Historia del léxico español y Humanidades digitales

STUDIA ROMANICA ET LINGUISTICA

condita a Peter Wunderli et Hans-Martin Gauger
curant Daniel Jacob, Elmar Schafroth, Edeltraud Werner,
Araceli López Serena, André Thibault et Manuela Caterina Moroni

SRL 53

PETER LANG

Dolores Corbella – Alejandro Fajardo –
Jutta Langenbacher-Liebgott (eds.)

Historia del léxico español y Humanidades digitales

PETER LANG

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available online at http://dnb.d-nb.de.

Proyecto FFI2016-76154-P (Ministerio de Economía y Competitividad. Gobierno de España)

Cover Design: © Olaf Gloeckler, Atelier Platen, Friedberg

Printed by CPI books GmbH, Leck

ISBN 978-3-631-75800-7 (Print)
E-ISBN 978-3-631-76280-6 (E-PDF)
E-ISBN 978-3-631-76281-3 (EPUB)
E-ISBN 978-3-631-76282-0 (MOBI)
DOI 10.3726/b14447

Peter Lang – Berlin · Bern · Bruxelles · New York ·
Oxford · Warszawa · Wien

This publication has been peer reviewed.

www.peterlang.com

Prólogo de los editores

El desarrollo tecnológico e informático ha producido en los últimos años recursos que, además de proporcionar una gran cantidad de información, están transformando los métodos de investigación y la difusión de sus resultados. Este avance se puede realizar gracias a una ingente cantidad de datos que nunca antes habían estado disponibles y también por las mejoras en la capacidad de procesarlos. El diseño de los nuevos recursos y la generación de conocimiento a partir de los estudios que hacen posible se engloban, de una manera general, bajo la denominación de *Humanidades digitales*.

El objetivo principal de este volumen es mostrar cómo están cambiando radicalmente los métodos de trabajo en historia del léxico español gracias a los enfoques más recientes de las Humanidades; con esta finalidad, damos a conocer de una manera muy concreta cuáles son las principales investigaciones que se están llevando a cabo. No cabe duda de que el desarrollo de recursos innovadores, tanto en España como en distintos lugares de América, está poniendo las bases para avances sin precedentes en el conocimiento de la evolución del léxico, en su análisis y en su descripción lexicográfica, que en los próximos años marcarán la ruta de estas disciplinas lingüísticas. Los estudios léxicos, una de cuyas dificultades principales hasta ahora era el manejo de una gran cantidad de unidades —cambiantes además a lo largo de la historia—, serán, sin duda, una de las ramas filológicas más beneficiadas por el procesamiento digital.

Numerosas fuentes archivísticas, hasta ahora de difícil acceso fuera del ámbito local, comienzan a estar a disposición de los investigadores en cualquier parte del mundo; la consulta de los fondos de las hemerotecas y bibliotecas digitales y la facilidad para realizar búsquedas en su contenido suministran en muy poco tiempo datos que antes eran ilocalizables. Por otra parte, la creación de corpus electrónicos diacrónicos, tanto de carácter general como especializado, aportan materiales que permiten llevar a cabo estudios sobre múltiples aspectos de interés para reconstruir la historia de la lengua en todo el ámbito hispánico.

Sin embargo, estos avances que han traído consigo un indiscutible progreso no están exentos de algunos problemas que se derivan de su concepción y de su funcionamiento. La dispersión de los recursos, la falta de interconexión, la carencia de estándares, la rápida obsolescencia, etc., dificultan el acceso a la información que contienen. Su diseño, por otra parte, no siempre resulta adecuado para determinadas investigaciones. Todo esto hace que sea necesaria una crítica especializada, desde el punto de vista del usuario filólogo, que contribuya

a señalar sus puntos débiles para poder mejorar las próximas versiones y los nuevos productos.

Las posibilidades de estudio del léxico con estos recursos son aplicables a todas las tareas e investigaciones que tradicionalmente han llevado a cabo los historiadores del léxico: elaboración de diccionarios diacrónicos y tesoros lexicográficos, estudios sobre la configuración histórica de los campos semánticos, de dialectología histórica, sobre los cambios morfosintácticos, fónicos, gráficos, etc.

En este volumen, especialistas de distintos aspectos de la historia de la lengua española dan a conocer de primera mano los proyectos que están desarrollando y abordan, desde diferentes perspectivas, los retos que plantean los nuevos métodos de trabajo. Los capítulos se han estructurado en tres secciones: una dedicada a los «Corpus y recursos actuales», que muestra algunos de los proyectos internacionales más destacados; un segundo bloque en el que se analizan con perspectiva crítica estos recursos; y una tercera parte donde se ofrecen ejemplos concretos de las posibilidades que presentan estas herramientas para el estudio del léxico español. En su conjunto, los distintos capítulos incluidos en esta monografía dan cuenta de los proyectos en curso de realización de ámbito panhispánico que han surgido por la necesidad de organizar y dar a conocer unos materiales extraídos de fuentes propiamente lingüísticas, literarias, lexicográficas o paralexicográficas, que resultan imprescindibles para plantear la historia del vocabulario en su diversidad diatópica, diastrática y temporal.

Los primeros capítulos están dedicados a tres de los proyectos de mayor proyección que han surgido en los últimos años, por los recursos que han captado y por la magnitud de datos que han puesto al alcance de los investigadores. La profesora de la Universidad de Santiago de Compostela y coordinadora del *Nuevo Diccionario Histórico del Español*, Mar Campos Souto, analiza en el capítulo «Las bases documentales del *NDHE*: Entre la realidad y el deseo» una buena parte de las herramientas utilizadas en este macroproyecto, desde las lexicográficas hasta los corpus académicos, sin desatender a las fuentes bibliográficas y las hemerotecas. Concebido como un diccionario en línea, el inicio de la redacción de los artículos del *NDHE* ha estado precedido de un intenso trabajo previo de digitalización y puesta en red de numerosos recursos por parte de la Real Academia Española, bien a través de su página web o de la correspondiente al Instituto de Investigación Rafael Lapesa. Estos materiales han surgido por iniciativa de la RAE (como, por ejemplo, el *Nuevo tesoro lexicográfico de la lengua española —NTLLE—*) o del equipo mismo del nuevo diccionario histórico (la digitalización del *Fichero general* o del *Diccionario de Autoridades* o la lematización, anotación y codificación de buena parte de los textos que integran el *CORDE* y el *CREA*, los dos grandes

corpus académicos del siglo XX, en parte relevados por las dos apuestas del presente siglo: el *CORPES XXI* y el *CDH*, que han desarrollado una interfaz acorde con las necesidades que en este momento demanda el lingüista pero también abierta a futuras propuestas y actualizaciones). Otras veces, la adopción, difusión o consulta de recursos no académicos se ha realizado por medio de convenios con otros equipos de investigación (como el grupo *CorLexIn*, de la Universidad de León) y entidades (tales como el Instituto de Estudios Canarios o la Biblioteca Nacional).

Otra de las grandes apuestas de recursos en red es el *Corpus del Español* dirigido por Mark Davies (Brigham Young University). Se trata, en realidad, de tres bases de datos: el *CE*, de carácter histórico, de unos cien millones de palabras, creado en 2002; el *CE-2* o *Web Dialectos*, de dos mil millones de palabras, accesible desde 2016, que contiene registros extraídos exclusivamente de Internet y que, como novedad, se enfoca hacia la variación geolectal (el léxico diferencial y la frecuencia de uso de las palabras en cada uno de los veintiún países de habla hispana); y un último recurso, denominado *NOW-Spanish*, de unos seis mil millones de unidades, previsto para finales de 2018, que se convertirá en un observatorio de los neologismos y del proceso del cambio léxico. La comparación con los corpus académicos (*CORDE*, *CREA* y *CORPES XXI*) permite al autor poner de relieve algunas de las mejoras incorporadas, como la posibilidad de recuperar la información por géneros dentro de cada sincronía, advertir los patrones del cambio en las colocaciones o, personalizando la búsqueda, comparar la frecuencia de los componentes de un campo semántico.

Las novedades del *Corpus Diacrónico y Diatópico del Español de América* (*CORDIAM*) constituyen la base del capítulo que presentan sus directoras, Concepción Company (Universidad Nacional Autónoma de México) y Virginia Bertolotti (Universidad de la República, Uruguay). El corpus surgió en 2012 a partir de la necesidad de hacer accesibles los materiales del español americano y de construir una nueva historia de la lengua fundamentada en la documentación manuscrita (*CORDIAM-documentos*). Recientemente ha ido adquiriendo unas dimensiones relativamente importantes al incorporar otros dos subcorpus: uno literario (*CORDIAM-literatura*) y otro de textos procedentes de las hemerotecas hispanoamericanas (*CORDIAM-prensa*). Junto a los metadatos habituales (cronología, localización, autor), el equipo de trabajo ha perfilado una serie de tipos textuales, atendiendo a la finalidad comunicativa del registro, de tal manera que, frente a otros corpus, no es lo cuantitativo lo que prevalece, sino la calidad de los materiales indexados y la red de relaciones que se puede establecer entre ellos.

El español de América y el léxico registrado en las definiciones de los vocabularios bilingües indoamericanos constituye el objetivo del *TELEAM*, un proyecto

de la investigadora Esther Hernández, del CSIC. Este tesoro léxico colonial contribuirá al mejor conocimiento de la historia de las voces y expresiones diferenciales que se crearon en el Nuevo Mundo. Se trata de presentar, de manera cronológicamente ordenada y sucinta, una información metalexicográfica de primer orden que apenas se ha tenido en cuenta porque ha pasado desapercibida y que, sin embargo, ofrece datos relevantes sobre la adaptación e incorporación de determinados indigenismos, su etimología, su significado, su registro o particular difusión diatópica. No resulta una novedad la presencia de «diccionarios de diccionarios» entre las producciones más características de la lexicografía española, continuando la estela con que Gili Gaya planteó, hace ya casi un siglo, su *Tesoro lexicográfico*. En este sentido, Dolores Corbella, de la Universidad de La Laguna, analiza la pervivencia de aquel modelo en la historiografía hispánica (incluyendo también la portuguesa, la gallega y la catalana), lo que ha supuesto para el desarrollo de la lexicografía regional dentro del español europeo y los proyectos realizados o en curso de elaboración en el español americano, entre los que destaca el *Tesoro Lexicográfico de Puerto Rico* digital, accesible en la web desde 2016. El futuro próximo será plantear una red de tesoros lexicográficos interconectados que facilite la labor del investigador, de tal modo que pueda acceder de forma inmediata a la recepción y tradición diccionarística de cualquier término, a su historia y a su circunscripción topolectal.

Ese binomio entre tradición e innovación puede advertirse asimismo en dos de los proyectos que se realizan en Alemania y en el CSIC sobre los materiales inéditos del *Diccionario del español Medieval* y del *Atlas Lingüístico de la Península Ibérica*. El primero, codirigido por Rafael Arnold y Jutta Langenbacher-Liebgott, al frente de un equipo de colaboradores de las Universidades de Rostock y Paderborn, retoma el trabajo que llevó a cabo durante casi cuatro décadas el profesor Bodo Müller en la Universidad de Heidelberg, con la finalidad de digitalizar casi un millón de testimonios léxicos de aquel proyecto pionero dentro de la lexicografía española, que venía a cubrir una época determinante en la conformación de la lengua. El segundo, dirigido por Pilar García Mouton, es el resultado de un proyecto del CSIC, que ha conseguido aunar las nuevas tecnologías (digitalización y cartografiado) para dar a conocer unos materiales que durante años se creyeron perdidos. Las características de aquel atlas, ideado por Menéndez Pidal y llevado a cabo por Navarro Tomás en la década de los años treinta del siglo pasado (y retomado después de la Guerra Civil), y el tiempo transcurrido desde entonces dan a las encuestas realizadas un valor excepcional al mostrar la riqueza léxica de la España rural de aquellos tiempos.

El *Diccionario de la alimentación y la culinaria medievales y renacentistas*, que presenta el profesor Rolf Eberenz en la Universidad de Lausana, aparece ya como

una espléndida realidad, con una planta magníficamente estructurada y una buena parte de los lemas redactados, tal como puede comprobarse en la página web del proyecto. En este caso, no es solo la limitación temporal lo que lo hace único (desde el s. XIII hasta principios del XVII), sino especialmente el que haya sido concebido como un diccionario temático con el que su autor demuestra que «los hábitos alimenticios constituyen un código cultural que tiene implicaciones lexicológicas».

Cierra este primer bloque la presentación del corpus literario *Texbox*, de José Calvo Tello, Ulrike Henny-Krahmer y Christof Schöch, de las Universidades de Wurzburgo y Tréveris. Se trata de una colección de acceso libre de textos literarios románicos en formato digital que integra, entre otros, dos subcorpus de novelas españolas editadas entre 1880 y 1940 y doce colecciones de cuentos seleccionados dentro del mismo marco temporal. Tras realizar una somera descripción de los pasos seguidos para la preparación del corpus, utilizando el formato TEI, los investigadores muestran algunas de las posibilidades que ofrece para análisis de tipo cualitativo y cuantitativo, tanto literarios como lingüísticos.

Este nuevo entorno de las Humanidades supone un cambio, no tanto en la naturaleza misma de la investigación como en el modo de afrontarla y en la posibilidad de poner a disposición de la comunidad de investigadores las fuentes mismas en las que se basan los proyectos. El problema que plantean estos recursos para la lingüística histórica parte de lo que podríamos denominar *desorden digital*, debido en un principio al ritmo incesante de actualización del soporte tecnológico, pero también a la falta de homogeneidad en el tratamiento de los textos, a la convergencia de materiales de procedencia muy dispar y, sobre todo, a la imposibilidad de transferencia e intercambio entre los distintos proyectos. Se están creando las autopistas de información, pero no siempre las rotondas que permitan la optimización de los resultados. Al análisis de estos temas están dedicados, en la segunda sección, los trabajos de Alejandro Fajardo (Universidad de La Laguna) y de Francisco Javier Herrero Ruiz de Loizaga (Universidad Complutense de Madrid).

Los avances, no obstante, resultan extraordinarios y sus aplicaciones al estudio del léxico permiten acceder a datos desconocidos hasta ahora para la lexicografía tradicional, como se comprueba en la tercera sección. Así lo demuestran Miguel Calderón Campos (de la Universidad de Granada) con el análisis de los andalucismos registrados en el *Corpus del reino de Granada* (*CORDEREGRA*), y María Jesús Torrens o Pedro Sánchez-Prieto Borja y Delfina Vázquez con la investigación que realizan a partir de los documentos del *Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid* (*ALDICAM-CM*). Ambos proyectos surgieron dentro de la red internacional *CHARTA*, un macrocorpus

global concebido con la idea de aunar esfuerzos para realizar ediciones unificadas de textos archivísticos de los siglos XII al XIX.

Facilitar el estudio del léxico del Siglo de Oro a partir de la transcripción literal de documentación notarial recogida en *CorLexIn* (*Corpus Léxico de Inventarios*) ha sido el objetivo del proyecto que desarrollan en la Universidad de León José Ramón Morala (director), Cristina Egido y otros investigadores de las Universidades de Burgos y Oviedo. El corpus textual, en fase muy avanzada de ejecución, presenta una acotación temporal que se circunscribe a los siglos XVI y XVII e incorpora una muestra representativa de textos (*relaciones de bienes*, esto es, inventarios y tasaciones, partijas y repartos de herencias, cartas de dote y arras, almonedas, testamentos, etc.) procedentes de todos los archivos históricos provinciales de España y buena parte de América.

Fuentes similares ha utilizado Miguel Ángel Puche Lorenzo en su trabajo sobre el léxico murciano del cuatrocientos. Como en muchas otras comunidades, los textos transcritos y disponibles en internet resultan escasos (excepto los proporcionados por la red *CHARTA* y por el proyecto *CorLexIn*), si bien ha existido entre los historiadores una larga trayectoria de edición de la documentación en formato papel. En los últimos años se cuenta, además, con los resultados del proyecto *Carmesí*, que ha puesto a disposición de los usuarios la digitalización de los manuscritos del período medieval conservados en los archivos históricos de Murcia.

El análisis de los registros lexicográficos académicos de la primera mitad del siglo XIX constituye el objetivo principal del capítulo presentado por Mª Ángeles Blanco Izquierdo (Centro de Estudios de la RAE), Gloria Clavería Nadal (Universidad de Barcelona) y Enrique Jiménez Ríos (Universidad de Salamanca). El *NTLLE* ha facilitado sobremanera la consulta del *DRAE* en sus sucesivas entregas, con lo que se ha avivado el interés por conocer cuáles fueron los motivos que llevaron a la renovación del léxico académico en sus distintas ediciones y cómo se plasmaron en la planta del diccionario los acuerdos adoptados y las innovaciones admitidas.

El último trabajo que incluye este libro atiende a otra fuente fundamental de cambio léxico que no siempre se ha tenido en cuenta por su dispersión: la prensa periódica. José Ignacio Pérez Pascual (de la Universidad de La Coruña) investiga un período (la «Edad de Plata», de 1885 a 1936) que tradicionalmente ha pasado inadvertido, pero que resulta extremadamente significativo por las revoluciones sociológicas, sociales y tecnológicas que se produjeron y que tuvieron como correlato un cambio profundo del vocabulario, con la adopción de numerosos préstamos y la creación de abundantes neologismos. Sin duda, tal

como demuestra Pérez Pascual, las hemerotecas digitales constituyen otro de los recursos imprescindibles para la reconstrucción de la historia del léxico.

En este siglo de siglas, el lexicógrafo empieza a contar con innumerables recursos que, sin duda, facilitan su labor pero que, a la vez, por su cantidad, pueden llegar a abrumar a cualquier «labrante de las palabras». La conjunción filología/informática parece ya indisoluble y los proyectos aquí presentados lo confirman. Nuestra intención al plantear la edición de este libro no ha sido otra que mostrar los caminos más logrados que los historiadores del léxico estamos transitando y las innumerables perspectivas y posibilidades que el futuro digital presenta para una disciplina que, en el siglo XXI, demanda un trabajo más colaborativo e interdisciplinar.

La Laguna/Paderborn, 31 de enero de 2018.

Índice

Lista de autores 15

I Corpus y recursos actuales 17

Mar Campos Souto
Las bases documentales del *NDHE*: Entre la realidad y el deseo 19

Mark Davies
Uso del *Corpus del Español* y los corpus relacionados para la lexicografía histórica española 47

Virginia Bertolotti y Concepción Company Company
El corpus para América: *CORDIAM* 75

Esther Hernández
Tesoro léxico de los americanismos contenidos en los vocabularios hispano-amerindios coloniales (1550–1800) [*TELEAM*] 107

Dolores Corbella
Del tesoro lexicográfico analógico al digital 133

Rafael Arnold, Stefan Serafin, Anna-Susan Franke y Jutta Langenbacher-Liebgott
Una nueva fuente para la historia del léxico español: el *DEMel* 165

Pilar García Mouton
Geolingüística y Humanidades digitales: el *Atlas Lingüístico de la Península Ibérica* (*ALPI*) 189

Rolf Eberenz
Hacia un diccionario de la alimentación y la culinaria medievales y renacentistas 203

José Calvo Tello, Ulrike Henny-Krahmer y Christof Schöch
Textbox: análisis del léxico mediante corpus literarios 223

II Crítica de los recursos en línea: el desorden digital 253

Alejandro Fajardo
Lexicografía histórica con corpus y recursos digitales: aspectos metodológicos 255

Francisco Javier Herrero Ruiz de Loizaga
Algunos problemas en la aplicación de los corpus informatizados al estudio de la diacronía del español, con especial atención a los procesos de gramaticalización 279

III Del corpus a los estudios léxicos 315

Miguel Calderón Campos
Andalucismos en el corpus del reino de Granada 317

Pedro Sánchez-Prieto Borja y Delfina Vázquez Balonga
El léxico en los documentos de la Comunidad de Madrid (ss. XVI–XIX) 341

María Jesús Torrens Álvarez
El proyecto *ALDICAM-CM* y el ejemplo de los fueros de Alcalá para el estudio de la historia del léxico 379

José R. Morala y Mª Cristina Egido
El proyecto *CorLexIn* y la variación diatópica en el léxico del Siglo de Oro 397

Miguel Ángel Puche Lorenzo
Estudio del léxico castellano a través de fuentes medievales murcianas 419

M.ª Ángeles Blanco Izquierdo, Gloria Clavería Nadal y Enrique Jiménez Ríos
Fuentes lexicográficas y estudio del léxico: el *Diccionario de la lengua castellana* de la Real Academia Española (1817–1852) 449

José Ignacio Pérez Pascual
Las publicaciones periódicas y el estudio del léxico de la «Edad de Plata» ... 477

Lista de autores

Rafael Arnold
Universität Rostock

Virginia Bertolotti
Universidad de la República, Uruguay

M.ª Ángeles Blanco Izquierdo
Centro de Estudios de la RAE

Miguel Calderón Campos
Universidad de Granada

José Calvo Tello
Universität Würzburg

Mar Campos Souto
Universidad de Santiago de Compostela

Gloria Clavería Nadal
Universidad Autónoma de Barcelona

Concepción Company Company
Universidad Nacional Autónoma de México

Dolores Corbella
Universidad de La Laguna

Mark Davies
Brigham Young University

Rolf Eberenz
Université de Lausanne

Mª Cristina Egido
Universidad de León

Alejandro Fajardo
Universidad de La Laguna

Anna-Susan Franke
Universität Paderborn

Pilar García Mouton
Instituto de Lengua, Literatura y Antropología, Consejo Superior de Investigaciones Científicas

Ulrike Henny-Krahmer
Universität Würzburg

Esther Hernández
Instituto de Lengua, Literatura y Antropología, Consejo Superior de Investigaciones Científicas

Francisco Javier Herrero Ruiz de Loizaga
Universidad Complutense de Madrid, Instituto Universitario Menéndez Pidal

Enrique Jiménez Ríos
Universidad de Salamanca

Jutta Langenbacher-Liebgott
Universität Paderborn

José R. Morala
Universidad de León

José Ignacio Pérez Pascual
Universidade da Coruña

Miguel Ángel Puche Lorenzo
Universidad de Murcia

Pedro Sánchez-Prieto Borja
Universidad de Alcalá

Christof Schöch
Universität Trier

Stefan Serafin
Universität Rostock

María Jesús Torrens Álvarez
Instituto de Lengua, Literatura y Antropología, Consejo Superior de Investigaciones Científicas

Delfina Vázquez Balonga
Universidad de Alcalá

I
Corpus y recursos actuales

Mar Campos Souto

Las bases documentales del *NDHE*: Entre la realidad y el deseo

Resumen: Este trabajo ofrece un examen de tres tipos de fuentes utilizadas en la redacción del *Nuevo diccionario histórico del español* (*NDHE*) de la Real Academia Española: las que se denominan fuentes tradicionales (repertorios o tesoros lexicográficos y ficheros), los corpus diacrónicos (en concreto, el *Corpus del Nuevo diccionario histórico del español*) y algunas hemerotecas y bibliotecas digitales. Este análisis pretende ofrecer una aproximación, desde la perspectiva práctica del trabajo lexicográfico, a sus diferentes características, posibilidades de explotación y limitaciones, así como sugerir algunas vías de mejora de las herramientas de consulta de estas fuentes.

Palabras clave: Lexicografía diacrónica, *NDHE*, Bases documentales

Abstract: This paper offers an examination of three types of sources used in the writing of the *New Historical Dictionary of Spanish (NDHE)* of the Royal Spanish Academy: those considered as traditional sources (lexicographical collections or thesauri and catalogues), the diachronic corpora (specifically, the *Corpus of the New Historical Dictionary of Spanish*) and some digitalized newspaper archives and libraries. This analysis aims to offer not only an approximation from the practical perspective of the lexicographical work, to its different characteristics and its possibilities of exploitation and limitations, but also to suggest some ways to improve the tools for consulting those sources.

