molina, Javier, *La imprescindible distinción entre texto y testimonio: el CORDE y los criterios de fiabilidad lingüística*, Scriptum Digital 5 (2016) (en prensa).

Oesterreicher, Wulf, *La «recontextualización» de los géneros medievales como tarea hermenéutica*, in: Jacob, Daniel/Kabatek, Johannes (edd.), *Lengua medieval y tradiciones discursivas en la Península Ibérica. Descripción gramatical – pragmática histórica – metodología*, Frankfurt am Main/Madrid, Vervuert/Iberoamericana, 2001, 199–232.

Oesterreicher, Wulf, *Mit Clio im Gespräch. Zu Anfang, Entwicklung und Stand der romanistischen Sprachgeschichtsschreibung*, in: Hafner, Jochen/Oesterreicher, Wulf (edd.), *Mit Clio im Gespräch. Romanische Sprachgeschichten und Sprachgeschichtsschreibung*, Tübingen, Narr, 2007, 1–35.

Osthoff, Hermann/Brugmann, Karl, *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*, Leipzig, Hirzel, 1878.

Parodi, Giovanni, *Lingüística de Corpus: de la teoría a la empiria*, Frankfurt/Madrid, Iberoamericana, 2010.

Price, Kenneth M., *Social Scholarly Editing*, in: Schreibman, Susan/Siemens, Ray/Unsworth, John, *A New Companion to Digital Humanities*, New York, Wiley, 2016, 137–149.

Pusch, Claus D./Kabatek, Johannes/Raible, Wolfgang (edd.), *Romance Corpus Linguistics II. Corpora and Diachronic Linguistics*, Tübingen, Gunter Narr, 2005,

Rosemeyer, Malte/Enrique-Arias, Andrés, *A match made in heaven. Using parallel corpora and multinomial logistic regression to analyze the expression of possession in Old Spanish*, Language Variation and Change, 2016 (en prensa).

Santiago Lacuesta, Ramón, *Originales y copias en la documentación del monasterio de Sahagún*, in: *Orígenes de las lenguas romances en el Reino de León. Siglos IX–XII*, León, Archivo Histórico Diocesano, 2004, 533–563.

Torrens Álvarez, María Jesús/Sánchez-Prieto Borja, Pedro (edd.), *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*, Bern et al., Lang, 2012.