Keywords: Historical lexicography, *NDHE*, Documentary databases

1 Introducción

En este capítulo se presentará un breve análisis de tres tipos de fuentes empleadas en la elaboración del *Nuevo diccionario histórico del español* (*NDHE*) de la Real Academia Española[1]. Esta descripción no pretende agotar la riqueza de bases documentales empleadas en el proyecto, sino que únicamente intenta ofrecer una aproximación a algunas de ellas, a sus diferentes características, posibilidades de explotación y limitaciones; y, relacionado con esto último, a la exigencia

1 Repertorio, en curso de elaboración, accesible en <http://web.frl.es/DH/org/login/Inicio.view>.

de continuar trabajando para mejorar sus opciones de consulta, así como para dotarlas de mayor fiabilidad filológica. Por consiguiente, este capítulo no se sitúa en la perspectiva del diseño o la planificación original de estos recursos, sino en la de su utilización para un diccionario histórico, es decir, en la modesta experiencia de unos usuarios con unos intereses y unas necesidades específicas. Examinaremos, en consecuencia, tres tipos de fuentes: las que hemos denominado tradicionales (que solo se mencionarán brevemente, aunque merecerían un estudio monográfico y exhaustivo); el *Corpus del Nuevo diccionario histórico del español* (*CDH*), como muestra de un corpus diacrónico; y algunas hemerotecas y bibliotecas digitales.

2 Fuentes tradicionales

2.1 Repertorios o tesoros lexicográficos

La consulta de los artículos publicados del *NDHE* pone de relieve la utilización de una heterogénea nómina de fuentes; así, por ejemplo, los primeros testimonios del artículo *varicela* apuntan hacia procedencias diversas (*Biblioteca digital*, *Hemeroteca digital*, la extensión diacrónica del *CDH* y el *Nuevo tesoro lexicográfico del español* de la Real Academia Española —*NTLLE*—). En efecto, en el *NDHE* se emplean como fuentes de información y documentación distintos tesoros lexicográficos (como el ya citado *NTLLE*, accesible en internet, o el *Nuevo tesoro lexicográfico del español* dirigido por L. Nieto y M. Alvar, publicado en papel) y los diccionarios históricos del español (los dos parciales elaborados por la Academia, así como los centrados en una variedad del español, como el *Diccionario histórico del español de Canarias*). En el marco del proyecto del *NDHE* se han elaborado versiones digitales o electrónicas de buena parte de estos repertorios (en algún caso, gracias al establecimiento de convenios con otras instituciones, como el Instituto de Estudios Canarios), versiones que, tras un breve período de prueba, se han puesto a disposición de la comunidad científica, en la red, con el fin de facilitar su acceso y difusión[2].

2 El 30 de marzo de 2012 se publicaron, en la página de la Fundación Rafael Lapesa, la versión digital del *Fichero general* de la Real Academia Española, el *Mapa de diccionarios académicos* y una versión electrónica de los fascículos publicados del *Diccionario histórico de la lengua española* de la Real Academia Española (1960–1996). El 3 de agosto de 2013 se incluyó la versión electrónica del *Diccionario de Autoridades* y el PDF de los dos tomos del primer *Diccionario histórico de la lengua española* (1933–1936). La versión electrónica del *Diccionario histórico del español de Canarias*, de C. Corrales Zumbado y D. Corbella, por su parte, se incorporó a esa página en diciembre de 2014.

Excepcionalmente, los repertorios lexicográficos brindan las primeras documentaciones de una voz en el *NDHE*; así sucede, por ejemplo, con *maríbula* que, curiosamente, pese a ser una voz cubana, se registra por vez primera en 1846 en el *Nuevo diccionario de la lengua castellana* de Vicente Salvá, hecho que se explica porque Salvá tuvo a su disposición un manuscrito, aún inédito, de voces cubanas que circuló por París[3]. Obviamente, el hecho de situar el primer testimonio de un vocablo en un repertorio lexicográfico induce a pensar que existe documentación previa, no disponible o no localizada. En otras ocasiones, los diccionarios históricos, construidos sobre una base textual, nos facilitan el acceso a unos documentos que, de otro modo, habría sido muy difícil —si no imposible— localizar[4].

2.2 Ficheros: el *Fichero general* de la Real Academia Española

Si bien se suele afirmar que la lingüística histórica siempre ha sido una lingüística de corpus, es innegable que la constitución de los grandes corpus textuales, nacidos, en buena medida, gracias a la decisiva aportación de las disciplinas computacionales, ha provocado una profunda transformación en el ámbito de los estudios lingüísticos y, en particular, en el de las investigaciones diacrónicas sobre el léxico. Previamente a la aparición de los corpus informatizados, la base textual de los diccionarios sustentados en ejemplos de uso se apoyaba en los datos extraídos de ficheros de diversa índole; en este sentido, los elaborados por la Real Academia Española a lo largo de su historia (como el *Fichero general*, el *Fichero de adiciones y enmiendas*, el *Fichero de hilo o de autoridades* o el *Fichero Rico y Sinobas*) constituyen un abundante venero del que beben los diccionarios académicos a lo largo de su historia[5].

En 2019 se publicarán los materiales inéditos del primer *Diccionario histórico* académico, compuesto por 29 legajos que contienen los artículos comprendidos entre *cia* y *efélide* (véase Seco 1980: 63, n. 37 y Campos Souto 2017: 167–168, n. 7).

3 Agradezco esta información a Armando Chávez Rivera.

4 Así, *ajabeba* 'flauta morisca', se documenta por primera vez en el *Libro de diferentes cuentas de entrada y distribución de las Rentas Reales y gasto de la casa Real en el Reynado de D. Sancho IV, Era 1331 y 1332, que son años 1293 y 1294* (1294, manuscrito del siglo XIII), obra –y testimonio– que figuran en el *Diccionario histórico* de 1960–1996, cita que ha permitido acudir a la fuente original, accesible en la *Biblioteca digital hispánica* de la Biblioteca Nacional de España.

5 Aunque supera los objetivos de este trabajo, es importante señalar que la nómina de estos ficheros explica no solo el canon textual sobre el que se levantan los diccionarios históricos del español del siglo XX, sino también, en cierto modo, el que preside la constitución de los corpus académicos —aunque con evidentes modificaciones,

El *Fichero general* (*FG*) de la Real Academia Española está conformado por más de diez millones de cédulas, que consignan testimonios léxicos y lexicográficos de las voces estudiadas; aunque, como se indica en la presentación de la versión electrónica de este recurso, «su período de máxima expansión se sitúa entre 1930 y 1996, fechas en que la Academia afrontó la redacción del *Diccionario histórico* en sus dos ediciones», investigaciones recientes muestran que las tareas de papeletización (vinculadas, en un principio, a la entonces ansiada —y tantas veces planeada— nueva edición del *Diccionario de autoridades*), se intensifican desde el segundo decenio del siglo XX (Campos Souto 2017: 169)[6]. La diferente calidad de las cédulas incluidas en el *FG* y las penosas tareas de cotejo de sus datos con los textos originales persuadieron a la Academia de la necesidad de afrontar un proyecto de informatización del *FG*, proyecto que se desarrolló a partir de 1993 y que alcanzó a medio millón de fichas. Sin embargo, probablemente debido a su elevado coste y a la aparición de los proyectos de conformación de los corpus, se suspendió finalmente en 1995.

El *FG* constituye aún hoy una fuente de indudable riqueza, debido a la generosa nómina de textos despojados y a la atención privilegiada que se brindó a las voces —o acepciones— menos frecuentes o, si se prefiere, a aquellas consideradas más distantes de las vigentes en los períodos en que se fueron confeccionando las cédulas. Por ese motivo, se continúa empleando como base documental en el *NDHE*, como se puede comprobar en el artículo consagrado a *partesana*, uno de cuyos primeros testimonios procede del *FG*; el texto corresponde al primer tomo del *Tratado de las campañas y otros acontecimientos de los ejércitos del emperador Carlos V*, de Martín García de Cereceda, que se cita por la edición de 1873–1876, de G. Cruzada Villaamil. En la actualidad, puede accederse a esa edición en la *Biblioteca digital hispánica* de la Biblioteca Nacional de España, pero, quizá por problemas derivados del programa de OCR, el vocablo no se recupera en la búsqueda, ni siquiera acotando la fecha de edición.

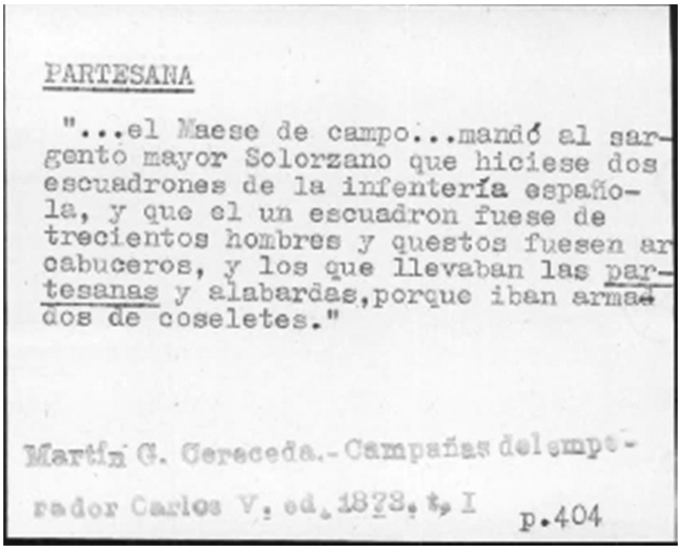
El protocolo establecido en el *NDHE* determina que, cuando se localiza un testimonio de interés en el *FG*, se debe intentar consultar directamente la edición citada en la cédula o, incluso, el manuscrito original[7]. Así sucede en el caso

---

debidas tanto a la ampliación del abanico temporal considerado como, fundamentalmente, a los cambios incluidos en la conformación del canon por los historiadores de la literatura—.

6 La versión electrónica del fichero puede consultarse en <http://web.frl.es/fichero.html>.

7 A nadie se le oculta que citar una obra del siglo XVI a través de una edición del XIX resulta claramente insatisfactorio, por más que el editor declare haber «cuidado mucho de no alterar el texto en lo más mínimo», si bien confiesa su objetivo de «arreglar, si


PARTESANA

"...el Maese de campo...mandó al sargento mayor Solorzano que hiciese dos escuadrones de la infentería española, y que el un escuadron fuese de trecientos hombres y questos fuesen arcabuceros, y los que llevaban las partesanas y alabardas, porque iban armados de coseletes."

Martín G. Cereceda.- Campañas del emperador Carlos V, ed. 1878, t. I p.404

Ilustración 1: Ficha n.º 38 de *partesana* en el *Fichero general*

del *Diálogo de la vida beata* de Juan de Lucena, de 1463 que, en el artículo de *caramillar*, por ejemplo, facilita la primera documentación de la voz, a través de la edición de 1950[8]; uno de sus testimonios, el manuscrito 6728 de la Biblioteca Nacional de España, copiado entre 1465 y 1467, es hoy accesible también en la *Biblioteca digital hispánica*. Estos dos casos demuestran que, pese a la existencia de una notable coincidencia, se observan también algunas discontinuidades en el establecimiento del canon textual entre el *FG* y sus herederos, los corpus: ambas obras, profusamente empleadas en el *Diccionario histórico* de 1960–1996, no se incorporaron al *CORDE* (ni, por tanto, al *CDH*)[9].

así puede decirse, la puntuación, de que por completo el códice carece, dejando su misma ortografía á aquellas palabras en que, al cambiarla, hubiera cambiado el sonido» (1873: XIII).

8 A través de la edición de G. M. Bertini (*Testi spagnoli del secolo XV,* Turín, Gheroni, 1950).

9 Las *Campañas* se citan en cuarenta artículos del *DH*-1960–1996 e incluso ofrecen los primeros testimonios en algunos casos (así, en *abestionar* 'abastionar, fortificar' o *anconitano, a* 'perteneciente o relativo a Ancona'). Por su parte, la obra de Lucena se

Por otra parte, en el *FG* se percibe el interés por papeletizar repertorios lexicográficos restringidos, por lo que abundan cédulas en que se recogen datos procedentes de diccionarios dialectales o vocabularios de lenguas de especialidad. Este hecho explica que la primera documentación de *güiro*, como sinónimo de *cabeza*, se obtenga del *FG* y, en concreto, del *Diccionario de americanismos* de Malaret, de 1925, fuente de este testimonio. En este ámbito (en el vaciado de fuentes lexicográficas restringidas), el *FG* continúa siendo un filón inagotable, pues en él se incluyen repertorios de tanta relevancia como el *Vocabulario matemático-etimológico* de F. Picatoste (1862), el *Diccionario militar* de J. Fernández Mancheño (1822), el *Vocabulario de mexicanismos* de J. García Icazbalceta (1894), el *Vocabulario andaluz* (1933) de A. Alcalá Venceslada, los *Hondureñismos* (1895) de A. Membreño o el *Vocabulario cubano* (1859) de J. García Arboleya, por citar solo algunos.

Pese a sus deficiencias (entre las que han de citarse las asignaciones erróneas de un conjunto de fichas a un lema equivocado), la versión digitalizada del *FG* ha incrementado notablemente las posibilidades de acceder a este recurso, aunque también es indiscutible que lo mejor habría sido disponer de una versión electrónica del *FG*, precisamente el propósito que perseguía el proyecto inacabado que se emprendió en 1993. Posteriormente, en el año 2006, la Academia retomó de algún modo esa idea y elaboró (hasta 2008) una base de datos de primeras documentaciones de las palabras del *Fichero*; en ella, se registraba sistemáticamente la información bibliográfica de las papeletas de punto rojo y se lematizaban todas las formas documentadas de cada voz[10]. En el camino hacia lo bueno, quizá sería posible, en un futuro próximo, intentar cruzar las informaciones contenidas en esta base de datos con las imágenes, lo que permitiría ofrecer una consulta mucho más refinada y rica, así como construir un

cita en noventa y cuatro artículos de este repertorio, si bien se emplean dos ediciones: la ya citada de Bertini (véanse, por ejemplo, los artículos *anjoíno, na*, *antojar* o *bacil*) o la incluida en los *Opúsculos literarios de los siglos XIV a XVI* (1892), citada en setenta y siete artículos (como *abastar*, *abundoso*, *ánima* o *antecámara*).

10 Las cédulas que incluyen el primer testimonio de una voz se identifican mediante un círculo o punto de color rojo en el ángulo superior derecho del *FG* (o, con menos frecuencia, de color azul), recurso que permitía localizar fácilmente ese primer testimonio al revisar las fichas en las gavetas.

lemario depurado y unificado del *FG*, además de ofrecer una nómina de textos citados. Este sería solo un modesto paso en la dirección, más ambiciosa, que persigue hoy la Academia: construir una consulta unificada de todos los recursos léxicos, lexicográficos y gramaticales que atesora.

3 Los corpus: el *CDH*

Como se ha señalado en otras ocasiones, la confección de diferentes corpus y el desarrollo de aplicaciones de consulta para facilitar su aprovechamiento constituye una de las causas determinantes en la revolución que ha experimentado la lexicografía en los últimos años[11]; la mera comparación entre la cifra de cédulas almacenadas en el *FG* —y su único modo de acceso, frente a las diversas posibilidades de consulta de un corpus— con el número de ocurrencias del *CORDE* o del *CDH* resulta suficientemente esclarecedor —sin mencionar, por otra parte, la notable reducción de tiempo y recursos implicados en su confección—. Kilgarriff *et al.* han recalcado, a su vez, la influencia determinante de la lexicografía en el desarrollo de los corpus, puesto que, en su opinión, esta disciplina «was the driving force in the development of corpus methods and corpus use» (2014: 14), si bien es indudable que esta aseveración cobra mayor fuerza en el ámbito de la lexicografía sincrónica que en la diacrónica. Sin embargo, diferentes estudios han mostrado las debilidades de estos corpus en distintos planos —y, particularmente, en el filológico—. Probablemente no existe hoy un consenso acerca de las características que, para el estudio del léxico diacrónico, debe poseer un corpus; probablemente, también, las prioridades marcadas por los lingüistas computacionales y los filólogos se sitúen en puntos distantes, pero no parece imposible enumerar algunas de las propiedades que debería presentar un corpus amplio, fiable, representativo y de fácil explotación para los historiadores del léxico del español:

a) Los textos del corpus han de estar lematizados.
b) Los textos deben someterse a un proceso de codificación textual.
c) La interfaz de consulta debe permitir una variada gama de búsquedas.
d) El corpus debe ser amplio y representativo.
e) Los corpus deben ser filológicamente fiables; en ese sentido, los criterios de transcripción o edición de los textos, así como las pautas empleadas para su

11 Véanse, entre otros, Béjoint (2007), Rojo (2009), Rafel i Fontanals (2011), Hanks (2012), Kilgarriff (2013) y Campos Souto (2016).

selección, deben ser explícitos y constar entre la documentación pública del corpus.

f) La copia digital de los testimonios base (o de los documentos transcritos) debe ser accesible.

Entre ese deseo y la realidad con la que trabajamos cotidianamente en el *NDHE* media cierta distancia, una distancia que se intenta suplir mediante algunas acciones que aspiran a paliar los problemas que presenta —como, por otra parte, otros bancos documentales— el *CDH*. Y, sobre todo, con la conciencia de que, pese a sus flancos débiles, los corpus hoy son nuestros principales aliados en la investigación sobre la historia del léxico, unos aliados que debemos conocer y, en la medida de nuestras posibilidades, mejorar.

3.1 Los textos han de estar lematizados

Los textos que conforman el *CDH* nuclear se sometieron a un proceso semiautomático de anotación lingüística (operación llevada a cabo por el Departamento de Tecnología de la Real Academia Española —que también se encargó de anotar los textos procedentes del *CREA*—), en tanto que los textos integrados en la extensión diacrónica del *CDH* poseen una preanotación morfosintáctica, realizada con herramientas de software libre (*Freeling*). Entendemos que la lematización constituye un punto de partida para el manejo de los datos en el quehacer lexicográfico; de hecho, la lematización del corpus se perfecciona a medida que se redacta el diccionario, pues en este proceso se reduce el notable grado de ambigüedad categorial que presenta la anotación y se depuran los posibles errores de asignación a lemas que se hayan deslizado previamente. Es esta una consecuencia de la integración del corpus en la herramienta de redacción del diccionario, que garantiza la interconexión entre ambos elementos, de tal modo que no solo las bases documentales alimentan el *NDHE* (suministrándole su primera cantera de ejemplos), sino que el *NDHE* contribuye a mejorar la calidad de los corpus[12].

3.2 Los textos deben someterse a un proceso de anotación o codificación textual

La interfaz de consulta del *CDH* permite un amplio abanico de consultas, entre las que se pueden mencionar, por ejemplo, la posibilidad de desechar los fragmentos escritos en otra lengua, así como la de prescindir —u obtener— de

12 Las características fundamentales de la herramienta de redacción del *NDHE*, denominada *ARDIDEs*, se describen en Salas Quesada y Torres Morcillo (2011 y 2015).

aquellos testimonios de una voz que se hayan anotado como citas o cambios de mano. Consultas de este cariz solo se pueden efectuar si los textos se han sometido a un exhaustivo proceso de codificación textual, codificación que debería, en la medida de lo posible, ajustarse a alguno de los estándares que faciliten el intercambio de datos —como la TEI—, puesto que la sujeción a los estándares permite garantizar que, aunque cada colección documental o corpus mantenga su identidad, pueda, al tiempo, ser susceptible de integrarse con facilidad en otras bases de datos —y, de este modo, se facilite su consulta y su reutilización en otros bancos de datos.

3.3 La interfaz de consulta debe permitir una variada gama de búsquedas

La interfaz de consulta del *CDH* (cuya primera versión se remonta a noviembre de 2009) se ha diseñado con el objetivo de ofrecer, de un modo gradual, un amplio abanico de consultas, pensando, en primer lugar, en las necesidades de los lexicógrafos —y en las de los filólogos o lingüistas en general[13]—. A modo de ejemplo, nos detendremos en una de las funcionalidades (la consulta de las coapariciones de un vocablo), de indudable valor para la redacción de palabras de elevada frecuencia. La pestaña de las coapariciones permite obtener todas las colocaciones, así como restringir la búsqueda de acuerdo con diferentes criterios (clase de palabras del colocativo, grado de probabilidad de que la ocurrencia sea producto del azar, etc.) u ordenar los resultados en función de la medida de asociación estadística preferida.

Así, la consulta de las coapariciones de *emanar* permite obtener una lista de sustantivos con los que se combina este verbo, que, a su vez, se pueden clasificar (con la inestimable ayuda de diccionarios como *Redes*, por ejemplo) en diferentes grupos, agrupaciones que, por su parte, nos permiten atisbar los posibles valores semánticos asociados al verbo: a) en su sentido físico ('desprenderse') figura en combinaciones con sustantivos que designan sustancias volátiles, gases o fluidos (como *perfume, aroma* o *gas*); b) con valor metafórico ('proceder, derivarse, venir originalmente'), se combina con sustantivos que designan atribuciones o

13 En el *Manual de consulta en línea del CDH* (alojado en la pestaña de ayuda del corpus: http://web.frl.es/CNDHE/org/publico/pages/ayuda/ayuda.view) se puede obtener una información pormenorizada sobre los tipos de búsquedas que se pueden efectuar por medio de la interfaz de consulta (desde las más simples —consulta por lema y forma— hasta las que permiten combinar criterios de consulta, obtener las coapariciones, elaborar subcorpus o extraer información estadística).

	Clase ▲	f ▼(2)	MI	LL SIMPLE	T-SCORE
dios	sustantivo	51	4.52	97.85	6.86
o	sustantivo	49	3.45	64.75	6.42
con	sustantivo	49	2.0	31.29	5.57
sus	sustantivo	46	3.58	63.7	6.34
disposición	sustantivo	45	7.89	175.35	6.7
gobierno	sustantivo	42	5.7	108.99	6.48
persona	sustantivo	40	5.12	90.46	6.16
voluntad	sustantivo	39	6.35	116.38	6.24
fuente	sustantivo	37	6.85	121.18	6.08
real	sustantivo	37	5.72	96.79	6.08
cuerpo	sustantivo	37	5.32	87.78	6.08
luz	sustantivo	36	5.97	99.39	6.0
norma	sustantivo	35	8.51	149.39	5.91
cosa	sustantivo	35	3.58	47.88	5.57
fuerza	sustantivo	33	6.08	93.02	5.74
toda	sustantivo	33	5.32	77.96	5.74
pueblo	sustantivo	33	4.95	71.14	5.57
otro	sustantivo	33	2.58	31.2	5.04
estado	sustantivo	32	4.95	68.9	5.48
olor	sustantivo	31	8.08	124.26	5.56

Ilustración 2: Muestra de la consulta de las coapariciones de *emanar* en el *CDH*

posiciones de preeminencia (como *autoridad* o *soberanía*); normas, preceptos y formas de regulación (*norma*, *provisión*); o sentimientos o estados de ánimo (como *tristeza* o *encanto*). A partir de esos listados de coapariciones, se puede acceder a los testimonios del corpus que las atestiguan.

3.4 El corpus debe ser amplio y representativo

Aunque el *NDHE* no es un diccionario sustentado únicamente en un corpus, el *CDH*, con sus tres capas de consulta, constituye el primer recurso documental al que deben recurrir los lexicógrafos; en particular, el *CDH* nuclear, que cuenta con más de 62 millones de ocurrencias, procedentes de 803 textos datados entre el siglo XII y el XXI, suministra la base textual inicial de la que los lexicógrafos extraen los testimonios de las entradas que redactan. En segunda instancia, se recurre a la extensión diacrónica y a la sincrónica del corpus, dado que el *CDH* se ha ampliado para integrar 223 millones de registros procedentes de textos datados entre el siglo XII y 1975 (tomados, a su vez, del *CORDE*), además de otros 123 millones de una tercera capa de consulta (obras fechadas entre 1975 y 2000, provenientes del *CREA*)[14]. Este proceso de unificación no ha estado exento de

14 El diseño inicial del corpus se explica en Pascual y Domínguez (2009); véase también Campos Souto y Pascual (2012: 153–159). La consulta del corpus, en sus tres capas, es accesible en http://web.frl.es/CNDHE/view/inicioExterno.view.

Distribución Periodo

Período	Freq	Fnorm.
1064-1500	67	1,53
1501-1700	265	2,68
1701-1800	12	0,65
1801-1900	396	7,69
1901-2005	485	2,35
1 - 5 of 5		página: 1

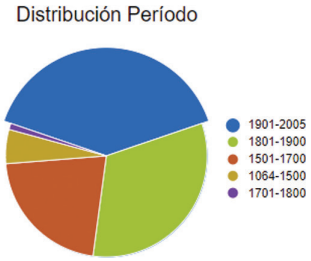


Gráfico 1: Distribución de ocurrencias por período en la extensión diacrónica del *CDH*

problemas, de modo que incluso en la versión actual se pueden detectar incómodas duplicaciones de textos (Pascual 2016: 65), situación que se intentará revertir en la próxima actualización del corpus, prevista para el año 2018.

Tanto el *CDH* nuclear como sus otras dos capas de consulta ofrecen problemas de representatividad[15]; en el diseño del *CDH* nuclear se primó la representación del español de los siglos XIX y XX, por lo que, como explican Pascual y Domínguez (2009), los registros de ambas centurias suponen un 48.73 % del corpus. Pero no es esta una característica exclusiva del *CDH* nuclear: si dirigimos nuestra atención a la capa de textos procedentes del *CORDE*, se observa una notable primacía de los textos del Siglo de Oro, que suponen un 37.8 % de las ocurrencias de esta sección del corpus.

Algunos estudios han puesto de manifiesto otros problemas particulares: así, por ejemplo, Octavio de Toledo y Huerta (2016: 62–63) ha concluido que la etapa comprendida entre 1675 y 1750 ha de considerarse un período «infrarrepresentado» en el *CORDE* que, además, muestra casi en exclusiva la lengua propia de una nómina restringida de autores de referencia (B. J. Feijoo, I. Luzán, G. Mayans y D. de Torres Villarroel); es evidente que la disponibilidad de ediciones de obras de esa etapa ha determinado de manera absoluta la representación de primer español moderno en los corpus. El criterio de la accesibilidad de los textos justifica en buena medida muchos de los problemas observados en los corpus, tanto en lo relativo a su representatividad como en lo referido a su calidad filológica. Por otra parte —como advierte asimismo Octavio de Toledo y Huerta—, la existencia de un canon textual condiciona también la inclusión

15 Para la cuestión de la representatividad de los corpus diacrónicos, véase Torruella (2016).

de las obras en los corpus[16]. En esa constitución del canon desempeña un papel significativo, a partir del siglo XVIII, la condición de académico del autor —o del editor— de una obra, criterio empleado con profusión en los catálogos académicos de autoridades durante el siglo XIX[17]. Si nos detenemos, finalmente, en la clasificación genérica o temática de las obras, se descubren nuevos desequilibrios; véanse, por ejemplo, las observaciones de Pascual con respecto a los textos médicos, en los que se percibe la preponderancia de los «escritos galénico-avicénicos de los siglos XV y XVI» (2016: 62)[18].

Ante estas circunstancias, parece necesario incluir algunos mecanismos compensatorios que permitan obtener unos resultados más ponderados. En ese sentido, la incorporación, en los corpus, de filtros de reducción proporcional de ejemplos (por épocas, géneros, zonas, etc.) podría, quizá, mejorar la representatividad de los datos. Por otro lado, la elaboración de corpus específicos (y su

16 Las consecuencias de la conformación de un determinado canon para la historia del español se han puesto de manifiesto en diversos estudios; véanse, por ejemplo, los trabajos de Fernández-Ordóñez (2006) y Pons Rodríguez (2006).

17 Así, por ejemplo, en el *Catálogo de los escritores que pueden servir de autoridad en el uso de los vocablos y de las frases de la lengua castellana* de la Real Academia Española (Madrid, Imprenta de Pedro Abienzo, 1874), la letra *A* señala la condición de académico de algunos de los autores elegidos; la proporción de académicos en la nómina de escritores se incrementa notablemente en el siglo XIX, en que se menciona, por ejemplo, a P. A. de Alarcón, A. López de Ayala, M. Bretón de los Herreros, R. de Campoamor, S. Catalina, el Duque de Rivas, J. N. Gallego, J. E. Hartzenbusch, A. Lista, A. Oliván, M. J. Quintana, J. Selgas, M. Tamayo y Baus, J. Valera, V. de la Vega y J. Zorrilla.

18 «No hay rastro de la ruptura de la anatomía vesaliana que se da a mediados del siglo XVI, representada por Valverde de Amusco y otros médicos. Tampoco lo hay de la fisiología que cultivan los novatores a finales del XVII. En el siglo XVIII, en el que tan importante es la anatomía española, no aparece ningún texto de esta disciplina y solo uno de medicina, del argentino P. Montenegro, *Materia médica misionera*, sujeto aún a la tradición. Tres son del siglo XIX, de los cuales uno es de Hernández Morejón, introductor de las corrientes europeas en España, pero la obra elegida, "Bellezas de la medicina práctica descubiertas en el Ingenioso caballero don Quijote…", es la menos interesante […]. Para el siglo XX se reservan siete textos, alguno de los cuales no son los más apropiados: es el caso de S. Echevarría, "Organoterapia y Opoterapia", G. Marañón, "Ensayo sobre la vida sexual", J. J. López Ibor, "Las neurosis como enfermedades del alma" y "El libro de la vida sexual". Los ejemplos podrían fácilmente ampliarse al campo jurídico, histórico, etc.» (Pascual 2016: 62–63).

inclusión en otras capas de consulta del corpus) podría solventar las lagunas detectadas en algunos dominios específicos, como el de los documentos[19].

3.5 Los corpus deben ser filológicamente fiables

Aunque la situación ha cambiado de modo significativo en los últimos años, aún muchos estudios sobre la historia del español asumen acríticamente las informaciones ofrecidas por los corpus diacrónicos, sin examinar el grado de calidad o fiabilidad que presentan. En el *CDH* se detectan los mismos problemas y deficiencias, en el plano filológico, que las denunciadas en otros corpus diacrónicos de nuestra lengua[20]. Según Lucía Megías (2008), tres son las principales carencias que presentan los corpus lingüísticos informatizados para poder convertirse en referencia para la filología hispánica:

1. No se indican los criterios que presiden la selección de ediciones que, por otra parte, muestran una heterogeneidad notable, dado que, por mostrar solo los polos del espectro editorial, en los corpus conviven transcripciones fieles de un manuscrito testimonio de una obra con ediciones efectuadas, generalmente en épocas pasadas, con criterios poco fiables[21].
2. La heterogénea presentación gráfica de los textos se alza como un obstáculo insalvable para el estudio diacrónico de los fenómenos fonético-fonológicos.
3. La confusión entre *texto* y *testimonio* conduce a numerosos errores en la fechación de los textos.

En la encrucijada entre la realidad y el deseo se sitúa la solución adoptada en el *NDHE*: la caracterización filológica de los textos, que facilita al lexicógrafo la

---

19 En los artículos publicados del *NDHE*, como *acetra*, puede rastrearse el uso de algunos corpus documentales, como el *CorLexIn* (véase Morala 2014); la nómina de corpus documentales empleados como fuente en el *NDHE* se ha ampliado, de tal modo que en la publicación que se ha efectuado en enero de 2018 de 1057 nuevos artículos —así como en la revisión de los artículos publicados con anterioridad— se puede apreciar la utilización que se efectúa de otros corpus de esta índole, como el *CODEA* o el *CORDEREGRA*.

20 Lucía Megías concluye que «en la relación entre nuevas tecnologías y filología, esta segunda ha quedado relegada a un segundo plano» (2006: 287). Véase también Enrique-Arias (2008), Rodríguez Molina (2010: 608 y 664), Rodríguez Molina y Octavio de Toledo y Huerta (2017) y Rojo (2010), quien advierte sobre las limitaciones del *Corpus del español* de Mark Davies.

21 En otra ocasión (Campos Souto 2016) ya hemos apuntado que sería deseable, además, recuperar el aparato crítico y la anotación de las ediciones críticas, puesto que su supresión impide acceder a una información vital para la adecuada interpretación de los datos.

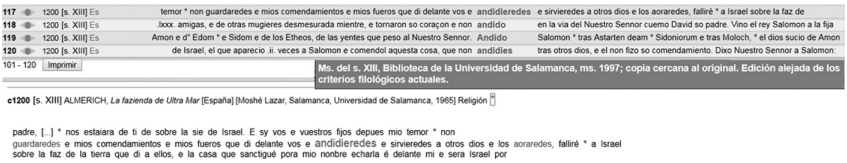


Ilustración 3: Muestra de la breve caracterización filológica de los textos medievales del *CDH* nuclear

información necesaria para valorar lingüísticamente una obra, actúa como un recurso atenuante de los problemas filológicos del corpus. Los primeros frutos de esa tarea pueden observarse en la versión 3.0 de la interfaz de consulta del corpus, publicada en abril de 2015, que permite acceder a una breve caracterización filológica de las obras de la capa medieval del *CDH* nuclear, así como ordenar los testimonios de una voz atendiendo a la fecha del testimonio base en que se apoya la edición o transcripción incorporada al corpus y no solo de acuerdo con la fecha asignada al texto[22].

En esta misma senda se inscriben otras iniciativas que, aunque nacidas posteriormente, discurren en paralelo a esta, como el utilísimo *Cordemáforo*, diseñado por Rodríguez Molina y Octavio de Toledo y Huerta (2017).

Por otra parte, la integración de los materiales en la herramienta de redacción del diccionario —integración que en la actualidad únicamente afecta a las tres capas de consulta del *CDH*— permite, además, aprovechar la experiencia adquirida en el examen de los testimonios, de tal modo que si se descubre que un texto presenta problemas filológicos, ese hecho se consigna en la ficha de nómina correspondiente y, en consecuencia, esa información se muestra en cada uno de los testimonios que se seleccione de esa obra en la redacción de cualquier artículo del *NDHE*.

3.6 La copia digital de los testimonios base (o de los documentos transcritos) debe ser accesible

Sería deseable que los corpus permitiesen acceder a las imágenes de los textos transcritos —bien directamente, bien a través de un enlace—, con el fin de facilitar la consulta directa de las fuentes en aquellos casos en que surja alguna duda sobre las

22 Esta opción de consulta es el fruto del trabajo efectuado, en una primera etapa, por un equipo del CSIC, coordinado por Mariano Quirós; posteriormente, la tarea de caracterización ha recaído en el equipo de lexicógrafos del *NDHE*.

lecturas presentadas en las ediciones o transcripciones incorporadas a estos bancos de datos[23]. En este sentido, el convenio firmado recientemente entre la Real Academia Española y la Biblioteca Nacional de España permitirá, al menos, establecer enlaces con los manuscritos base de las ediciones contenidas en el *CDH*; según la planificación establecida, esta nueva funcionalidad estará disponible en 2019.

4 Fuentes digitales

La irrupción de las bibliotecas y las hemerotecas digitales ha transformado radicalmente la investigación sobre la historia del léxico español, por más que algunos de estos recursos (como, por ejemplo, *Google libros*) se hayan construido con una notable insensibilidad hacia la filología. No obstante, la consulta de estas fuentes es hoy imprescindible en una obra como el *NDHE*; precisamente, con el fin de calibrar el peso de la información suministrada por las bibliotecas y hemerotecas digitales en el *NDHE*, efectuamos un sencillo experimento: analizar la procedencia de los primeros testimonios de las acepciones léxicas que se documentan en nuestro repertorio después del año 1700[24].

Los resultados, obviamente, estaban determinados por el modestísimo tamaño de la muestra objeto de estudio —los 1448 artículos publicados del *NDHE* y las 2545 acepciones léxicas espigadas—, así como el mismo sistema de redacción adoptado en el proyecto, que determina el protagonismo de que gozan, en los artículos publicados, los vocablos relacionados con el léxico de las armas, de las enfermedades, de la indumentaria, de los instrumentos musicales y de medida —y, por consiguiente, de sus respectivas familias léxicas—. Pese a esas restricciones, los datos muestran la relevante aportación de las fuentes digitales a este conjunto de artículos.

4.1 Primeros datos

La consulta realizada en la base de datos arroja un total de 1886 acepciones léxicas con primeros testimonios posteriores al año 1700 (el abanico temporal abarca desde 1703 hasta 2014, fecha en que se registra *antiébola*). La procedencia de estas 1886 primeras documentaciones se distribuye del siguiente modo:

---

23 El *CODEA+2015*, como es bien sabido, se caracteriza por su triple presentación de los textos, dado que facilita el facsímil del documento, así como la edición paleográfica y crítica del texto.

24 Los primeros resultados de ese análisis se presentaron en el *Seminario sobre fuentes digitales*, celebrado en San Millán en octubre de 2016 (Campos Souto, 2018). Descartamos, por lo tanto, las acepciones lexicográficas (es decir, aquellas que solo se atestiguan en diccionarios).

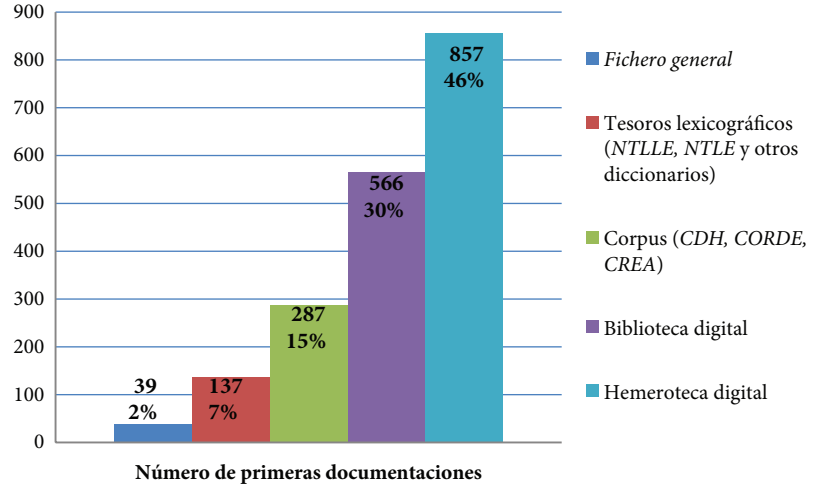


Gráfico 2: Procedencia de los primeros testimonios de las acepciones léxicas posteriores a 1700

La sola lectura de este gráfico ya demuestra la preeminencia de las bibliotecas y las hemerotecas digitales en la provisión de los primeros testimonios de las acepciones analizadas. Pero es evidente que un análisis más demorado permite extraer otras conclusiones no menos interesantes.

4.2 Procedencia *Biblioteca digital*

Si efectuamos una distribución cronológica de los primeros testimonios (566) extraídos de las bibliotecas digitales, podemos apreciar su especial relevancia en los siglos XVIII y XIX[25]:

25 En el *NDHE* se recurre, en primer lugar, a la *Biblioteca digital hispánica* de la Biblioteca Nacional de España, aunque también se emplea profusamente *Google libros*. Ocasionalmente, se consultan fondos procedentes de otras bibliotecas digitales, para comprobar lecturas o datos concretos, pues el hecho de que muchas de ellas no permitan la búsqueda por palabra u ofrezcan un modo de consulta menos ágil y rápido que las mencionadas con antelación (dado que, por ejemplo, impiden la consulta o la descarga de toda la obra y solo permiten ir folio a folio o página a página), las convierte en un recurso de segundo orden.

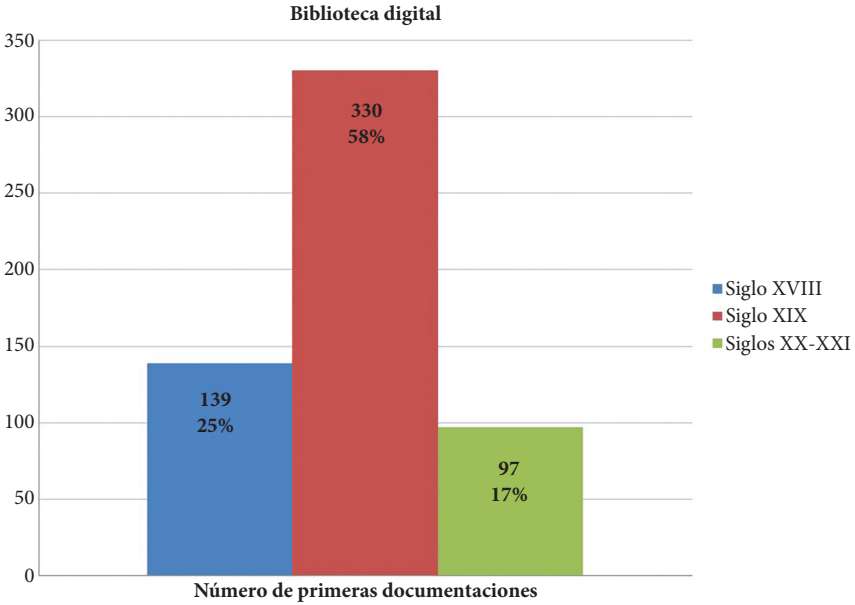


Gráfico 3: Distribución, por siglos, de los primeros testimonios con fuente Biblioteca digital

Pese a su indudable riqueza, la utilización de estas fuentes introduce un notable sesgo a favor de los documentos localizados en España, pues solo ocasionalmente en el siglo XVIII (y, con una intensidad un poco mayor, en el siglo XIX) se recogen documentaciones procedentes de otras áreas del español; únicamente 9 de los 139 testimonios del siglo XVIII provienen de Filipinas, Perú o México (y, en varias ocasiones, para referirse a realidades propias de esos países; así se puede apreciar en los artículos de *lantaca* y *mosquete*)[26]. En el siglo XIX, la proporción no mejora sustancialmente, pues únicamente 36 de los 330 testimonios

26 *Lantaca*, como 'arma de artillería de poco calibre, mayor que el esmeril', se registra por primera vez en 1734, en la *Relación de los sucesos de Mindanao, en las islas Philipinas*; posteriormente figura en obras o crónicas relativas a las Filipinas. *Mosquete*, en la acepción de 'patio interior de un teatro, localizado detrás de las bancadas situadas delante del escenario', propia de México, se atestigua en el *Reglamento u Ordenanzas del teatro* de 1786 y después figura en diversos artículos publicados, ya en el siglo XIX, en el *Diario de México*.

Tabla 1: Muestra de textos científicos y técnicos del siglo XVIII

Autor	Obra	Fecha	Artículos con primeras documentaciones
Boix y Moliner, Miguel Marcelino	*Hippocrates aclarado*	1716	catoco (adj.) hidrofobia
Suárez de Ribera, Francisco	*Febrilogia chyrurgica*	1720	antihidrofóbico, a hidrofóbico, a ('persona que tiene hidrofobia') tetánico
Suárez de Ribera, Francisco	*Arcanismo Antigalico, o Margarita Mercurial*	1721	pulmoníaco, a pulmonario, a
Suárez de Ribera, Francisco	*Tesoro medico, o Observaciones medicinales reflexionadas*	1724	alfanjada
Suárez de Ribera, Francisco	*Escuela medica convincente triumphante, sceptica dogmatica, hija legitima de la experiencia y razon*	1727	hidrofóbico, a ('perteneciente o relativo a la hidrofobia')
Suárez de Ribera, Francisco	*Breviario medico y chyrurgico, de nuevos y raros secretos*	1739	antirreumático, a

se ubican fuera de España. La accesibilidad de los materiales se revela nuevamente —al igual que en el análisis de los corpus— como un elemento crucial en la utilización de las fuentes digitales, una accesibilidad que, sin duda, tiene un efecto no deseado: la sobrerrepresentación de una variante diatópica sobre las otras.

Igualmente esclarecedor resulta el análisis del tipo de obras que facilitan los primeros testimonios; conviene subrayar, en este sentido, el peso de textos fundamentales en la constitución del canon de la ciencia y de la técnica, entre los que se pueden citar, por ejemplo, en el ámbito de la medicina, los debidos a F. Suárez de Ribera o del novator M. M. Boix y Moliner:

Debe señalarse, por otra parte, la trascendencia de las traducciones, que proporcionan con frecuencia los primeros testimonios de los préstamos; entre los títulos tomados de las bibliotecas digitales destacan aquellos que vierten al español obras esenciales para la ciencia y la técnica, escritas originalmente en latín o, con mayor frecuencia, en francés, en los siglos XVIII y XIX; a modo de ejemplo pueden mencionarse las incluidas en la siguiente tabla, correspondientes al siglo XVIII:

Tabla 2: Muestra de traducciones de textos científicos y técnicos del siglo XVIII

Autor	Obra	Fecha	Artículos con primeras documentaciones
Clavijo Fajardo, José	*Traducción de la Historia natural, de Buffon*, I	1785	eudiómetro
Galisteo y Xiorro, Félix	*Traducción del Tratado de las enfermedades venéreas […]. Escrito en idioma latino por Mr. Astruc*. Tomo I	1772	leprosería sífilis
Izuriaga y Ezpeleta, Martín Joseph	*Traducción de la Cirugía completa, de C. Musitano* […]. I	1741	anticatarral
Palau y Verdera, Antonio	*Traducción de la Parte práctica de Botánica, de Carlos Linneo […]*, I	1784	abroquelado, a
Palau y Verdera, Antonio	*Traducción de la Parte práctica de Botánica, de Carlos Linneo […]*, VII	1787	pelta ('órgano esporífero plano y poco prominente en los líquenes')
Piñera y Siles, Bartholomé	*Traducción de los Elementos de medicina práctica, de G. Cullen*, II	1791	antisifilítico, a
Suárez y Núñez, Miguel Gerónimo	*Traducción del Arte de hacer el papel según se practica en Francia y Holanda, en la China y en el Japón […], y del Arte de hacer los cartones, de Mr. de la Lande*	1778	rodela ('pieza de forma circular y plana, con un orificio en el medio, generalmente de metal, que sirve para mantener apretados una tuerca o un tornillo, asegurar los cierres de una junta o evitar el roce de dos objetos') pistola ('instrumento que sirve para mantener un calor constante y moderado en los recipientes en que se fabrica papel')

4.3 Procedencia *Hemeroteca digital*[27]

Una de las señales del cambio de rumbo de la investigación sobre la historia del léxico del español contemporáneo (y, consecuentemente, de la filología) radica precisamente en el recurso a la prensa como una de las más valiosas fuentes de información; si bien en el anterior proyecto de *Diccionario histórico* se emplean profusamente periódicos como el *ABC* para atestiguar voces y acepciones en el siglo XX, la prensa apenas se usa para las centurias anteriores, por obvios motivos de disponibilidad[28]. En el *NDHE*, sin embargo, se recurre de manera regular y sistemática a las publicaciones periódicas, circunstancia que se refleja en el peso de las documentaciones procedentes de las hemerotecas, que supera al de las espigadas en las bibliotecas digitales y experimenta un incremento continuado con el devenir del tiempo; en el siguiente cuadro podemos apreciar esa progresión:

Tabla 3: Número de primeras documentaciones suministradas por las hemerotecas digitales

Hemeroteca digital	Número de primeras documentaciones
Siglo XVIII	49
Siglo XIX	393
Siglos XX-XXI	415

En coherencia con lo observado en la *Biblioteca digital*, la preeminencia de los documentos localizados en España es abrumadora: en el siglo XVIII solo tres testimonios se adscriben a Argentina y a México. La distribución geográfica de la documentación de los siglos XIX, XX y XXI se refleja, a su vez, en la siguiente tabla:

27 La *Hemeroteca digital* de la Biblioteca Nacional de España constituye un recurso de enorme valor para el *NDHE*; pese a no estar lematizada y a la cantidad nada despreciable de lecturas erróneas que arroja (producto de errores de OCR), el hecho de que la búsqueda por palabra permita acceder a un fragmento de texto facilita el trabajo de los lexicógrafos (opción que, en cambio, no está disponible en la *Biblioteca digital hispánica*). Otras hemerotecas de enorme interés, utilizadas con frecuencia en el *NDHE*, son, por ejemplo, la *Hemeroteca digital de México*, *Jable* o la *Biblioteca virtual de prensa histórica*.

28 Entre las excepciones pueden citarse el *Diario de Madrid* (artículo *anascotín*), el *Mercurio histórico y político* (*analizar*, 1764) y, sobre todo, los treinta y cuatro artículos en que se menciona el *Mercurio peruano*, gracias a la existencia de una edición facsimilar de la Biblioteca Nacional de Perú (Lima, 1964–1966).

Tabla 4: Distribución de primeras documentaciones suministradas por las hemerotecas digitales por país y siglo

Países	Siglo xix	Siglo xx	Siglo xxi
Argentina	3	20	2
Chile		1	1
Colombia		1	2
Costa Rica			1
Cuba		1	
El Salvador			1
España	378	360	5
Filipinas	1		
México	7	6	2
Perú		2	1
Uruguay	1		
Venezuela	2		

Por lo que respecta a las características de las publicaciones, en el siglo XVIII puede apreciarse (v. tabla 5) la influencia de los periódicos oficiales, como la *Gazeta de México*, la *Gaceta de Madrid* o el *Mercurio histórico y político*, junto al notable peso de aquellas cabeceras que representan una innovación en el periodismo español y que servirán de puerta de entrada a las novedades que, en el plano cultural y científico-técnico, se vivían en Europa, como el *Diario curioso, erudito y comercial, público y económico* (publicado entre 1758 y 1781).

En el siglo XIX, en cambio, comienzan a adquirir importancia algunas publicaciones especializadas, si bien serán las cabeceras de orientación general las que suministren más testimonios al *NDHE* (v. tabla 6): solo mencionaremos, entre los nuevos títulos de este siglo, *El Eco del Comercio*, *El Heraldo*, *El Clamor Público*, *La Correspondencia de España*, *La Época*, *La Iberia* y *El Guardia Nacional*, en tanto que, a finales de siglo, cobran gran relieve diarios como *La Vanguardia* y el *ABC*, que siguen siendo un recurso de primer orden en las centurias siguientes por su accesibilidad.

En definitiva, este primer análisis sirve para poner de manifiesto algunas de las indudables ventajas que ofrecen las fuentes digitales para el estudio histórico del léxico, unas fuentes que suponen una contribución decisiva al *NDHE*, pero, al tiempo, desvela también algún problema no menor derivado de su utilización. El lujo de contar con bibliotecas y hemerotecas digitales tan ricas como *la Biblioteca digital hispánica* o la *Hemeroteca digital* de la Biblioteca Nacional de España presenta, como reverso de la medalla, algunos inconvenientes, como la carencia de

Tabla 5: Muestra de primeras documentaciones suministradas por la *HD* en el siglo XVIII

Título	Otras denominaciones	Sección	Artículos
Diario Noticioso, Curioso, Erudito y Comercial Público y Económico (Madrid)			broquel ('pendiente, adorno', 1769) guitarra ('arte o técnica de tocar la guitarra', 1758) violín ('arte o técnica de tocar el violín', 1758)
Diario Noticioso, Curioso, Erudito y Comercial Público y Económico (Madrid)		Miguel Terracina, *Traducción de Historia general de los viajes de Prévost*	balafo (1764) gongom^{2} (1764) teodolito (1764)
	Diario de Madrid (Madrid)		cotillería ('establecimiento', 1791) estoqueador (1789) guitarrería ('establecimiento', 1794) oboe ('arte o técnica de tocar el oboe', 1789)
Espíritu de los mejores diarios literarios que se publican en Europa (Madrid)			electrómetro ('parrrayos', 1788) termométrico, a ('perteneciente o relativo al termómetro', 1788) violoncello (1790)
Gaceta de Madrid (Madrid)			encotillado, a (1786)
Gazeta de México (México)			fagot^{2} ('registro de algunos instrumentos de viento', 1794) pestífero, a ('persona que tiene una enfermedad epidémica', 1784)
Mercurio Histórico y Político (Madrid)		Noticias de Francia. El Sr. de Morveau, Fiscal en el parlamento de Borgoña	guardarrayos (1776)
Mercurio Histórico y Político (Madrid)		Noticias de España: Madrid	bombardera ('embarcación', 1783)

Tabla 5: Continúa

Título	Otras denominaciones	Sección	Artículos
Mercurio Histórico y Político (Madrid)			grippe (1775)

ciertos datos relevantes para la valoración de un texto (como la distinción entre fecha de redacción y fecha de edición), la dispersión de los recursos o la sobreabundancia de testimonios localizados en el español europeo, frente al español americano[29]. Por otra parte, la accesibilidad de los materiales condiciona de manera decisiva (en este tipo de bases documentales y, también, en los corpus) la obtención de los testimonios que, en un repertorio como el *NDHE*, sustenta el estudio de la historia de cada palabra. Una accesibilidad que, además, no deja de ser limitada, pues estas fuentes documentales carecen, por lo general, de lematización.

5 Conclusiones

Este rápido recorrido por algunas de las fuentes documentales empleadas en el *NDHE* nos ha permitido apreciar que la insatisfacción ante sus flancos débiles debe actuar (y, de hecho, ha actuado) como un estímulo para su mejora permanente. Sin embargo, conviene reflexionar detenidamente sobre la pluralidad y heterogeneidad de las fuentes que deben manejarse en una obra de este tipo. Algunos intelectuales han alertado últimamente sobre los riesgos derivados del «exceso de información», una sobreabundancia que no parece nueva, pues ya el psicólogo David Lewis, en 1996, creía haber descubierto el síndrome de fatiga informativa (o *Information Fatigue Syndrome*)[30]. Este problema puede trasladarse a nuestro ámbito de trabajo: ¿cuántos recursos diferentes debe consultar un estudioso de la historia del léxico —y, en particular, el redactor de un diccionario

29 Por no mencionar el hecho de que su incremento continuo debido a la digitalización de nuevas obras y publicaciones periódicas –del que debemos felicitarnos como usuarios de estas fuentes– añade un elemento de caducidad casi permanente a cualquier indagación sobre la historia del léxico del español de los últimos tres siglos. Por otra parte, la imposibilidad de acceder a publicaciones sujetas a derechos de autor conduce, en muchas ocasiones, a trazar una historia distorsionada del léxico en los siglos XX y XXI.

30 En la obra *Dying for information? An investigation into the effects of information overload in the UK and worldwide* (Londres, Reuters, 1996).

Tabla 6: Muestra de primeras documentaciones suministradas por publicaciones periódicas generales en el siglo XIX

Título	Artículos
La Dinastía (Barcelona)	antipestoso, a (1896) bugle ('persona que toca el bugle', 1896) contrafagot ('persona que toca el contrafagot', 1896)
El Guardia Nacional (Barcelona)	flageolet (1839)
El Clamor Público (Madrid)	griposo, a (1852) mosquetón[2] ('gancho', 1849) violonchelista (1844)
La Correspondencia de España (Madrid)	antipulmonar (1877) helicón ('persona que toca el helicón', 1875) xilófono (1867) xilofonista (1882)
El Eco del Comercio (Madrid)	afusilamiento (1834) fusilador, a (1839) pulmonista ('persona que grita mucho', 1843) sablear ('atacar o acometer [a alguien] con un sable', 1838)
La Época (Madrid)	balafón (1863) corselete ('prenda femenina que ciñe el talle', 1861) panderetazo (1859) sarampiónico, a ('persona que tiene sarampión', 1893)
El Heraldo (Madrid)	bombardeador, a (1842) concertina (1854) saxófono (1846) violonchelo ('persona que toca el violonchelo', 1844)
La Iberia (Madrid)	ametrallador, a ('que ametralla', 1854) bongó (1889) diftérico, a ('que causa difteria', 1894) neuropulmonar (1885) zambombazo ('noticia impactante', 1876)

histórico—? ¿Cuántos corpus, tesoros lexicográficos, hemerotecas o bibliotecas digitales diferentes han de pasar por sus manos para evitar el riesgo de perder algún dato relevante? ¿Cómo se puede contribuir a la mejora de las condiciones de trabajo de los historiadores del léxico español? ¿Internet se ha convertido solo en un trasunto digital de los recursos tradicionales en papel, diseminados por todo el mundo? ¿Cómo podemos frenar esa dispersión e ir hacia bancos de datos unificados, flexibles y potentes? ¿Es posible comenzar a caminar, con paso firme,

hacia plataformas o hacia agregadores de contenidos filológicos o lexicográficos? Quizá hoy no tengamos todas las respuestas, ni sepamos trazar todavía el mapa que nos guíe por una ruta que presumimos accidentada pero, si sabemos hacia dónde queremos ir, quizá estemos más cerca de averiguar cómo podemos llegar.

Referencias bibliográficas

Béjoint, Henri (2007): «Informatique et lexicographie de corpus: les nouveaux dictionnaires», *Révue française de linguistique appliquée* XII/1, 7–23.

Campos Souto, Mar (2016): «Lexicografía del futuro para la lengua del pasado», en Rosalía Cotelo (coord.), *Entre dos coordenadas: la perspectiva diacrónica y diatópica en los estudios léxicos del español*. San Millán de la Cogolla: Cilengua, 33–71.

Campos Souto, Mar (2017): «Hacia una crónica del *Diccionario histórico* de la lengua española de 1933–1936: Los materiales del Archivo de la Real Academia Española», *BRAE* XCVII/CCCXV, 161–201.

Campos Souto, Mar (2018): «Bibliotecas y hemerotecas digitales en el *NDHE*», *Cuadernos del Instituto de Historia de la Lengua*, 11, 237–255.

Campos Souto, Mar/José A. Pascual (2012): «Lexicografía, filología e informática: una alianza imprescindible. A propósito de la situación del *NDHE*», en Dolores Corbella *et al.* (eds.), *Lexicografía hispánica del siglo XXI: Nuevos proyectos y perspectivas. Homenaje al profesor Cristóbal Corrales Zumbado*. Madrid: Arco/Libros, 151–170.

CDH = Instituto de Investigación Rafael Lapesa de la Real Academia Española (2013): *Corpus del Nuevo diccionario histórico (CDH)* [en línea]. <http://web.frl.es/CNDHE> [último acceso: 10/09/2017].

CODEA+2015 = *GITHE* (Grupo de Investigación Textos para la Historia del Español): *CODEA+ 2015* (*Corpus de documentos españoles anteriores a 1800*) [en línea]. <http://corpuscodea.es/> [último acceso: 10/08/2017].

CORDE = Real Academia Española: Banco de datos (*CORDE*) [en línea]. *Corpus diacrónico del español*. <http://www.rae.es> [último acceso: 10/07/2017].

CORDEREGRA = Calderón Campos, Miguel/M.ª Teresa García Godoy (dirs.): *Corpus diacrónico del español del Reino de Granada*. 1492–1833. <http://www.corderegra.es> [último acceso: 10/02/2016].

CorLexIn = Morala Rodríguez, José R. (dir.): *Corpus Léxico de Inventarios (CorLexIn)*. <http://web.frl.es/CORLEXIN.html> [último acceso: 10/09/2017].

CREA = Real Academia Española: Banco de datos (*CREA*) [en línea]. *Corpus de referencia del español actual*. <http://www.rae.es> [último acceso: 10/08/2017].

Enrique-Arias, Andrés (2008): «Biblias romanceadas e historia de la lengua», en Concepción Company/José Moreno de Alba (eds.), *Actas del VII Congreso Internacional de Historia de la Lengua Española. Mérida (Yucatán), 4–8 de septiembre de 2006*. Madrid: Arco/Libros, 1781–1794.

Fernández-Ordóñez, Inés (2006): «La Historiografía medieval como fuente de datos lingüísticos. Tradiciones consolidadas y rupturas necesarias», en José Jesús de Bustos Tovar y José Luis Girón Alconchel (eds.), *Actas del VI Congreso Internacional de Historia de la Lengua Española (Madrid, 29 de septiembre-3 de octubre de 2003)*. Madrid: Arco/Libros, II, 1779–1807.

Hanks, Patrick (2012): «The impact of corpora on dictionaries», en Paul Baker (ed.), *Contemporary Corpus Linguistics*. Londres-Nueva York: Continuum, 214–236.

Kilgarriff, Adam (2013): «Using corpora [and the web] as data sources for dictionaries», en Howard Jackson (ed.), *The Bloomsbury Companion to Lexicography*. Londres: Bloomsbury, 77–96.

Kilgarriff, Adam *et al.* (2014): «The Sketch Engine: Ten years on», *Lexicography: Journal of ASIALEX* 171, 7–36.

Lucía Megías, José Manuel (2006): «Informática textual: nuevos retos para la edición y difusión de los textos (bibliotecas virtuales y bancos de datos textuales)», en Ramón Santiago, Ana Valenciano y Silvia Iglesias (eds.), *Tradiciones discursivas: Edición de textos orales y escritos*, Madrid: Universidad Complutense de Madrid, 251–302.

Lucía Megías, José Manuel (2008): «El hipertexto ante el reto de los textos medievales: nuevas reflexiones sobre informática humanística», en Aurelio González y Lilian von der Walde Moreno (eds.), *Textos, motivos y contextos medievales*. México: El Colegio de México-Universidad Autónoma de México, 9–14.

Morala, José Ramón (2014): «El *CorLexIn*, un corpus para el estudio del léxico histórico y dialectal del Siglo de Oro», *Scriptum Digital* 3, 5–28.

NTLLE = Real Academia Española (2001): *Nuevo tesoro lexicográfico de la lengua española*. <http://ntlle.rae.es/ntlle/SrvltGUISalirNtlle> [último acceso: 15/09/2017].

Octavio de Toledo y Huerta, Álvaro S. (2016): «El aprovechamiento del *CORDE* para el estudio sintáctico del primer español moderno (ca. 1675–1825)», en Johannes Kabatek y Carlota de Benito (eds.), *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín/Boston: Walter de Gruyter, 57–89.

Pascual, José A. (2016): «La Filología en vago y en vilo entre los datos», en Emilio Blanco (ed.), *Grandes y pequeños de la literatura medieval y renacentista*. Salamanca: Ediciones del *SEMYR*, 55–84.

Pascual, José A./Carlos Domínguez (2009): «Un corpus para un nuevo diccionario histórico del español», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid/Fráncfort: Iberoamericana/Vervuert, 79–33.

Pons Rodríguez, Lola (2006): «Canon, edición de textos e historia de la lengua cuatrocentista», en Lola Pons (ed.), *Historia de la lengua y crítica textual*. Madrid/Fráncfort: Iberoamericana/Vervuert, 69–125.

Rafel i Fontanals, Joaquim (2011): «Lexicografía e informática. Aplicación a la lengua catalana», en *Pirinioetako hizkuntzak: oraina eta lehena*, Bilbao: Euskaltzaindia, 557–575.

Redes = Bosque, Ignacio (dir.) (2004): *Redes: Diccionario combinatorio del español contemporáneo*. Madrid: SM.

Rodríguez Molina, Javier (2010): *La gramaticalización de los tiempos compuestos en español antiguo: cinco cambios diacrónicos* (tesis doctoral). Madrid: Universidad Autónoma de Madrid.

Rodríguez Molina, Javier/Álvaro Octavio de Toledo y Huerta (2017): «La imprescindible distinción entre texto y testimonio: el *CORDE* y los criterios de fiabilidad lingüística», *Scriptum Digital* 6, 5–68.

Rojo, Guillermo (2009): «Sobre la construcción de diccionarios basados en corpus», *Tradumática* 7. <http://www.fti.uab.cat/tradumatica/revista/num7/articles/02/02.pdf>.

Rojo, Guillermo (2010): «Sobre codificación y explotación de corpus textuales: otra comparación del *Corpus del español* con el *CORDE* y el *CREA*», *Lingüística* 24, 11–50.

Salas Quesada, Pilar/Abelardo Torres Morcillo (2011): «*ARDIDEs*: Aplicación de Redacción de un Diccionario Diacrónico del Español», *Revista de lexicografía* XVII, 133–159.

Salas Quesada, Pilar/Abelardo Torres Morcillo (2015): «Aproximación a los fundamentos del *NDHE* a través de las herramientas informáticas usadas en su elaboración y presentación», *Estudios de lexicografía* 3, 15–69 [accesible en <http://www.cilengua.es/sites/cilengua.es/files/page/docs/2015_monografico_ndhe_rae.pdf>].

Seco, Manuel (1980): *Las palabras en el tiempo: los diccionarios históricos*. Madrid: Real Academia Española.

Torruella, Joan (2016): «Tres propuestas en el ámbito de la lingüística de corpus», en Johannes Kabatek y Carlota de Benito (eds.), *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín/Boston: Walter de Gruyter, 90–112.

Mark Davies

Uso del *Corpus del Español* y los corpus relacionados para la lexicografía histórica española

Resumen: El *Corpus del Español* original de 100 millones de palabras (siglos XIII–XX) permite a los investigadores llevar a cabo investigaciones avanzadas que incluyen colocativos, n-gramas, sinónimos, así como comparaciones a través de períodos históricos, que no son posibles con otros corpus como *CORDE*. Además, los datos del siglo XX permiten comprender la variación en el léxico basada en el género textual, lo que no es posible con otros corpus como *CREA*. En 2016, el *Corpus del Español* se amplió considerablemente para incluir dos mil millones de palabras de 20 países de habla hispana, lo que permite que la investigación sobre la variación regional del léxico sea mucho más profunda que con otros corpus como *CORPES*. Finalmente, el próximo corpus *NOW-Spanish* (que se lanzará a finales de 2018), con seis mil millones de palabras en español, es completamente original en la forma en que los investigadores pueden rastrear los cambios léxicos en curso (incluso semana por semana).

Palabras clave: Corpus Léxico, Histórico, Géneros, Dialecto

Abstract: The original 100 million word *Corpus del Español* (1200s–1900s) allows researchers to carry out advanced research involving collocates, n-grams, synonyms —as well as comparisons across historical periods— in ways that are not possible with other corpora such as *CORDE*. In addition, the data from the 1900s allows insight into genre-based variation in lexis, which is not possible with other corpora like *CREA*. In 2016 the *Corpus del Español* was greatly enlarged to include two billion words of data from 20 Spanish-speaking countries, and it allows research on regional variation in lexis that is much more powerful than with other corpora like *CORPES*. Finally, the upcoming *NOW-Spanish* corpus (to be released in late 2018) is completely unique in the way that researchers can track ongoing changes in lexis (even week-by-week) in a six billion-word corpus of Spanish.

Keywords: Corpus, Lexis, Historical, Genres, Dialect

1 Introducción

Hasta hace unos veinte años, la lexicografía histórica solía llevarse a cabo mediante minuciosos y laboriosos análisis de textos impresos. Sin embargo, a partir de la llegada de grandes corpus en línea, la tarea de quienes estudian el

cambio y la variación léxica se ha convertido en algo mucho más fácil. Los investigadores pueden lograr ahora en solo unos minutos lo que antes podría haber necesitado días o incluso semanas.

En este trabajo, consideraremos con cierto detalle cómo el *Corpus del Español* se puede utilizar para examinar la variación y el cambio léxico[1]. Nos centraremos en las dos partes del corpus: el corpus de 100 millones de palabras «histórico/basado en el género» que se lanzó en 2002, y el corpus de dos mil millones de palabras «basado en web/dialectal» que se lanzó en 2016. También trataremos brevemente un corpus de cinco mil millones de palabras que se lanzará en 2018, y que permitirá a los investigadores examinar los neologismos y el cambio léxico actual con increíble detalle. A medida que analicemos estos diferentes corpus, también trataremos brevemente los corpus similares de la Real Academia Española: *CORDE* (textos históricos), *CREA* (principalmente textos de finales del siglo XX) y *CORPES* (textos de principios del siglo XXI).

2 Los corpus textuales

El *Corpus del Español* se compone de dos partes diferentes, ambas diseñadas para dar respuesta a diferentes preguntas sobre el cambio y la variación del lenguaje. La primera parte es el corpus «histórico/género». Se completó en 2002 y fue revisado a fines de 2007. Está compuesto por aproximadamente 100 millones de palabras, desde el español antiguo hasta finales del siglo XX, tiene aproximadamente 18 millones de palabras del siglo XIII al XV, 42 millones de palabras del siglo XVI al XVIII, y alrededor de 40 millones de palabras de los siglos XIX y XX. Está compuesto de textos de una amplia gama de géneros, incluidos más de cinco millones de palabras de transcripciones de conversaciones habladas de finales del siglo XX. Para el siglo XX, están divididos en partes iguales entre los siguientes géneros: hablado, ficción, prensa y académico. Los detalles completos sobre cada uno de los casi 14 000 textos se pueden encontrar a través del enlace «Textos» en el sitio web del corpus, y los usuarios pueden descargar un archivo de Excel que enumera todos esos textos.

La segunda parte del *Corpus del Español* fue lanzada en 2016. Está orientada a estudios sobre la variación dialectal en español o que necesitan un corpus mucho más grande que el corpus original de 100 millones de palabras. El corpus contiene más de dos mil millones de palabras, lo que lo hace 100 veces más grande que la porción del siglo XX del corpus original. Además, los usuarios pueden

1 Para estudios similares sobre el inglés, vid. Davies en prensa a y b.

realizar comparaciones entre los 21 países del corpus (España, México, Colombia, Argentina, etc.) para ver la frecuencia por dialecto y buscar directamente todas las palabras que son más frecuentes en un país que en otro.

En este artículo, compararemos la parte histórica (y basada en el género) del *Corpus del Español* con el *CORDE* (*Corpus diacrónico del español*) y el *CREA* (*Corpus de referencia del español actual*) de la Real Academia Española. A continuación, en las Secciones 3–11, compararemos la funcionalidad de la parte histórica del *Corpus del Español* con el *CORDE*, por lo que daremos una visión general del *CORDE*.

El *CORDE* se creó a finales de los años 90 del siglo XX y fue el primer gran corpus histórico del español. Está compuesto por aproximadamente 250 millones de palabras de texto, con buena representación a lo largo de los diferentes períodos históricos, y un buen equilibrio entre géneros, que incluyen poesía, escritos históricos, literatura, materiales didácticos, etc. Con 250 millones de palabras, es de dos a tres veces más grande que el *Corpus del Español*. Sin embargo, como veremos, el tamaño no es todo. Sin el tipo correcto de arquitectura e interfaz de corpus, los datos textuales están, en esencia, «atrapados» dentro del corpus y no están disponibles para los usuarios finales.

3 Descripción general del uso de corpus para la lexicografía histórica

Los corpus históricos, para ser útiles, deberían permitir a los usuarios realizar investigaciones sobre varios aspectos diferentes del cambio léxico[2]. Estos incluyen lo siguiente:

- Encuentra resultados representativos. En el nivel más básico, los usuarios pueden buscar una palabra o frase, encontrar la primera aparición de la palabra o frase y ver todas las ocurrencias en contexto.
- Frecuencia (más avanzada). Los usuarios pueden ver fácilmente la frecuencia de una palabra o frase a lo largo del tiempo, con frecuencias normalizadas[3].
- Frecuencia (aún más avanzada). En lugar de tener que decirle al corpus qué palabras o frases específicas buscar, el corpus puede generar una lista de palabras cuya frecuencia coincide con ciertos criterios, como los nombres que

2 Para estudios similares con corpus históricos del inglés, vid. Davies 2012a y 2012b.

3 En otras palabras, frecuencia por millar o por millón de palabras de texto, para tener en cuenta el diferente tamaño del corpus en diferentes períodos históricos.

se incorporaron al léxico en el siglo XVII o todas las palabras que se usan al menos cinco veces más en el siglo XIII que en el XIV

- Forma de palabra (morfológica). Los usuarios deben poder buscar por prefijos, sufijos y raíces, y ver la frecuencia de cada forma coincidente en los diferentes períodos históricos, así como la frecuencia general de todas las formas en cada período histórico.
- Significado de la palabra (semántica): simple. Los usuarios pueden encontrar las colocaciones más frecuentes (palabras cercanas) de una palabra o frase determinada, lo que obviamente proporciona una muy buena comprensión del significado de la palabra. Prácticamente cualquier arquitectura e interfaz de corpus permite a los usuarios ver las palabras cercanas caso por caso, pero los corpus realmente útiles resumen toda esta información sobre las colocaciones para todas las ocurrencias de una palabra o frase determinada.
- Semántico (más avanzado). Suponiendo que el corpus pueda encontrar colocaciones, debería ser posible compararlas a través de períodos históricos o entre diferentes géneros. Los cambios en las colocaciones a lo largo de períodos sirven a menudo como marcadores de cambio semántico.
- Semántica (aún más avanzada). En lugar de simplemente buscar palabras y frases, los usuarios pueden buscar por campo semántico. P. ej., si un repertorio está integrado en el corpus, o si los usuarios pueden crear listas personalizadas de palabras, entonces podrían crear una búsqueda donde cualquier palabra en un campo semántico es parte de la consulta. Un ejemplo de esto podría ser [miembro de la familia] seguido de [sinónimo de *pedir*] seguido de [sinónimo de *limpiar*], o [hora del día] cerca de [sinónimo de *lúgubre*]. Del mismo modo, se podría comparar la frecuencia de todas las palabras o frases en un campo semántico completo, y comparar la frecuencia y la distribución de cada miembro a lo largo del tiempo.

En los siguientes apartados, proporcionaré ejemplos concretos de cómo estos dos grandes corpus históricos, el *Corpus del Español* y el *CORDE,* pueden usarse (o no) para investigar la amplia gama de fenómenos enumerados anteriormente. (Para obtener una visión general previa de la funcionalidad del *Corpus del Español* con arquitectura e interfaz más antiguas, v. Davies 2002, 2005a, 2005b, 2008, 2010). Algunos lectores pueden, así, comenzar a obtener una perspectiva completamente nueva de lo que se puede hacer con los corpus históricos. Si han utilizado corpus con arquitecturas e interfaces limitadas, quizá los hayan usado solo para encontrar las ocurrencias de una palabra o frase específica. Sin embargo, una vez que una persona ha utilizado un corpus de los que permiten una amplia gama de consultas, se da cuenta de que hay inmumerables temas de lingüística histórica que podrían estudiarse con un corpus completo.

4 Encontrar y mostrar resultados representativos

En el nivel más básico, un corpus debería permitir al investigador buscar una palabra o frase, encontrar la primera ocurrencia de la palabra o frase o ver todas los resultados en contexto (y quizás limitar las ocurrencias a un periodo histórico determinado). Los programas para realizar tales búsquedas son abundantes y bastante rápidos: 1–2 segundos (en el caso de la búsqueda simple) incluso para un corpus de 100 millones de palabras.

El *CORDE* puede hacer estas búsquedas básicas bastante bien. P. ej., supongamos que el usuario quiere encontrar todas las apariciones de la palabra *braueza*. Después de enviar la búsqueda, el usuario ve que hay 273 ocurrencias en 86 documentos. Al hacer clic en «Obtención de ejemplos», el usuario ve entradas de «Palabra clave en contexto» (KWIC, *Keyword in Contex*t) como las siguientes:

CONCORDANCIA			AÑO
a que de suso dicha es mouiendose con saña: o con	brauezа: o con mal querençia como quiera que pena çi	**	1491
ixo el rey salamon atal es la yra del rey como la	brauezа del leon que ante el su bramido todas las otr	**	1491
que ge las deuen por si mismos. Llanamente & syn	brauezа ninguna deuen los onbres vnos a otros demanda	**	1491
sto no serie derecha si la diesse con sanna o con	brauezа por malquerencia que ouiesse contra el. E por	**	1256 - 1263
tra esta ley, & fuer cruel contra sus pueblos por	brauezа o por cobdicia o por auaricia, sea descomulga	**	c 1250 - 1260
verdat & Apuestas & ssin tuerto de njnguno & ssin	brauezа assi como ssi lo ouiessen a ffablar en conçei	**	a 1260
trar al Rey deue lo ffazer omjldosa mjente & ssin	brauezа. Et otrossi non deue denostar njn Amenazar A	**	a 1260
diçiendole que aquel era el que mas temian por su	brauezа segun que se muestra en algunas pinturas era	**	1661
ridad las estrellas, Sin escuridad la sombra. Sin	brauezа los leones, A los tigres qual palomas, Sin fo	**	1598

Gráfico 1: Palabra clave en la visualización de contexto con el *CORDE*

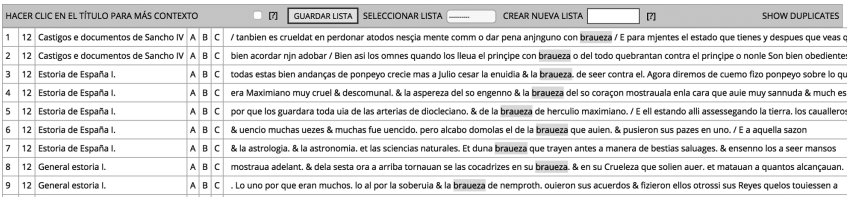
Sin embargo, una restricción importante a los usuarios del *CORDE* es que limita la visión de la palabra en contexto solo cuando aparece menos de mil veces en el corpus. De esta manera, cuando se encuentran miles de palabras no hay una manera fácil de verlas en contexto.

Una búsqueda básica de palabras y frases simples funciona de manera similar con el *Corpus del Español*. P. ej., después de hacer la búsqueda, el usuario verá lo siguiente:

		CONTEXTO	TODOS	s13	s14	s15	s16	s17	s18	s19	s20
1		BRAUEZA	144	48 7.15	32 11.99	27 3.31	32 1.88	5 0.41			

Gráfico 2: Listado de frecuencia con el *Corpus del Español*

Se muestra aquí la frecuencia absoluta de la palabra en cada siglo (p. ej., 32 ocurrencias en el siglo XIV), así como la frecuencia normalizada por siglo (p. ej.,



HACER CLIC EN EL TÍTULO PARA MÁS CONTEXTO [?] GUARDAR LISTA SELECCIONAR LISTA ..... CREAR NUEVA LISTA [?] SHOW DUPLICATES

1	12	Castigos e documentos de Sancho IV	A	B	C	/ tanbien es crueldat en perdonar atodos nesçia mente comm o dar pena anjnguno con braueza / E para mjentes el estado que tienes y despues que veas q
2	12	Castigos e documentos de Sancho IV	A	B	C	bien acordar njn adobar / Bien asi los omnes quando los lleua el prinçipe con braueza o del todo quebrantan contra el prinçipe o nonle Son bien obedientes
3	12	Estoria de España I.	A	B	C	todas estas bien andanças de ponpeyo crecie mas a Julio cesar la enuidia & la braueza. de seer contra el. Agora diremos de cuemo fizo ponpeyo sobre lo qu
4	12	Estoria de España I.	A	B	C	era Maximiano muy cruel & descomunal. & la aspereza del so engenno & la braueza del so coraçon mostrauala enla cara que auie muy sannuda & much es
5	12	Estoria de España I.	A	B	C	por que los guardara toda uia de las arterias de diocleciano. & de la braueza de herculio maximiano. / E ell estando alli assessegando la tierra. los caualleros
6	12	Estoria de España I.	A	B	C	& uencio muchas uezes & muchas fue uencido. pero alcabo domolas el de la braueza que auien. & pusieron sus pazes en uno. / E a aquella sazon
7	12	Estoria de España I.	A	B	C	& la astrologia. & la astronomia. et las sciencias naturales. Et duna braueza que trayen antes a manera de bestias saluages. & ensenno los a seer mansos
8	12	General estoria I.	A	B	C	mostraua adelant. & dela sesta ora a arriba tornauan se las cocadrizes en su braueza. & en su Crueleza que solien auer. et matauan a quantos alcançauan.
9	12	General estoria I.	A	B	C	. Lo uno por que eran muchos. lo al por la soberuia & la braueza de nemproth. ouieron sus acuerdos & fizieron ellos otrossi sus Reyes quelos touiessen a

Gráfico 3: Pantalla de palabras clave en contexto con el *Corpus del Español*

un pico de alrededor de 12 ocurrencias por millón de palabras en el siglo XIV). Además, uno puede ver la palabra clave en contexto para cualquier palabra, no solo aquellas con baja frecuencia (como es el caso del *CORDE*).

En el *Corpus del Español*, al hacer clic en los números en cualquier columna se mostrará la palabra clave en el contexto de ese siglo, o se pueden ver todas las entradas a la vez haciendo clic en TOTAL. Luego se puede hacer clic en otras entradas para ver el contexto extendido (aproximadamente 200 palabras en total).

La visualización de palabra clave en contexto en el *Corpus del Español* permite algunas funcionalidades importantes que no son posibles con el *CORDE*. Primero, los usuarios pueden guardar ocurrencias (en tres grupos diferentes de ocurrencias) haciendo clic en A, B o C. Posteriormente, pueden organizar estas listas de ocurrencias (lo que incluye moverlas entre diferentes listas), lo que facilita en gran medida el almacenamiento y análisis de datos.

Hasta este punto, entonces, las búsquedas en los dos corpus son bastante similares. El *CORDE* tiene la ventaja de ser el corpus más grande, mientras que el *Corpus del Español* tiene la ventaja de mostrar la frecuencia en cada siglo y de mostrar la palabra clave en contexto para todas las palabras, independientemente de la frecuencia.

5 Frecuencia de palabra: datos básicos

Sin embargo, además de obtener simplemente todas las apariciones de una palabra o frase determinada, los usuarios a menudo quieren saber cuán frecuente eran en diferentes siglos o en determinados períodos históricos. Es en este punto donde el *CORDE* comienza a mostrar algunas debilidades serias. P. ej., después de buscar *braueza* y luego seleccionar «Ver estadística», el usuario visualiza:

Estadísticas

Año	%	Casos
1627	20.95	35
1547	20.35	34
1610	17.96	30
1632	8.98	15
1566	4.79	8

País	%	Casos
ESPAÑA	86.44	236
MÉXICO	10.98	30
PERÚ	1.46	4
BOLIVIA	0.36	1
COLOMBIA	0.36	1

Tema	%	Casos
22.- Verso narrativo	26.37	72
12.- Prosa narrativa	20.51	56
19.- Prosa histórica	17.58	48
21.- Verso lírico	12.45	34
16.- Prosa de sociedad	7.69	21

Gráfico 4: Frecuencia (por año) con el *CORDE*

Esta tabla nos dice los años específicos en los cuales la palabra o frase es más común, pero es imposible ver la frecuencia por década o por siglo. No sirve de nada mostrar que la palabra fue la más frecuente en 1627 si, de hecho, es mucho menos común en el siglo XVII que en el XIII o el XIV. El otro problema importante es que las cifras no están relativizadas. En otras palabras, vemos la frecuencia absoluta por año, pero una palabra o frase puede ser más común en ese año simplemente porque hay más palabras para ese año en el corpus. Cualquier comparación seria de frecuencia requiere que los resultados se relativicen a lo largo de períodos históricos para que podamos tener en cuenta los diferentes tamaños del corpus en diferentes periodos históricos, y ver la frecuencia de la palabra o frase por millón de palabras.

El *Corpus del Español* permite este tipo de búsqueda con bastante facilidad. P. ej., con *braueza* en el *Corpus del Español*, podemos ver la «visualización de tabla» (como en la tabla 2 anterior), o una visualización en gráfico:

SECCIÓN	s13	s14	s15	s16	s17	s18	s19	s20
OCURRENCIAS	48	32	27	32	5	0	0	0
POR MILL	7.15	11.99	3.31	1.88	0.40	0.00	0.00	0.00

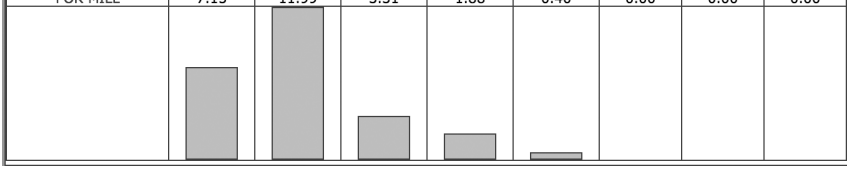


Gráfico 5: *Corpus del Español*: Frecuencia de palabra y frase por siglo

Se nos muestra aquí la frecuencia absoluta (p. ej., 32 ocurrencias en siglo XIV), así como la importante frecuencia relativa («por millón»), que tiene en cuenta el tamaño de la sección en millones de palabras. P. ej., hay 32 ocurrencias en los 3 millones de palabras del siglo XIV, o 10.8 ocurrencias por millón de palabras. Un gráfico como este es la única manera de ver realmente los cambios en la frecuencia de una palabra, frase o construcción, y solo es posible con el *Corpus del Español*.

6 Frecuencia de palabra: comparación de períodos históricos

En lugar de tener que decirle al corpus qué palabras o frases específicas buscar, un corpus con arquitectura e interfaz bien diseñadas generaría una lista de palabras cuya frecuencia coincide con ciertos criterios. P. ej., podría encontrar todos los sustantivos que entraron en la lengua del siglo XVII o las palabras que se usaron al menos cinco veces más en el siglo XIII que en el XIV. Tal consulta es completamente imposible con el *CORDE*. Todo lo que se puede hacer es buscar palabras y frases específicas. Si el *CORDE* reconoce la frecuencia de todas las palabras y frases en todos los períodos históricos, ciertamente no permite a los investigadores usar esa información mediante una consulta.

Con el *Corpus of Spanish*, por otro lado, tales consultas son bastante sencillas. P. ej., se puede buscar simplemente [nn *] (sustantivos) y seleccionar [siglo XIII] (1200–1299) «SECCIÓN 1» para comparar con [siglo XIV] (1300–1399) «SECCIÓN 2». En uno o dos segundos, el usuario ve la siguiente lista[4]:

Tabla 1: *Corpus del Español*. Comparación de la frecuencia de palabras por siglo (todas las palabras a la vez)

Siglo XIII	Siglo XIV
capitolo, ascendente, ladeza, saturnus, orizon, morauedis, roque, xaque, acendent, armella, murcia, baldouin, ascendent, segonda, gudufre, significador, fferrando, corualan, alffil, zonte, dond, iudga, tiemplo, boymonte, catamiento, infortunas, uinie, caput, camyaron, sacrifficio, declinacion, sacrifficios, yguador, juppiter, algarue, linnas, deillos, hierusalem, decima.	armadas, osso, ome, avia, paris, ahe, collado, camjno, hercoles, ley, çima, rrayzes, armada, elena, falcon, yuierno, verano, gonçales, encarnaçion, pase, bjen, venja, avras, falcones, jnfante, façer, puerco, ynfanta, ynfante, ynperio, vyno, venjdo, sembrar, fojas,ençima, talante, mjel, menalao, syenpre, dolençia, ssiete, avedes, castilla, muria, aujdo, peça, arroyo, vuas, çiençia, termjno, tenjan.

4 Téngase en cuenta que en la versión web hay frecuencias –en bruto y normalizadas– para cada palabra, así como enlaces para ver la palabra en contexto, como se muestra en el Gráfico 2. En la Tabla 1, hemos simplificado la visualización.

SEC 1 (1900s): 22,822,256 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	SECTOR	2540	0	111.3	0.0	11,129.5
2	TELEVISIÓN	2119	0	92.8	0.0	9,284.8
3	SECTORES	1360	0	59.6	0.0	5,959.1
4	FÚTBOL	1208	0	52.9	0.0	5,293.1
5	LÍDER	1119	0	49.0	0.0	4,903.1
6	PROTEÍNAS	753	0	33.0	0.0	3,299.4
7	IMPACTO	746	0	32.7	0.0	3,268.7
8	AEROPUERTO	721	0	31.6	0.0	3,159.2
9	NARCOTRÁFICO	705	0	30.9	0.0	3,089.1
10	INFLACIÓN	704	0	30.8	0.0	3,084.7
11	LÍDERES	666	0	29.2	0.0	2,918.2
12	CAMPEONATO	657	0	28.8	0.0	2,878.8

SEC 2 (1800s): 19,297,249 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	VENTURA	1237	25	64.1	1.1	58.5
2	NÚM	1025	21	53.1	0.9	57.7
3	RODRIGO	697	19	36.1	0.8	43.4
4	VULGO	662	20	34.3	0.9	39.1
5	PAJE	611	21	31.7	0.9	34.4
6	MATILDE	669	24	34.7	1.1	33.0
7	APOSENTO	1174	44	60.8	1.9	31.6
8	HONRA	1652	63	85.6	2.8	31.0
9	MENESTER	1545	60	80.1	2.6	30.5
10	MARGARITA	554	23	28.7	1.0	28.5
11	ERMITAÑO	553	23	28.7	1.0	28.4
12	CONDESA	1829	78	94.8	3.4	27.7

Gráfico 6: *Corpus del Español*: Comparación de la frecuencia de las palabras (sustantivos en siglo XIX/siglo XX)

Obviamente, solo algunas de las palabras de esta lista son significativas. Muchas palabras son simplemente variantes ortográficas, otras son sustantivos propios que pueden aparecer en un puñado de textos de un siglo, pero no de otro.

Por supuesto, tales búsquedas no están limitadas solo al español antiguo, también se pueden llevar a cabo para períodos históricos más recientes. La siguiente tabla muestra (a la izquierda) nombres que son comunes en el siglo XX, pero no en el XIX y (a la derecha) aquellos que son comunes en el siglo XIX pero no en el siglo XX. Esta tabla muestra todas las palabras que son más comunes en un período que en otro, incluso si no aparecen en el segundo período. P. ej., no hay resultados de televisión o aeropuerto en el siglo XIX (lo que probablemente no sea sorprendente).

Para «comparar manzanas con manzanas», podría ser útil indicar que una palabra debe darse con al menos una frecuencia determinada en cada uno de los dos períodos. P. ej., la tabla 7 muestra los nombres que aparecen al menos diez veces en los siglos XIX y XX, pero en los que hay un aumento significativo a lo largo del tiempo[5].

En resumen, debido a la arquitectura del *Corpus del Español*, donde el corpus «reconoce» la frecuencia de cada palabra y frase en cada período histórico, tales comparaciones son bastante simples. En cambio, con el *CORDE* el corpus aparentemente no reconoce la frecuencia de palabras y frases en cada sección

5 Téngase en cuenta que, debido a que el etiquetado no es siempre perfecto, hay algunas entradas extrañas, como *lic* en el siglo XX o el nombre propio de *Matilde* en el siglo XIX, pero la mayoría de las palabras son relevantes.

SEC 1 (1900s): 22,822,256 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	DÉCADA	3647	16	159.8	0.8	192.7
2	DÓLARES	3081	19	135.0	1.0	137.1
3	FINALES	2644	19	115.9	1.0	117.7
4	CÉLULAS	1693	16	74.2	0.8	89.5
5	POSIBILIDADES	1350	15	59.2	0.8	76.1
6	DÉCADAS	1112	14	48.7	0.7	67.2
7	CINE	1613	21	70.7	1.1	64.9
8	UTILIZACIÓN	758	10	33.2	0.5	64.1
9	LIC	2004	27	87.8	1.4	62.8
10	DEPORTE	712	10	31.2	0.5	60.2
11	EQUIPO	3185	46	139.6	2.4	58.5
12	OBJETIVOS	809	13	35.4	0.7	52.6

SEC 2 (1800s): 19,297,249 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	VENTURA	1237	25	64.1	1.1	58.5
2	NÚM	1025	21	53.1	0.9	57.7
3	RODRIGO	697	19	36.1	0.8	43.4
4	VULGO	662	20	34.3	0.9	39.1
5	PAJE	611	21	31.7	0.9	34.4
6	MATILDE	669	24	34.7	1.1	33.0
7	APOSENTO	1174	44	60.8	1.9	31.6
8	HONRA	1652	63	85.6	2.8	31.0
9	MENESTER	1545	60	80.1	2.6	30.5
10	MARGARITA	554	23	28.7	1.0	28.5
11	ERMITAÑO	553	23	28.7	1.0	28.4
12	CONDESA	1829	78	94.8	3.4	27.7

Gráfico 7: *Corpus del Español*: Comparación de la frecuencia de las palabras (sustantivos en siglo XIX/siglo XX)

(hasta que se buscan, una palabra específica cada vez), por lo que tal listado sería completamente imposible.

7 Búsqueda de formas de palabras (morfológicas)

Idealmente, los usuarios deberían poder ir más allá de las palabras y frases exactas y buscar por prefijos, sufijos y raíces. Esto les permitiría ver la frecuencia de cada forma coincidente en los diferentes períodos históricos, así como también la frecuencia general de todas las formas en cada período histórico.

El *CORDE* tiene serios problemas en búsquedas orientadas morfológicamente porque el motor de búsqueda no fue diseñado para ser utilizado en investigaciones de orientación gramatical. En el mejor de los casos, el corpus produce resultados, aunque no son demasiado útiles. P. ej., supongamos que un usuario busca *des*m?ento* en los siglos del XIII al XV. El corpus indica que hay «797 casos en 81 documentos». Luego, puede recorrer todas las 797 ocurrencias, una por una, y contar manualmente el total de cada forma diferente (*desfazimiento*, *desaffiamiento*, etc.) y ver con qué frecuencia aparece cada una, pero esto llevaría una o dos horas. Se podría seleccionar «Recuperar/agrupaciones» para ver las cadenas de dos, tres y cinco palabras más frecuentes (*destroymiento de*, *destruimiento de*, etc.), y se podría tardar una hora, aproximadamente, en encontrar las palabras más frecuentes que coincidan con este patrón. Estas búsquedas, además, solo funcionan cuando el número total de ocurrencias para una forma determinada aparece mil veces o menos en el corpus. Para una búsqueda como **azo* (*puñetazo*, *portazo*, etc.), el corpus simplemente afirma que «No se pueden ver estadísticas. Demasiados documentos». Una vez más, esto se debe a que su motor de

búsqueda fue diseñado para permitir a los usuarios encontrar y leer documentos completos (como con Google) y (en este caso, al menos) no es adecuado para la investigación gramatical.

Con el *Corpus del Español*, sin embargo, las búsquedas orientadas morfológicamente son fáciles y rápidas. P. ej., un usuario que desee encontrar las formas más frecuentes para *des*m?ento* entre los siglos XIII–XVII, aproximadamente en un segundo, verá lo siguiente:

		CONTEXTO	TODOS	s13	s14	s15	s16	s17
1		DESCUBRIMIENTO	864			13	515	336
2		DESABRIMIENTO	160			2	80	78
3		DESTRUYMIENTO	121	89	17	15		
4		DESTROYMIENTO	97	94	2	1		
5		DESTRUYMJENTO	87	14	45	28		
6		DESVANECIMIENTO	77			2	15	60
7		DESFALLESÇIMJENTO	74		3	71		
8		DESTERRAMIENTO	65	53	10	2		
9		DESAGRADECIMIENTO	57			3	40	14
10		DESCOMEDIMIENTO	46				33	13
11		DESCOMULGAMIENTO	42	41		1		
12		DESTERRAMJENTO	42	4	14	24		

Gráfico 8: *Corpus del Español*: frecuencia de formas verbales (*des*m?ento* en los siglos XIII–XVII)

La interfaz muestra la frecuencia de cada forma en cada siglo (p. ej., 53 casos de *desterramiento* en el XIII), así como en el total de los siglos seleccionados (en este caso, del XIII al XV) en la columna de la derecha. Los usuarios pueden seleccionar las formas y los siglos que sean de interés, y luego hacer clic para ver las palabras en contexto.

Además de recuperar las frecuencias individuales para todas las formas coincidentes (como en el Gráfico 8), también es posible ver el total agregado para todas las formas coincidentes en cada siglo, como hemos indicado en el Gráfico 5. Finalmente, como se describió en la sección 6, también podemos comparar la frecuencia de las formas en diferentes secciones del corpus. P. ej., supongamos que un usuario del *Corpus del Español* quiere ver qué palabras que terminan en **ismo* son más comunes en el siglo XX (izquierda) y el siglo XIX (derecha), respectivamente. En menos de un segundo, vería lo siguiente:

SEC 1 (1900s): 22,822,256 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	PROTAGONISMO	175	0	7.7	0.0	766.8
2	METABOLISMO	148	0	6.5	0.0	648.5
3	URBANISMO	127	0	5.6	0.0	556.5
4	SURREALISMO	118	0	5.2	0.0	517.0
5	ATLETISMO	111	0	4.9	0.0	486.4
6	MARXISMO	108	0	4.7	0.0	473.2
7	FASCISMO	108	0	4.7	0.0	473.2
8	CUBISMO	105	0	4.6	0.0	460.1
9	HINDUISMO	104	0	4.6	0.0	455.7
10	NERVIOSISMO	97	0	4.3	0.0	425.0
11	EXPRESIONISMO	79	0	3.5	0.0	346.2

SEC 2 (1800s): 19,297,249 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	PAUPERISMO	71	0	3.7	0.0	367.9
2	CACIQUISMO	123	14	6.4	0.6	10.4
3	DESPOTISMO	293	38	15.2	1.7	9.1
4	ABISMO	1030	186	53.4	8.1	6.5
5	PAGANISMO	88	16	4.6	0.7	6.5
6	FANATISMO	281	65	14.6	2.8	5.1
7	PATRIOTISMO	330	78	17.1	3.4	5.0
8	FATALISMO	50	19	2.6	0.8	3.1
9	SENTIMENTALISMO	57	22	3.0	1.0	3.1
10	CATOLICISMO	304	138	15.8	6.0	2.6
11	CATECISMO	147	68	7.6	3.0	2.6

Gráfico 9: *Corpus del Español*: Comparación de formas (**ismo* en siglos XIX/XX)

Esto muestra, p. ej., que *despotismo* se encuentra 293 veces en el siglo XIX, pero solo 38 veces en el XX, y que *fascismo* se encuentra 108 veces el siglo XX, pero ninguna en el XIX. La capacidad de comparar formas de palabras en diferentes siglos es una característica de gran alcance del *Corpus del Español*, pero no es posible con el *CORDE*.

8 Significado de la palabra (semántica): colocaciones básicas

Como les gusta señalar a los lingüistas de corpus, «se puede decir mucho sobre una palabra con las otras palabras con las que se junta» (Firth 1957). A veces, las colocaciones (palabras cercanas) simplemente confirman lo que ya sabemos. P. ej., las colocaciones nominales más comunes (palabras cercanas) para *selva* son *árboles*, *vegetación*, *sierra*, *bosque*, etc. Para una palabra menos concreta, a menudo es necesaria más perspicacia. P. ej., los sustantivos más comunes que aparecen con formas de *lúgubre* son *acento*, *voz*, *silencio*, *noche*, *eco*, *gemido*, etc.

La clave del significado, entonces, se suele encontrar en las colocaciones o en las palabras cercanas. Prácticamente cualquier arquitectura e interfaz de corpus permite a los usuarios buscar una palabra y luego ver esa palabra en contexto. El usuario del corpus siempre puede recorrer los ejemplos uno por uno, tomar notas sobre palabras cercanas comunes y luego tratar de usar estas colocaciones para discernir el significado. Sin embargo, esto puede consumir mucho tiempo para palabras comunes. Un enfoque más productivo consistiría en hacer que el corpus encuentre todas las colocaciones por sí mismo y luego presentarlas al usuario en orden de frecuencia.

En términos de cambio histórico, lo deseable sería poder encontrar las colocaciones de una palabra dada en diferentes períodos históricos. Al observar los

cambios en las colocaciones a lo largo del tiempo, podemos obtener información sobre los cambios en el significado y el uso de la palabra.

Consideremos brevemente cómo el *CORDE* y el *Corpus del Español* permiten a los usuarios encontrar y procesar las colocaciones para obtener una idea del significado de las palabras. En el caso del *CORDE*, supongamos que queremos examinar las casi 38 000 colocaciones de todas las formas de *duro* (*dura*, *duros*, etc.). Suponiendo que a un usuario le toma unos 20 segundos encontrar cada ocurrencia en contexto y anotar (lo que supone que son) las palabras cercanas relevantes, dedicaría alrededor de 26 días (a ocho horas diarias) a repasar todos los ejemplos relevantes; y esto, suponiendo que el usuario no decidiera cambiar el ancho de la «ventana de colocaciones» o buscase un tipo diferente de colocación, en cuyo caso tendría que emplear, más o menos, otro mes.

El *CORDE* permite a los usuarios ver «agrupaciones» para una palabra determinada, como las de la forma simple *duro* en el siglo XIII:

De 2 palabras	%	Casos	De 3 palabras	%	Casos
duro la	6.97	30	*duro en el*	2.55	11
duro el	6.27	27	*duro la batalla*	1.86	8
duro en	5.34	23	*duro punto nado*	1.16	5
duro fasta	3.95	17	*duro esta batalla*	1.16	5
duro esta	3.48	15	*duro fasta el*	1.16	5

Gráfico 10: *CORDE*: *duro* + colocaciones

Pero tal listado es de poco valor, porque como el *CORDE* no tiene ninguna forma de discriminar qué palabras son relevantes, nos da colocaciones como *dura la*, *dura en*, etc. (ya que *la*, *en*, etc., concurren con frecuencia con casi cualquier término), que proporcionan poca o ninguna comprensión del significado de la palabra. En cualquier caso, solo lista las nueve colocaciones más frecuentes, lo que no es suficiente para obtener información completa sobre el significado. Estas búsquedas son mucho más fáciles con el *Corpus del Español*. Los usuarios simplemente introducen la «palabra del nodo» (p. ej., *duro*, o *lúgubre*, o *selva*) y pueden, opcionalmente, seleccionar la categoría gramatical de las colocaciones;

en aproximadamente dos o tres segundos tienen todas las colocaciones en orden. P. ej., supongamos que un usuario quiere encontrar colocaciones relacionadas con el concepto '*duro*' en español antiguo. Después de introducir [= duro] (*duro*, *duras*, etc.,) y esperar unos dos segundos, aparece una lista como la siguiente:

		CONTEXTO	TODOS	s13	s14	s15	s16	s17
1		PIEDRA	188	33	7	52	56	40
2		QUEBRANTAR	91	89			2	
3		MIENTRAS	84			1	48	35
4		PIEDRAS	81	10	3	23	28	17
5		DURA	77	13		15	30	19
6		GOLPES	75		12	1	53	9
7		FUERT	71	71				
8		HIERRO	53			3	36	14
9		PEÑA	48			2	21	25
10		MÁRMOL	47				26	21
11		HADO	42				40	2
12		PEÑAS	42		1	1	24	16

Gráfico 11: *Corpus del Español*: colocaciones de *duro* en siglos XIII–XVIII

Esta tabla muestra la frecuencia de cada colocación en cada siglo (aquí solo se muestran los siglos XIII al XVII). P. ej., *piedra* ocurre cerca de [*duro*] 33 veces en el siglo XIII y 7 veces más en el XV. Hay 188 apariciones totales de piedra entre los siglos XIII y XV, por lo que los 40 casos cerca de [*duro*] son aproximadamente el 1,8 % de todos los resultados. Esto se traduce en una puntuación de información mutua de 3.53, que muestra que la relación entre las dos palabras es significativa. Por lo tanto, con el *Corpus del Español*, podemos hacer en 2–3 segundos lo mismo que llevaría hacer con el *CORDE* un mes o más.

9 Cambio semántico: comparación de colocaciones en diferentes períodos históricos

Si tenemos una arquitectura e interfaz de corpus que nos permite encontrar fácilmente colocaciones (lo que es posible, como se ha visto, con el *Corpus del Español*), podemos utilizar esta información de manera ingeniosa para examinar el cambio semántico. La idea básica es que si las palabras «cercanas» a una

determinada palabra cambian con el tiempo, puede deberse a que la misma palabra ha cambiado de significado (o al menos se está utilizando de una manera diferente). P. ej., la siguiente tabla muestra (a la izquierda) los sustantivos que aparecen con [*duro*] en el siglo XX, pero que no son muy comunes en el XIX, y (a la derecha) en el XIX, pero no en el XX:

SEC 1 (1900s): 22,822,256 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	CRÍTICAS	27	0	1.2	0.0	118.3
2	MADERAS	19	0	0.8	0.0	83.3
3	LÍNEA	18	0	0.8	0.0	78.9
4	REPRESIÓN	16	0	0.7	0.0	70.1
5	DISCO	13	0	0.6	0.0	57.0
6	REGALO	10	0	0.4	0.0	43.8
7	LUCHA	10	1	0.4	0.1	8.5
8	COMPETENCIA	15	2	0.7	0.1	6.3
9	BATALLA	12	2	0.5	0.1	5.1
10	GOLPE	51	12	2.2	0.6	3.6
11	MESES	12	3	0.5	0.2	3.4
12	FORMA	12	3	0.5	0.2	3.4
13	GOBIERNO	13	4	0.6	0.2	2.7

SEC 2 (1800s): 19,297,249 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	MILES	30	0	1.6	0.0	155.5
2	RENTA	24	0	1.2	0.0	124.4
3	MILLONES	20	0	1.0	0.0	103.6
4	TRANCE	18	0	0.9	0.0	93.3
5	ART	18	0	0.9	0.0	93.3
6	CARÁCTER	17	0	0.9	0.0	88.1
7	ENTRAÑAS	14	0	0.7	0.0	72.5
8	DUROS	13	0	0.7	0.0	67.4
9	REALES	13	0	0.7	0.0	67.4
10	PESOS	12	0	0.6	0.0	62.2
11	PUÑADO	10	0	0.5	0.0	51.8
12	CANTIDAD	10	0	0.5	0.0	51.8
13	ALMA	10	0	0.5	0.0	51.8

Gráfico 12: *Corpus del Español*: comparación de las colocaciones de *duro*, siglos XIX/XX

P. ej., *disco* se produce cerca de *duro* 13 veces en el siglo XX, pero no hay resultados (lo que no sorprende) en el XIX. *Entrañas*, por otro lado, aparece 14 veces con [*duro*] en el XIX, pero no hay resultados en el siglo XX. Si *disco* está en el corpus en el siglo XIX y *entrañas* en el siglo XX, ¿por qué su frecuencia como colocación con *duro* cambia tanto de un siglo a otro? ¿Es porque el uso de *duro* puede haber cambiado en algo? Mostremos otro ejemplo: la siguiente es una lista parcial de las colocaciones adjetivales de *mujer* (y *mujeres*) en los siglos XX (izquierda) y XIX (derecha):

SEC 1 (1900s): 22,822,256 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	INTERNACIONAL	17	0	0.7	0.0	74.5
2	SEXUAL	15	0	0.7	0.0	65.7
3	CUBANAS	15	0	0.7	0.0	65.7
4	EMBARAZADAS	14	0	0.6	0.0	61.3
5	IMPORTANTE	13	0	0.6	0.0	57.0
6	PROFESIONALES	13	0	0.6	0.0	57.0
7	LABORAL	12	0	0.5	0.0	52.6
8	EX	12	0	0.5	0.0	52.6
9	ÁRABE	10	0	0.4	0.0	43.8
10	DERECHOS	10	0	0.4	0.0	43.8
11	DIFERENTES	15	1	0.7	0.1	12.7
12	MADURA	15	1	0.7	0.1	12.7
13	SEXUALES	12	1	0.5	0.1	10.1

SEC 2 (1800s): 19,297,249 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	HONRADA	58	0	3.0	0.0	300.6
2	HONRADAS	35	0	1.8	0.0	181.4
3	DIGNA	24	0	1.2	0.0	124.4
4	LIVIANA	21	0	1.1	0.0	108.8
5	INFAME	19	0	1.0	0.0	98.5
6	DESVENTURADA	19	0	1.0	0.0	98.5
7	PERDIDAS	18	0	0.9	0.0	93.3
8	VANO	18	0	0.9	0.0	93.3
9	MALDITA	16	0	0.8	0.0	82.9
10	SEMEJANTE	15	0	0.8	0.0	77.7
11	DESDICHADA	14	0	0.7	0.0	72.5
12	LIVIANAS	13	0	0.7	0.0	67.4
13	INFELICES	13	0	0.7	0.0	67.4

Gráfico 13: *Corpus del Español*: comparación de colocaciones de *mujer*, siglos XIX/XX

Obsérvese cómo los adjetivos del siglo XIX (derecha) se refieren a las «virtudes morales» de las mujeres, que están casi por completo ausentes en el siglo XX (izquierda); en este siglo, por otro lado, son mucho más prosaicos y se refieren a clasificaciones sobre la nacionalidad, el empleo, etc. En este caso, los datos del corpus proporcionan información interesante del cambio en la forma de ver a las mujeres en estos dos siglos; podemos obtener esta útil información con una simple búsqueda de 1–2 segundos en el corpus.

Aplicado al español antiguo o en los textos de los siglos XVI al XVIII, se podría adoptar un enfoque similar. Usando la interfaz para el *Corpus del Español,* simplemente es necesario indicar qué palabras o conceptos son de interés, especificar el tipo de colocación (sustantivo, verbo, etc., si corresponde), y luego hacer clic una o dos veces más para mostrar qué dos períodos históricos deben ser comparados. En dos o tres segundos, se recopilan y resumen todos los datos relevantes. Usando el *CORDE*, en cambio, las búsquedas como esta serían muy difíciles o imposibles, ya que la arquitectura del *CORDE* no está diseñada para encontrar colocaciones.

10 Cambios léxicos en un campo semántico: sinónimos

Con la arquitectura de corpus adecuada, los usuarios podrían buscar por campos semánticos, en lugar de simplemente buscar palabras y frases. Así, en el *Corpus del Español*, los índices onomasiológicos se integran en la arquitectura del corpus; esto nos permite, en el nivel más básico, encontrar la frecuencia histórica de todas las palabras relacionadas con un concepto en particular. P. ej., los usuarios pueden introducir = *mujer* y ver la frecuencia de todos los sinónimos de *mujer* a lo largo del tiempo (y, p. ej., en diferentes géneros en el siglo XX).

Esta lista parcial de resultados muestra, p. ej., que *doncella* y *moza* han disminuido desde el siglo XVI (por cada millón de palabras), mientras que *chica* y *muchacha* han aumentado desde el siglo XVIII hasta el XIX/XX. En cuanto a los períodos medievales, lo que obviamente se necesitaría es algún tipo de «índice histórico» que por ahora no está disponible. Pero, en la medida en que lo estuviera, la arquitectura del corpus podría incorporarlo fácilmente.

Además de buscar la frecuencia de palabras sueltas, la información semántica de los repertorios o las listas de palabras personalizadas y definidas por el usuario se pueden integrar directamente en la sintaxis de la consulta. P. ej., en el *Corpus del Español*, es posible que los usuarios creen (a través de la interfaz web) listas personalizadas de palabras de un campo de interés semántico particular, como términos navales, palabras relacionadas con las emociones, una lista de términos relacionados con la estructura familiar o una lista de palabras relacionadas con

	CONTEXTO	TODOS	s13	s14	s15	s16	s17	s18	s19	s20	ACAD	PER	FIC	ORAL
1	MUJER [S]	35395	19	6	687	5707	8233	1962	11836	6945	478	803	3740	1924
2	SEÑORA [S]	32812	161	955	1584	9385	8635	1704	7465	2923	43	314	1819	747
3	PERSONA [S]	24445	430	215	2637	5576	3430	2044	5368	4745	787	854	786	2318
4	JOVEN [S]	11639			4	238	465	607	7193	3132	362	729	1408	633
5	ESPOSA [S]	9400	121	36	219	1712	2809	365	2694	1444	196	480	572	196
6	DAMA [S]	8287	3	2	133	1412	3879	332	2020	506	46	134	266	60
7	MUCHACHA [S]	2266				115	135	132	994	890	13	49	663	165
8	DONCELLA [S]	3025	3	1		1003	762	224	967	65	8	11	42	4
9	CHICA [S]	1725	29	43	95	97	59	96	390	916	19	98	406	393
10	SEÑORITA [S]	1534			1	1	7	134	962	429	9	28	261	131
11	HEMBRA [S]	1397	1	3	261	311	144	97	241	339	207	22	89	21
12	MOZA [S]	1592	3		1	326	584	116	538	24	1	1	20	2

Gráfico 14: *Corpus del Español*: comparación de sinónimos de *mujer*

un concepto teológico particular. Esta lista personalizada de palabras se puede usar como parte de la sintaxis de la consulta. P. ej., si un usuario [andrés.gómez] crea una lista de 100 palabras relacionadas con 'emociones' en español antiguo y otra lista de 70 palabras relacionadas con 'relaciones familiares' (*padre*, *hermanastro*, *nuera*, etc.), podría encontrar cada vez que aparezca una palabra en la Lista 1 cerca de la Lista 2. De esta forma, se pueden llevar a cabo búsquedas semánticas de gran alcance en el corpus.

El *Corpus del Español* puede incorporar este tipo de consultas orientadas semánticamente debido a la arquitectura subyacente del corpus, que se basa en bases de datos relacionales. Con este tipo de bases de datos relacionales es posible agregar cualquier cantidad de conjuntos de datos nuevos (repertorios, listas de palabras definidas por el usuario, etc.) y luego integrarlos sin problemas en la sintaxis de la consulta. La arquitectura para el *CORDE*, en cambio, no es «abierta» y no admite la incorporación de otros conjuntos de datos. Solo se puede buscar palabras o frases individuales, pero nada que se aproxime a un campo semántico completo o algo similar.

11 Desvío sincrónico # 1: género

Hasta este punto, hemos discutido los cambios históricos en el léxico y el significado, pero apenas hemos hecho alusión al género y a cómo este afecta a la frecuencia y al significado de las palabras. Sin embargo, obviamente, hay que considerar la importancia del género. P. ej., analícense los siguientes tres cuadros, que muestran la frecuencia de tres palabras del español moderno que son más comunes en el género de ficción (*borracho*), académico (*proporcionar*) y el marcador de discurso *bueno* en español hablado (como en «lo haremos. Bueno, pero no es tan fácil»):

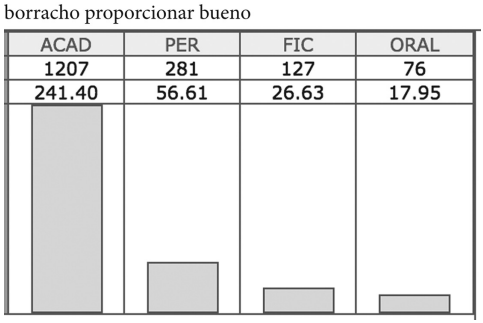


Gráfico 15: Frecuencia de palabras por género en español moderno

Estadísticas

Año	%	Casos	País	%	Casos	Tema	%	Casos
1996	10.08	124	*ESPAÑA*	64.81	818	*6.- Salud.*	22.59	289
2000	8.86	109	*MÉXICO*	14.18	179	*1.- Ciencia y Tecnología.*	20.95	268
2001	8.78	108	*ARGENTINA*	8.00	101	*2.- Ciencias sociales, creencias y pensamiento.*	15.71	201
1995	7.39	91	*CHILE*	2.77	35	*5.- Ocio, vida cotidiana.*	11.64	149
1997	7.07	87	*PERÚ*	2.45	31	*4.- Artes.*	10.47	134
2002	5.93	73	*EE. UU.*	1.66	21	*3.- Política, economía, comercio y finanzas.*	10.39	133

Gráfico 16: Frecuencia por género/dominio en *CREA*

Afortunadamente, el *Corpus del Español* nos permite encontrar rápida y fácilmente la frecuencia de cualquier palabra, frase, morfema o construcción sintáctica en los diferentes géneros del español moderno (siglo XX). Los corpus de la Real Academia Española no proporcionan cifras de frecuencia útiles por género. El *CREA* es el corpus «sincrónico» de la RAE y, de forma similar a lo que hemos visto en el Gráfico 4 para el corpus diacrónico *CORDE*, no proporciona frecuencias «normalizadas» por género. P. ej., muestra que el 22.59 % de las ocurrencias de *borracho* proviene de textos de 'salud' y que 11.64 (aproximadamente la mitad de Salud) proviene de textos que tratan de 'ocio, vida cotidiana'.

Pero si el sub-corpus para 'salud' tiene dos veces el tamaño del corpus 'ocio', entonces cualquier cosa ocurriría aproximadamente el doble en 'salud' y, por tanto, los porcentajes no tendrían sentido. Un corpus tiene que mostrar la frecuencia normalizada (p. ej., por millón de palabras) para que los resultados adquieran algún valor.

Al igual que con la parte histórica *del Corpus del Español* (v. Gráfico 8), también podemos encontrar la frecuencia por género en el *Corpus del Español* para

	CONTEXTO	TODOS	s13	s14	s15	s16	s17	s18	s19	s20	ACAD	PER	FIC	ORAL
1	[HERMOSO] [S]	22329	559	432	1249	6328	6059	1392	4919	1391	117	226	869	179
2	[BELLO] [S]	12404	39	160	261	1632	3375	1366	3844	1727	315	516	631	265
3	[VALIENTE] [S]	6714	62	156	448	2256	1818	353	1397	224	15	46	129	34
4	[LINDO] [S]	4742	61	30	293	857	1143	213	1125	1020	11	133	431	445
5	[BONITO] [S]	2549			10	93	58	63	731	1594	18	120	338	1118
6	[GALÁN] [S]	3887			30	713	2038	169	795	142	6	60	71	5
7	[ELEGANTE] [S]	2423		1	58	215	119	238	1169	623	126	92	322	83
8	[GALLARDO] [S]	2856		14	2	606	1173	238	704	119	8	83	22	6
9	[VALEROSO] [S]	2878			63	1465	774	181	349	46	10	8	27	1
10	[ATRACTIVO] [S]	1501				26	40	227	498	710	151	300	154	105

Gráfico 17: Frecuencia por sinónimo

SEC 1 (FICT): 4,769,873 PALABRAS

	PALABRA/FRASE	OCURR 1	OCURR 2	P/M 1	P/M 2	PROPORCIÓN
1	[CANDADO]	9	0	1.9	0.0	188.7
2	[CUELLO]	8	0	1.7	0.0	167.7
3	[RELOJ]	6	0	1.3	0.0	125.8
4	[CASA]	5	0	1.0	0.0	104.8
5	[BICICLETA]	5	0	1.0	0.0	104.8
6	[RUIDO]	4	0	0.8	0.0	83.9
7	[BOTELLA]	3	0	0.6	0.0	62.9
8	[AMOR]	3	0	0.6	0.0	62.9
9	[CERROJO]	3	0	0.6	0.0	62.9
10	[MANO]	3	0	0.6	0.0	62.9
11	[CRUZ]	2	0	0.4	0.0	41.9

SEC 2 (ACAD): 4,999,945 PALABRAS

	PALABRA/FRASE	OCURR 2	OCURR 1	P/M 2	P/M 1	PROPORCIÓN
1	[PARTE]	12	0	2.4	0.0	240.0
2	[FORMA]	11	0	2.2	0.0	220.0
3	[CORDILLERA]	10	0	2.0	0.0	200.0
4	[ELECTRÓN]	8	0	1.6	0.0	160.0
5	[POLIMERASA]	8	0	1.6	0.0	160.0
6	[CARBONO]	7	0	1.4	0.0	140.0
7	[ÁCIDO]	7	0	1.4	0.0	140.0
8	[SERIE]	7	0	1.4	0.0	140.0
9	[TRANSPORTE]	7	0	1.4	0.0	140.0
10	[AMINOÁCIDO]	6	0	1.2	0.0	120.0
11	[SISTEMA]	6	0	1.2	0.0	120.0

Gráfico 18: Colocaciones por género

los sinónimos de una palabra determinada. P. ej., el gráfico 17 muestra la frecuencia de sinónimos de *guapo* en los cuatro géneros principales (académico, periodístico, ficción y oral). Podemos observar que algunas palabras (como *lindo* y *bonito*) son mucho más frecuentes en los géneros informales (especialmente hablados) que en los géneros más formales, como el académico.

Finalmente, el uso y el significado de una palabra también se ve afectado por el género. P. ej., el gráfico 18 muestra las colocaciones de *cadena* en ficción (a la izquierda) y académica (a la derecha). Obsérvese que, en la ficción, *cadena* generalmente se refiere a una cadena física (es decir, algo que uno puede sostener en la mano), mientras que en el género académico su uso es más metafórico y, a menudo, se refiere a sustantivos abstractos (‘serie’, ‘sistema’, etc.).

Como podemos ver, el género constituye un factor crucial en términos de frecuencia y significado. *El Corpus del Español* nos permite mostrar los resultados por género de manera fácil y rápida, búsqueda que no es posible con los corpus *CREA* y *CORDE*.

Tabla 2: Frecuencia de palabras de *-idad* en el *Corpus del Español* original (2002)

> 100 resultados	**NONE**
50–100	sonoridad, inmovilidad, expresividad, obscuridad
20–49	morbilidad, unicidad, posmodernidad, adaptabilidad, cotidianeidad, discrecionalidad, selectividad, salubridad, elegibilidad, insensibilidad, disconformidad, mensualidad, afectividad, deformidad
10–19	perpetuidad, receptividad, morosidad, anormalidad, hipersensibilidad, disparidad, extremidad, interioridad, permeabilidad, corresponsabilidad, perversidad, susceptibilidad, viscosidad, emotividad, natividad, plasticidad, asiduidad, promiscuidad, salinidad, virilidad, homogeneidad, inutilidad, frivolidad, perplejidad, voracidad
< 10	empleabilidad, trazabilidad, centralidad, consanguinidad, invisibilidad, inhabilidad, estadidad, habitabilidad, alteridad, sociabilidad, probidad, banalidad, materialidad

12 El nuevo *Corpus del Español* (Web/Dialectos)

Si bien es muy útil el componente original «Histórico/Género» de 100 millones de palabras, el *Corpus del Español* presenta algunas limitaciones. La primera es el tamaño: solo hay 20 millones de palabras del siglo XX y esta cifra resulta cuantitativamente escasa para una investigación en profundidad del léxico. La segunda es que el *Corpus del Español* original termina en 1999, no contiene textos del siglo XXI. Y la tercera es que el corpus original no permitía a los usuarios comparar el léxico en diferentes países.

Para abordar estas tres limitaciones creamos una gran extensión del *Corpus del Español* en 2014–2015. El nuevo corpus alcanza un tamaño de dos mil millones de palabras, lo que significa que es 100 veces el tamaño de la parte del español del siglo XX en el *Corpus del Español* original (2002). Todos los textos basados en la web se recopilaron en 2014–2015, por lo que el corpus representa mejor el español reciente. Finalmente, permite a los usuarios comparar frecuencias léxicas en 21 países diferentes de habla hispana.

En términos de tamaño, un corpus que es 100 veces más grande proporciona datos mucho más ricos y mayor información sobre la variación léxica. P. ej., considérese la Tabla 2, que muestra aquellas palabras que terminan en *-idad*, que tienen una frecuencia de entre 2000 y 3000 ocurrencias en el nuevo corpus de dos mil millones de palabras. La tabla muestra cuántas veces aparecen en el corpus de más de 100 millones de palabras (donde 20 millones de palabras son del siglo XX).

Solo unas pocas palabras tienen una frecuencia de al menos 50 ocurrencias en el corpus anterior, la mayoría ocurre menos de 20 veces. Con un número tan limitado de apariciones, es muy poco lo que se puede hacer para investigar el significado y uso de las palabras o investigar su frecuencia en todos los géneros. Además, en algún punto hay tan pocos resultados que la misma palabra *frecuencia* se vuelve problemática. P. ej., una voz que aparece solo 15 veces puede aparecer en solo 3–4 textos diferentes, y la palabra podría no encontrarse en absoluto en un corpus que tuviera una composición textual ligeramente diferente. En algún punto, los términos con baja frecuencia son simplemente «ruido».

Además de la frecuencia de voces aisladas, el nuevo (2016) corpus de dos mil millones de palabras también proporciona datos mucho más ricos para palabras en contexto con otras palabras. P. ej., en el *Corpus del Español* original, solo hay 67 adjetivos diferentes que aparecen cinco veces en el contexto gente + ADJ (*gente joven*, *pobre*, *rica*… *ponderosa*, *popular*), mientras que hay aproximadamente 1445 adjetivos diferentes en el corpus nuevo de dos mil millones de palabras (p. ej., *visionaria*, *desmotivada*, *sumisa*, *enfadada*, *gritona*, *caritativa*). Del mismo modo, solo hay 24 nombres diferentes que aparecen cinco veces o más antes de una forma de *rico* en el corpus anterior (*países*, *comerciantes*, *alimentos*), mientras que hay alrededor de 650 nombres diferentes en el corpus de dos mil millones de palabras (*vocabulario*, *parientes*, *subsuelo*, *guion*, *piedras*). Como ejemplo final, solo hay unos diez sustantivos que aparecen al menos cinco veces en cuatro palabras después de una forma de *destrozar* en el corpus anterior (p. ej., *corazón*, *alma*, *guerra*, *árboles*, *puerta*), pero hay más de 570 nombres en el nuevo corpus (p. ej., *canción*, *confianza*, *películas*, *cadenas*, *pedacitos*, *hacha*). En la mayoría de los casos, hay 50–60 veces más colocaciones en el nuevo corpus de dos mil millones de palabras, lo que, por supuesto, proporciona a los investigadores datos mucho más detallados y útiles sobre el significado y el uso de estas palabras.

13 Búsqueda y estudio de neologismos

Desde el punto de vista del cambio léxico, el corpus nuevo también proporciona datos mucho más ricos para los neologismos. Esto es consecuencia del incremento del corpus (100 veces más grande que la parte del siglo XX del *Corpus del Español* original), pero también se debe a que el nuevo corpus contiene textos muy recientes. Los dos millones de textos fueron extraídos de la Web en 2014–2015, una década y media después de haber sido introducidos los últimos textos en el *Corpus del Español* original.

Hay muchos miles de palabras que aparecen al menos 500 veces en el corpus más reciente, pero menos de diez veces (en muchos casos nada) en el corpus

más antiguo y pequeño. Por supuesto, muchas de estas son palabras relacionadas con la tecnología (p. ej., *blog*, *web*, *internet*, *celular*, *navegador*, *correo electrónico*, *clic*, *tweet*), pero otras son palabras relacionadas con otros campos (p. ej., *documental*, *implementación*, *fiscalía*, *inversionista*, *globalización*, *operativo*, *migrante*, *sostenibilidad*, *biodiversidad*).

Como una prueba más de la extensión de estos neologismos, los sustantivos que terminan en *–idad* y que se anotan a continuación aparecen al menos 700 veces en el nuevo corpus, pero menos de cinco (generalmente 0 veces) en el corpus antiguo: *interoperabilidad, catolicidad, viralidad, masividad, proactividad, escalabilidad, colonialidad, transversalidad, transexualidad, digestibilidad, sincronicidad, ciberseguridad, emocionalidad, accidentalidad, inexequibilidad, tipicidad, ruralidad*. Las palabras que terminan en *-ismo* incluyen: *kirchnerismo, chavismo, fujimorismo, extractivismo, multiculturalismo, uribismo, massismo, madridismo, emprendedurismo, agnosticismo, biomagnetismo, ateísmo, vaginismo, veganismo, ventajismo, macrismo, bruxismo, cateterismo*.

Las palabras que terminan en *-ción* y aparecen al menos 700 veces en el corpus nuevo y más amplio, pero que prácticamente desaparecen en el corpus más antiguo y reducido son: *autenticación, fidelización, virtualización, mercantilización, victimización, geolocalización, encriptación, dispensación, precarización, suplementación, judicialización, visibilización, abducción, desafección, compartición, cosificación, procrastinación, remediación, iteración, resignificación, disrupción, deslocalización, feminización, extranjerización, redirección, evicción, invisibilización, bancarización, demonización, gamificación, previsualización, propiciación*.

Esto no quiere decir que todas estas palabras constituyan neologismos, solo que rara vez aparecen en el corpus más antiguo y pequeño y que, por el contrario, resultan frecuentes en el corpus más moderno y de mayor dimensión. Los lexicógrafos experimentados, por supuesto, querrían investigar tales palabras con más detalle. Pero el punto principal es que el nuevo corpus de dos mil millones de palabras proporciona, además, los datos para la investigación de voces donde ha habido un relativo aumento de frecuencia en los últimos 15–20 años.

14 Desvío sincrónico # 1: dialectos

Además de identificar neologismos, una de las ventajas del nuevo corpus es que permite a los investigadores obtener la frecuencia de cada palabra o frase en 21 países diferentes de habla hispana[6]. Para algunos países, como España, el nuevo

6 Para datos similares de un corpus de inglés análogo, vid. Davies/Fuchs 2015.

corpus contiene hasta 459 millones de palabras (con 261 para México, 183 para Argentina y 180 para Colombia). Incluso entre los países con menor representación cuantitativa en el corpus, cada uno tiene al menos 30 millones de palabras (p. ej., 39 para El Salvador, 39 para Honduras, 37 para la República Dominicana, 36 para Puerto Rico, 35 para Nicaragua, 33 para Paraguay y 32 para Costa Rica).

Las siguientes son solo algunas de las palabras que se registran con una frecuencia más alta en un país que en los otros[7]:

- Caribe
  Puerto Rico *ay bendito, chavos, chiringa, mahones, habichuela* (+DR), *zafacón* (+DR); Cuba *guajiro, jimaguas, babalao, bitongo, pedir botella*; Rep Dom *mangú, fucú, tutumpote, mangulina, mofongo* (+PR).
- México y América Central
  México *ándale, híjole, órale, güero, (muy) padre, chamaco* (CAm/Car), *pinche*, *popote, charola* Guatemala *huipil, canche, muchá, patojo, chafa* (+HN), *chirmol, canche*; El Salvador *cipote, chero, pupusa, cuilio, bayunco, piscucha*; Honduras *catracho, papada*; Nicaragua *chavalo, maje* (+CAm), *pinol, pinolillo, chigüín, vigorón, gallo pinto* (+CR), *idiay* (+CR) Pánama *fulo, chombo, guandul*; Costa Rica *chinear, guila, chunche*.
- América del Sur
  Colombia *cachaco, cachifo, verraquera, estar mamado, guandoca, biche*; Venezuela *bojote, coroto, catire, gafo, macundales, arepa, cachapa, cambur, caraotas, jojoto*; Ecuador *chumar, chulla, montuvio, omoto*; Perú *anticucho, jebe, chupe, pisco, jora, chompa* (+CL/EC), *choclo* (+CL/EC); Bolivia *opa, colla, chuño, lagua*; Chile *pololo*, pololear, achuntar, bencina, bacán, fome, huaso*; Paraguay *ñembo, ñanduti, karai, yopará, mitai*; Uruguay *tropero, hacer * sota, con fritas*; Argentina *pibe, fiaca, morfar, falopa, sobre el pucho, falluto, cafishi.*
- España *ordenador, aparcar, enfadar, gafas, zumo, chulo, guay, coger, bolígrafo, patata, melocotón, echar de menos, vale*.

Además de poder analizar la frecuencia de una palabra o frase específica en los 21 países, también es posible hacer que el corpus produzca una lista de todas las palabras que son más comunes en un país (o conjunto de países) que en otro. P. ej., la siguiente tabla muestra palabras de *-ismo* que son más comunes en Venezuela (izquierda) que en Colombia, México, Argentina o España (derecha):

7 Téngase en cuenta que los países y la región entre paréntesis indican que la palabra también tiene una alta frecuencia relativa en estas otras zonas. La lista de palabras se corresponde con los datos léxicos aportados por Lipski 1996.

SEC 1 (Venezuela): 98,170,248 WORDS

	WORD/PHRASE	TOKENS 1	TOKENS 2	PM 1	PM 2	RATIO
1	JALABOLISMO	44	1	0.4	0.0	452.0
2	ESCUALIDISMO	21	1	0.2	0.0	215.7
3	MADURISMO	141	10	1.4	0.0	144.8
4	GOMECISMO	41	3	0.4	0.0	140.4
5	VENTAJISMO	812	63	8.3	0.1	132.4
6	PUNTOFIJISMO	165	14	1.7	0.0	121.1
7	RIQUISMO	22	3	0.2	0.0	75.3
8	HIPISMO	295	41	3.0	0.0	73.9
9	CASTRO-COMUNISMO	55	8	0.6	0.0	70.6
10	ECO-SOCIALISMO	27	4	0.3	0.0	69.3
11	VENEZOLANISMO	27	4	0.3	0.0	69.3
12	CASTROCOMUNISMO	58	9	0.6	0.0	66.2

SEC 2 (Argentina, Colombia, España...): 1,008,426,240 WORDS

	WORD/PHRASE	TOKENS 2	TOKENS 1	PM 2	PM 1	RATIO
1	MASSISMO	1249	1	1.2	0.0	121.6
2	PRIISMO	362	0	0.4	0.0	35.9
3	EMPRENDEDORISMO	251	0	0.2	0.0	24.9
4	GARANTISMO	254	1	0.3	0.0	24.7
5	PARKINSONISMO	219	0	0.2	0.0	21.7
6	NEOPLATONISMO	218	1	0.2	0.0	21.2
7	CHARRISMO	204	1	0.2	0.0	19.9
8	SOCORRISMO	196	0	0.2	0.0	19.4
9	LIBERTARISMO	184	0	0.2	0.0	18.2
10	CRISTINISMO	557	3	0.6	0.0	18.1
11	COPERNICANISMO	173	0	0.2	0.0	17.2
12	MOTOTAXISMO	171	0	0.2	0.0	17.0

Gráfico 19: Comparación por dialecto en el corpus Web/Dialecto

15 Comparación con el *CORPES*

Al analizar los datos históricos del *Corpus del Español* original, los comparamos con los datos del corpus *CORDE* y, cuando discutimos los datos sincrónicos del *Corpus del Español* original, lo comparamos con el corpus *CREA*. Existe un tercer corpus de la Real Academia Española que es relevante, en términos de una comparación con la parte más moderna de dos mil millones de palabras «Web/Dialectos» del *Corpus del Español*, que hemos discutido en las Secciones 12–14; este tercer corpus es el *CORPES* (*Corpus del Español del Siglo XXI*).

En cuanto a posibilidades de estudios sobre léxico, existen algunas diferencias importantes entre el *CORPES* y la nueva extensión de dos mil millones de palabras del *Corpus del Español* (que llamaremos *CDE-2* en esta sección). La primera diferencia importante es el tamaño: el *CORPES* tiene aproximadamente 175 millones de palabras, que es menos del 10 % del tamaño del *CDE-2*. Esto es, en caso de que una palabra tenga 200 ocurrencias en *CDE-2* (probablemente un número suficiente para muchos tipos de investigación), en el *CORPES* probablemente tendría menos de 20 ocurrencias, lo que resulta mucho más problemático. En segundo lugar, los materiales que integran el *CDE-2* son más recientes, lo que resulta importante para detectar neologismos. Solo alrededor del 17 % del *CORPES* es posterior a 2010, mientras que el 100 % de los datos del *CDE-2* corresponde a ese período (los textos se recopilaron en 2014–2015).

También hay diferencias importantes en cuanto a la arquitectura e interfaz del corpus. Aunque el *CORPES* puede generar gráficos de frecuencia útiles para palabras individuales o para una frase exacta, no es posible encontrar la frecuencia de cadenas coincidentes en una búsqueda como «*menos * que*» (p. ej., *menos valor/intenso/arriesgado/sano que*) o «adjetivo + *ojos*» (p. ej., *ojos bellos/nublados/*

cristalinos/cariñosos). En relación con esto, con el *CORPES* no es posible obtener la frecuencia de colocaciones significativas (es decir, es típico del *CORPES* mostrar artículos o preposiciones como las colocaciones más importantes de una palabra, como es también el caso del *CREA* y el *CORDE*, v. el Gráfico 10). El *CORPES* tampoco permite la comparación de colocaciones entre palabras diferentes (p. ej., *potente* y *poderoso*, o *iluminar* y *alumbrar*), aunque puedan ser búsquedas útiles para poner de manifiesto las diferencias de significado. Y, por último, el *CORPES* no permite comparaciones entre todas las palabras en diferentes países para encontrar las que son más comunes en un país que en otro (v. Gráfico 19). Por tanto, aunque el *CORPES* tiene una arquitectura e interfaz más avanzada que el *CORDE* o el *CREA*, todavía resulta bastante limitada en los tipos de búsquedas que permite.

16 Datos muy recientes y datos continuamente actualizados

En 2016 lanzamos un corpus de textos en inglés llamado *NOW* («Noticias en la web»). Este corpus crece automáticamente en tamaño en alrededor de 5–6 millones de palabras por día y está basado en aproximadamente 10 000 nuevas URL diarias extraídas de Google News. Con este corpus, pueden rastrearse los cambios en la frecuencia y el uso de palabras en el transcurso de meses, semanas e incluso días, lo que es obviamente muy útil para observar cambios extremadamente recientes en el lenguaje. P. ej., los investigadores pueden rastrear la frecuencia en el tiempo de las palabras «nuevas» en el idioma (que han surgido a partir de 2007), como *Brexit, manspreading, makerspace, gig economy, dadbod, momager, swatting, walkscore, trigger warning, mommy porn, normcore, listicle, sufferfest, catfishing, sapiosexual, nomophobia, omnishambles, humblebrag, FOMO, precariat, filter bubble, range anxiety, collaborative consumption, churnalism, birther, truther, staycation, glamping, locavore, voluntourism, freegan.*

Hemos recopilado los textos para un corpus similar en español (que llamaremos aquí *NOW-Español*), que se lanzará en mayo de 2018. Cuando se publique contendrá aproximadamente 4800 millones de palabras desde enero de 2012 hasta mayo de 2018, y luego crecerá en tamaño alrededor de 140 millones de palabras cada mes (alrededor de 1600 millones de palabras cada año). De la misma manera que ya es posible para el inglés, los usuarios de *NOW-Español* podrán seguir la frecuencia de cualquier palabra, frase o construcción sintáctica a lo largo del tiempo (incluso a nivel de meses y semanas), y también tendrán las listas de neologismos en los textos que generará el corpus automáticamente.

Aunque *NOW-Español* aún no está disponible, podemos utilizar algunos datos preliminares para mostrar cómo *NOW-Español* se compara con la extensión de

dos mil millones de palabras en el *Corpus del Español* que, como hemos indicado, contiene textos recopilados de 2014–2015. P. ej., las siguientes palabras que terminan en *-idad* aparecen al menos 10 veces en *NOW-Español*, pero no figuran en los datos del *Corpus del Español* de 2014–2015: *poliautoinmunidad, bancabilidad, promovilidad, superelasticidad, sucrosidad, sucreñidad, inexportabilidad, teleseguridad*. Las palabras terminadas en *-ismo* incluyen *narcouribismo, uribestialismo, cadivismo, cuñadismo, carnismo, larretismo, raspacupismo, chaveztialismo, urbanudismo, figurinismo*. Y las palabras terminadas en *-ción* incluyen *narcorevolución, uberización, multihabitación, posdesmovilización, incoración, cibervictimización, kilometración, deaferentación, electrorreducción, multilateración, desdiabolización, termonebulización, retermalización, hipsterización*. Una vez más, no hay garantía de que todas estas palabras sean neologismos (es decir, que sean novedades de los últimos 3–4 años), solo de que aparecen en el español actual y no figuran en el *Corpus del Español* de dos mil millones de palabras. Los lexicógrafos tendrán acceso a esta rica información para observer los cambios más recientes en el lenguaje.

17 Conclusión

Los grandes corpus accesibles mediante Internet han ayudado a revolucionar el campo de la lexicografía histórica. Con solo unos pocos clics del ratón, los investigadores pueden buscar corpus que contengan cientos de millones (y ahora miles de millones) de palabras de miles (y ahora millones) de textos.

Cada corpus tiene sus propias fortalezas y debilidades. Los corpus *CORDE* y *CREA* de la Real Academia Española son bastante sólidos en términos de «corpus textual», pero tienen arquitecturas e interfaces muy limitadas y obsoletas que limitan el acceso de los investigadores a estos datos. Por otro lado, la parte «histórica/de género», original del *Corpus del Español* (que fue lanzado en 2002) constituye un corpus de dimensiones más reducidas que el *CORDE* o el *CREA*, pero permite varios tipos de investigación que no pueden llevarse a cabo con estos dos corpus, lo que incluye recuperar la frecuencia de palabras, frases y subcadenas (p. ej., prefijos o sufijos) por siglo o género, seguir los patrones de cambio de colocaciones para observar las modificaciones en el significado y uso, y utilizar el índice integrado y las listas de palabras personalizadas para ver cómo las voces compiten por el «espacio semántico» a lo largo del tiempo (o en diferentes géneros).

Uno de los usos más interesantes de los grandes corpus en línea es poner de relieve los cambios muy recientes en el lenguaje, como puede hacerse con la nueva extensión del *Corpus del Español*, que contiene dos mil millones de palabras de

21 países diferentes de habla hispana de 2014–2015. A partir de mayo de 2018, será posible utilizar *NOW-Español* para analizar los cambios que se producen virtualmente en «tiempo real», como ya es posible para *NOW-English*. En algún momento, incluso, sería posible crear corpus que monitoricen continuamente trillones de palabras de las redes sociales para rastrear con increíble precisión cómo las palabras y frases se propagan a lo largo del tiempo a través de distintas comunidades de habla y dialectos[8].

En definitiva, los grandes corpus en línea nos permiten realizar muchos tipos de investigaciones que apenas eran imaginables hace 15 o 20 años y, con las mejoras y avances en la tecnología, podemos darnos cuenta de que solo estamos en el comienzo de lo que se puede llegar a hacer.

Referencias bibliográficas

Davies, Mark (2002): «Un corpus anotado de 100.000.000 de palabras del español histórico y moderno», en *Actas del XVIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Valladolid: SEPLN, 21–27.

Davies, Mark (2005a): «Advanced research on syntactic and semantic change with the Corpus del Español», en Claus Pusch *et al.* (eds.), *Romance Corpus Linguistics II: Corpora and Diachronic Linguistics*. Tubinga: Guntar Naar, 203–214.

Davies, Mark (2005b): «The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation», *International Journal of Corpus Linguistics* 10, 301–328.

Davies, Mark (2008): «Spanish and Portuguese corpus linguistics», *Studies in Hispanic and Lusophone Linguistics* 1, 149–186.

Davies, Mark (2010): «Creating useful historical corpora: A comparison of *CORDE*, the Corpus del Español, and the Corpus do Português», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorromances: nuevas perspectivas desde la lingüística de corpus*. Fráncfort/Madrid: Vervuert/Iberoamericana, 137–166.

Davies, Mark (2012a): «Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English», *Corpora* 7, 121–157.

8 Véase la Sección 7 de Davies 2015.

Davies, Mark (2012b): «Examining Recent Changes in English: Some Methodological Issues», en Terttu Nevalainen y Elizabeth Closs Traugott (eds.), *The Oxford Handbook of the History of English*. Oxford: Oxford Univ. Press, 263–287.

Davies, Mark (2015): «Corpora: An introduction», en Douglas Biber y Randi Reppen (eds.), *Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 11–31.

Davies, Mark (en prensa a): «Using large online corpora to examine lexical, semantic, and cultural variation in different dialects and time periods», en Eric Friginal *et al.* (eds.), *Corpus-Based Sociolinguistics*. Londres: Routledge.

Davies, Mark (en prensa b): «Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design», en Carla Suhr, Terttu Nevalainen e Irma Taavitsainen (eds.), *From data to evidence in English language research* (Digital Linguistics). Leiden: Brill.

Davies, Mark/Robert Fuchs (2015): «Expanding Horizons in the Study of World Englishes with the 1.9 Billion Word Global Web-Based English Corpus (*GloWbE*)», *English World-Wide* 36, 1–28.

Firth, J. R. (1957): *Papers in Linguistics 1934–1951*. Londres: Oxford University Presiglos.

Lipski, John (1996): *El español de América*. Madrid: Cátedra.

Virginia Bertolotti y Concepción Company Company

El corpus para América: *CORDIAM*

Resumen: El *Corpus Diacrónico y Diatópico del Español de América* (*CORDIAM*) fue creado para poder historiar el español de América y para hacer lingüística histórica general con datos de muchas variedades del español. Estos datos, exclusivamente americanos, se procesan a través de una interfaz amigable y eficiente, producto del trabajo conjunto de informáticos e investigadores en lingüística histórica e historia de la lengua. Este trabajo muestra las características lingüísticas y textuales del *CORDIAM*, sus características informáticas y da cuenta, además, del proceso de toma de decisiones para la creación de cada uno de los tres subcorpus que conforman el *CORDIAM*: *CORDIAM-Documentos*, *CORDIAM-Prensa* y *CORDIAM-Literatura*. Muestra también cómo los textos incluidos representan facetas diversas de la cultura en América.

Palabras clave: Lingüística de corpus, Corpus escrito, Lingüística histórica, Español en América

Abstract: The *Corpus Diacrónico y Diatópico del Español de América* (*CORDIAM*) is presented in this paper. It was created to make the history of Spanish in America and historical linguistics with data from many dialects of Spanish. These exclusively American data are processing by a friendly and efficient interface, which is the result of the joint work of computer engineers and investigators both in historical linguistics and in language history. This work shows the linguistic and textual characteristics of *CORDIAM*, its computing characteristics and, in addition, explains the decisions underlying process for the creation of each of the three subcorpus of *CORDIAM*, which are *CORDIAM-Documents*, *CORDIAM-Press*, and *CORDIAM-Literature*. It also shows how the included texts exhibit different aspects of the American culture.

Keywords: Corpus linguistics, Written corpus, Historical linguistics, Spanish in America

1 Presentación y objetivos

Los *corpus lingüísticos*, escritos y orales, diacrónicos y sincrónicos, las *ediciones críticas de textos*, antiguos y modernos, así como los *atlas lingüísticos*, son, como es sabido, los tres tipos de productos fundamentales que constituyen infraestructura para la investigación lingüística, y son, por ello, soportes esenciales para abrir nuevos horizontes de investigación, hallar nuevas evidencias, descriptivas

y teóricas, sobre las lenguas y dejar caminos pavimentados a las futuras generaciones de estudiosos.

El *Corpus Diacrónico y Diatópico del Español de América* (*CORDIAM*) fue creado con tres objetivos principales: *a*) historiar el devenir del español en América, *b*) poder realizar una dialectología histórica del español de América, y *c*) completar y enriquecer la historia general de la lengua española, sin parcelaciones geográficas o dialectales, cuando estas no sean requeridas, puesto que hablar y escribir español es un hecho integral, común a varios cientos de millones de hispanohablantes. Para lograr estos tres objetivos, el *CORDIAM* tiene la necesaria amplitud y diversidad geográfica, la necesaria profundidad histórica y la imprescindible diversidad tipológica textual.

En cuanto a la diversidad geográfica, el *CORDIAM* incluye textos de los actuales 19 países que integran Hispanoamérica (Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Ecuador, El Salvador, Guatemala, Honduras, México, Nicaragua, Panamá, Paraguay, Puerto Rico, Perú, República Dominicana, Uruguay y Venezuela), además de contener textos de otros 5 países americanos, que formaron parte del dominio colonial de España desde inicios del siglo XVI a inicios del XIX (Belice, Estados Unidos, Guyana, Jamaica y Trinidad y Tobago). El *CORDIAM* abarca, por lo tanto, 24 actuales países del continente americano. Por lo que respecta a la profundidad temporal, el *CORDIAM* recubre cuatro siglos de diacronía; el primer documento es de 1494 y el último de 1905; *grosso modo*, abarca desde la llegada europea a América hasta la consolidación de los actuales estados nación de Hispanoamérica. Por último, en cuanto a la diversidad textual, se trata, de un corpus variado en cuanto a clases textuales, ya que está conformado por tres subcorpus que contienen una amplia diversidad textual en su interior: *CORDIAM-Documentos*, *CORDIAM-Literatura* y *CORDIAM-Prensa*.

En suma, como el título de este trabajo indica, la especificidad de este corpus es su contenido *estrictamente americano*, con documentos, textos y obras escritos exclusivamente en América y redactados casi absolutamente por nacidos en América, especificidad americana enriquecida, además, por su extensa diacronía más por su amplia variedad textual. Es posible decir que el *CORDIAM*, en el desarrollo y avance cualitativos y cuantitativos actuales, constituye ya el corpus de consulta y referencia obligada para acercarse a fenómenos lingüísticos caracterizadores del español americano, además de, como dijimos, analizar y comprender mejor hechos de lengua del español general.

Fue concebido para brindar datos fiables y robustos a los investigadores de la lengua española, pero también, sin duda, como todo corpus de lengua, el *CORDIAM* permite acercarse y analizar las culturas que se manifiestan en lengua

española o los múltiples ángulos culturales codificados en ella, ya que la lengua, hecho sabido, es el soporte de la actividad humana, porque atraviesa y estructura la vida cotidiana toda de cada ser humano en todas y cada una de sus facetas, o, en otras palabras, como dice Maturana (1996) «somos en el lenguaje y a través de él». En efecto, el *CORDIAM* provee cuantiosa evidencia cultural, de interés para las ciencias sociales, para las humanidades e incluso para la creación ficcional, ya que incluye múltiples historias vitales con visos literarios o cinematográficos.

A modo de ejemplo, mostramos en (1)–(6) algunos breves fragmentos que representan facetas diversas de la cultura en América, tales como la intimidad (1), la economía (2), la vida cotidiana (3), la visión de mundo (4), las noticias en la prensa (5) o la creación literaria (6). Con ellos el lector puede hacerse una primera idea de los contenidos, los textos y la lengua del *CORDIAM*.

(1) Regalo y bien de mis/ojos. ¡Lo q*ue* me cuestas de/lagrimas y sospriro!/Miras si q*ue*res salir /[5] esta noche, q*ue* en ti está./Y si q*u*ires, abisa de/palabra a Juanilla, q*ue*/ba con tu padre, sólo con/desile q*ue* esperas (México, año 1689, documentos entre particulares, *CORDIAM*).

(2) que se llaman chacaras de coca y es una yerva que lo comen/los naturales desta tierra y no la tragan mas que/mascalla y es de grande estima entre ellos y esta en/poder de nosotros que ella no es gente que las sustentan/sino muy poco ella es la mejor moneda que ay/en esta tiera porque por ella se alla quanto estas/yndias tienen oro y plata y ropa y ganados y quanto/tienen dan por ella (Perú, año 1576, Documentos entre particulares, *CORDIAM*).

(3) dijo q*ue* los a uisto estar retosando en su casa,/tirandole los cabellos el a ella y dandole de puñetes, y q*ue*/los a reñido, y que deuajo destas acciones y sospechas /[25] se leuanto la delatante una noche y por no auer candela/ensendio unas ojas de mais (Estados Unidos, año 1728, documentos jurídicos, *CORDIAM*).

(4) auía Dios uençido a las guacas y los/españoles a los yndios; enpero que agora daua/la buelta el mundo y que Dios y los españoles/quedarían vençidos desta uez, y todos los españoles /[30] muertos, y las çiudades dellos anegados; y que/la mar auía de creçer y los auía de aogar por que/dellos no huuiese memoria (Perú, año 1600, documentos cronísticos, *CORDIAM*).

(5) El 22 del próximo pasado entraron á Salta algunas tropas de las que tuvo á sus órdenes el finado Güemes, é hicieron en aquella ciudad un horroroso saqueo (Argentina, año 1821, prensa, textos informativos, *CORDIAM*).

(6) A la voz del sacristán/en la iglesia se colaron/dos princesas de Guinea/con vultos azabachados./Y, mirando tanta fiesta,/por ayudarla cantando,/soltando los sestos, dieron/albricias a los muchachos (México, año 1679, literatura, Sor Juana Inés de la Cruz, *Tercero nocturno. Segundo villancico*, *CORDIAM*).

El *CORDIAM* es un corpus de acceso libre: www.cordiam.org. Los textos que lo conforman tienen una interfaz con el usuario a través de un potente motor de

búsqueda. El *CORDIAM* es un corpus abierto en la medida en que la incorporación de nuevos datos al motor de búsqueda es permanente.

El *CORDIAM* está radicado en la Academia Mexicana de la Lengua, institución que acogió el proyecto desde sus inicios. Ha contado, además, en diversas etapas de su desarrollo con financiamiento diverso, tales como el del Consejo Nacional para la Ciencia y la Tecnología (CONACYT) de México y el de la Universidad de la República del Uruguay. Tampoco sería posible contar con este corpus sin el trabajo y fructífero diálogo interdisciplinario con investigadores del Instituto Politécnico Nacional (IPN) de México, que han realizado el diseño y la interfaz del *CORDIAM*.

El *CORDIAM* se inscribe claramente en el ámbito disciplinario que se ha dado en llamar *Humanidades Digitales*, ya que se basa en el diálogo entre la Lingüística Histórica, la Filología y la Informática. Es, además, un proyecto colaborativo en que participan varias decenas de investigadores de diversas instituciones y países; ha dado lugar, asimismo, a interacción 2.0 a través de una página cuyo objetivo es poner en relación contenidos de interés general del corpus *CORDIAM* con usuarios de la web y mostrar la vigencia de los hechos históricos para entender mejor el mundo cotidiano actual (cf. *Red del Español de América*, *REDEA*, http://redea.cordiam.org/).

La creación y desarrollo del *CORDIAM* se ubica en un momento de gran desarrollo de la lingüística de corpus, en español y en otras lenguas, en que tenemos una experiencia acumulada y una mirada crítica sobre cómo realizar un nuevo corpus histórico. Lo anterior nos permitió un diseño que, creemos, supera algunas de las carencias o limitaciones de otros corpus históricos del español, como veremos en los apartados 2 y 3.

El *CORDIAM* se integra, sin duda, a una tradición de conformación de corpus digitales históricos en español, inaugurada por el *Corpus Diacrónico del Español* (*CORDE*) en la década de los 90, fundador de la tradición de lingüística de corpus para el español, que incluye textos desde los inicios del español hasta el año 1974, aunque contiene menos de 10 % de textos americanos. En este siglo, otros corpus, que se encuentran en etapas muy desiguales de desarrollo filológico e informático, han venido a fortalecer el conocimiento histórico, textual y dialectal de la lengua española, a los cuales, como es lógico, se suma el *CORDIAM*. Estos nuevos corpus son: el *Corpus Léxico de la Navegación y la Gente de Mar* (*AGILEX*), constituido por documentos tomados del Archivo General de Indias de los siglos XVI al XVIII, el *Corpus de Biblias* (*BIBLIA MEDIEVAL*), compilación de traducciones medievales de la biblia al español, el *Corpus Hispánico y Americano en la Red: Textos Antiguos* (*CHARTA*), que reúne textos archivísticos en español de los siglos XII al XIX, el *Corpus Diacrónico del Español del Reino*

de Granada (*CORDEREGRA*), que reúne documentos de la antigua administración de Granada entre 1492 y 1833, el *Corpus Documental de las Islas Canarias* (*CORDICan*), el *Corpus Léxico de Inventarios* (*CorLexIn*), consistente en una diversidad de textos notariales de los siglos XVI al XVIII, tales como inventarios de bienes de vivos y difuntos, cartas de dote, testamentos, etc., y el *Corpus del Español* (*CE*), también conocido como *Corpus Davies*, que se distingue por estar constituido por un universo de palabras muy grande.

El *CORDIAM* comparte muchas de las propiedades de estos corpus, como era de esperar, pero cuenta también con especificidades relevantes en el tratamiento y la interfaz informáticos, así como innovaciones en la concepción de las categorías de búsqueda, además de su amigabilidad y el escaso tiempo de entrenamiento que su uso requiere. La pantalla inicial del *CORDIAM* contiene diversos archivos de presentación de contenidos, de colaboradores, de fondos documentales, así como una breve *guía rápida*, que rápidamente familiariza al usuario con el adecuado uso informático y filológico del corpus.

Existen, al menos, dos modos de concebir la creación de un corpus histórico lingüístico, las cuales no necesariamente son excluyentes entre sí y tampoco se contraponen: privilegiar la cantidad de textos que se suben o privilegiar la calidad filológica de los textos. Es posible afirmar que el *CORDIAM* ha privilegiado en todo momento la calidad filológica sobre la cantidad. Las razones son varias: en primer lugar, la amplia experiencia investigadora de quienes colaboran en su construcción, ya que todos los materiales, en el caso de *CORDIAM-Documentos*, provienen del trabajo directo en archivo por parte de investigadores en historia de la lengua y de filólogos; no hemos subido, en ningún caso, antiguas compilaciones documentales ya existentes, que aunque pudieran ser de óptima calidad filológica, carecemos de la información sobre el control paleográfico y filológico que establecieron sus compiladores. Por lo que respecta a los corpus de *CORDIAM-Literatura* y *CORDIAM-Prensa*, estos pasan unos filtros de selección explícitos, que presentaremos más adelante. El *CORDIAM* supone, además, una propuesta acerca de cómo compatibilizar la calidad filológica con el aprovechamiento informático. Como mostraremos más adelante, su arquitectura interna está diseñada de manera tal que permite dar acceso explícito al usuario a diversas variables; el acceso a los datos es inmediato y replicable bajo las mismas condiciones —tomando en cuenta, por supuesto, su carácter de corpus abierto—.

El objetivo general de este trabajo es presentar en forma resumida la gestación de este corpus y su estado actual, poniendo énfasis en las características diferenciales del *CORDIAM* frente a otros corpus. Los objetivos particulares son dos: en primer lugar, dar cuenta de las características filológicas, textuales y de variación lingüística que lo integran; en segundo lugar, describir las propiedades

informáticas y de organización del conocimiento diseñadas específicamente para este corpus.

Este trabajo además de esta breve introducción está organizado en tres grandes apartados, que corresponden, en esencia, a los dos objetivos particulares. En el apartado 2, describimos las características filológicas y los criterios tipológico-textuales de conformación, clasificación y organización del corpus. En el apartado 3 exponemos detalladamente las propiedades informáticas, haciendo hincapié, como ya dijimos, en aquellas prestaciones del *CORDIAM* que lo diferencian de otros corpus históricos existentes. En el apartado 4, se describe el proceso de creación del *CORDIAM*. Cierran en 5, a manera de conclusiones, unas consideraciones finales y el señalamiento de próximas etapas en la mejora de este corpus, las cuales permitirán hacer una mejor historia del español en América, pero también una mejor lingüística histórica en general.

2 Características lingüísticas y textuales del *CORDIAM*

La arquitectura *lingüística* de este corpus considera cuatro ejes de variación, en cuyo interior pueden existir otros ejes menores de estructuración: 1. variación diatópica, 2. variación diacrónica, 3. variación autoral, y 4. variación textual. Transversalmente, el *CORDIAM* permite la combinatoria de todas ellas en búsquedas complejas, mediante la combinación de los cuatro ejes o subtipos de esos ejes.

2.1 El ámbito variacional diatópico

Este ámbito incluye dos ejes menores interdependientes: lugar de escritura y lugar de nacimiento del autor del texto en cuestión. En cuanto al primero, todos los textos incluidos en el *CORDIAM* han sido producidos, como ya dijimos, en un contexto americano. Los documentos que integran el *CORDIAM* cuentan con una identificación del país americano actual en que fueron producidos, aunque, como ya señalamos, no todos estos países sean actualmente considerados parte de Hispanoamérica. Proporciona también el *CORDIAM* otra información geográfica menor, que es de interés diacrónico, cultural y filológico: la ciudad, pueblo o villa en que se escribió el documento, texto u obra. En el caso de *CORDIAM-Documentos* se asienta tanto el topónimo histórico como el correspondiente topónimo actual, lo cual, creemos, permitirá al investigador situar el documento en el proceso histórico de conformación de las zonas dialectales de América. Los otros dos subcorpus, Literatura y Prensa, sólo contienen el topónimo actual y carecen del topónimo histórico, porque, las más de las veces,

ambos topónimos, histórico y actual, coinciden, ya que las obras literarias y los periódicos fueron impresos, en su mayor parte, en las capitales de la actual Hispanoamérica, las cuales, en lo general, han mantenido su nombre desde antiguo.

Por lo que toca a los autores de los documentos, los textos hemerográficos y las obras literarias, todos son nacidos en América, con excepción, como es lógico, de aquellos que escribieron en el siglo XVI, puesto que este periodo corresponde, como es sabido, a la llegada de los primeros pobladores europeos. Para garantizar la americanidad del *CORDIAM*, en *CORDIAM-Prensa* solo hemos incluido textos aparecidos en periódicos americanos o reproducidos por estos, excluyendo, por ejemplo, las corresponsalías, en la medida en que no es posible corroborar el carácter americano de los escritores. Los autores de las obras incluidas en *CORDIAM-Literatura* son, en todos los casos, americanos o, para el caso del siglo XVI, habitantes de América desde larga data que escribieron en América; así, para este periodo sí hemos incorporado los cronistas de Indias que escribieron desde América, como Bernal Díaz del Castillo, pero hemos excluido aquellos cronistas que escribieron desde España, no obstante haber vivido en América, tal es el caso de la *Historia general y natural de las Indias* de Gonzalo Fernández de Oviedo.

El estricto control de los dos ejes de este ámbito variacional diatópico, lugar americano y escritor americano o asentado y escribiendo en América, otorga el indudable carácter de corpus de América al *CORDIAM*.

2.2 El ámbito variacional diacrónico

El *CORDIAM* comienza inmediatamente después de la llegada de Colón a América y culmina, dependiendo de los países, en las postrimerías del siglo XIX o muy al comienzo del siglo XX, esto es, cierra cuando ya están constituidos los estados nacionales y se han estabilizado las independencias.

En el caso de *CORDIAM-Prensa*, las restricciones diacrónicas son consecuencia de la aparición de la prensa en América. El primero de los periódicos incluidos en el *CORDIAM* es de 1722, *La Gaceta de* México, el primero publicado en América. A partir de esa fecha, lentamente al principio y en forma explosiva en el siglo XIX, la prensa periódica se desarrolla en América. Hemos incluido, entonces, prensa de los siglos XVIII y XIX.

En el subcorpus *CORDIAM-Literatura* hemos incluido obras literarias entre los siglos XVI al XIX. Cabe hacer algunas precisiones para este subcorpus. En *CORDIAM-Literatura* el criterio ha sido, por sobre cualquier otro eje variacional, la recepción, es decir, el del canon, guiados por la pregunta: ¿qué obras conforman el corpus de literatura hispanoamericana? El criterio de inclusión fue,

por tanto: ser parte del canon + haber vivido el autor en América + haber sido escrita en América la obra en cuestión. Somos conscientes de que se trata de un criterio no lingüístico pero que preserva la identidad americana del subcorpus *CORDIAM-Literatura*. Si sacáramos a los autores no americanos que después del siglo XVI publicaron en América, tendríamos que sacar obras fundamentales del periodo, por ejemplo, el «peruano» Concoloncorvo, o el «novohispano» Agustín de Salazar y Torres. Hay que señalar que no son demasiados casos, en realidad, pues no llegan a una decena en todo el subcorpus *CORDIAM-Literatura*, pero son casos relevantes.

2.3 El ámbito variacional autoral

En los tres corpus que conforman el *CORDIAM* la atribución de autoría es, por diferentes razones, una cuestión problemática. Sin embargo, dado que este corpus ha sido construido desde la óptica del investigador como usuario central, hemos hecho los mayores esfuerzos por recuperar datos de los autores —o de sus características étnicas o de su origen geográfico o de su sexo— en todas las ocasiones en que fue posible.

Así, *CORDIAM-Documentos* incluye datos sobre el sexo, el origen geográfico y el origen étnico del autor del documento. En los casos en que esos datos no son confiables, hemos marcado *sin datos* (s/d) y en los casos en que hay documentos en los que interviene más de un autor se consigna como *varios*. En este subcorpus de documentos, no incluimos el nombre del autor, aunque en algunos casos era recuperable, porque la finalidad básica de este subcorpus no es establecer la identidad de la persona que escribe sino el qué y el cómo lo escribe, esto es, la finalidad de un corpus no literario y no periodístico es, como se sabe, buscar y establecer las grandes rutinas lingüísticas generales de un dialecto o de unos dialectos, sin importar quién escribió ni si se trata de una creación individual o colectiva; en no pocos casos, el nombre del autor está asentado en la síntesis del documento.

En el caso de *CORDIAM-Prensa* se consigna el nombre del autor o de la autora, si este es recuperable; muy pocas veces, por cierto, en la prensa del siglo XVIII, algo más, pero no mucho, en la decimonónica. En todos los demás casos, esto es, en textos no firmados o firmados con nombres genéricos como *un lector*, *el verdadero patriota* —o diversos nombres que pueden asimilarse a los actuales avatares/identidades ficcionales en los comentarios en la prensa digital— se consigna como autor *anónimo*.

Las decisiones para *CORDIAM-Literatura* fueron ligeramente distintas y estuvieron determinadas por las características propias de la creación literaria y

de qué buscaría un usuario en este ámbito textual. En este subcorpus hay mayor información autoral que en los subcorpus Documentos y Prensa, hecho lógico en tanto que la literatura es el ámbito de la creación individual, de la explotación estilística de las posibilidades gramaticales e interesa, por lo general, en su análisis quién es el autor, su sexo, si usaba o no seudónimos, etc. Por ello, el autor de las obras literarias es consignado de acuerdo con la firma de la obra, coincida o no con su nombre de pila, y como *anónimo* en los casos en que así corresponde. En *CORDIAM-Literatura* hemos incluido el sexo del autor, ya que es sabido que hay autoras que firmaban como hombres o, más comúnmente, autores que firmaban como mujeres. Aunque la finalidad esencial del *CORDIAM*, como dijimos al inicio de este trabajo, es conocer mejor el español en América y aportar más y mejores evidencias para el español general y para la teoría lingüística, nos pareció pertinente incluir este tipo de información literaria extralingüística, ya que puede ayudar a conocer más finamente algunos de los entramados culturales americanos.

2.4 El ámbito variacional textual

El cuarto eje de variación considerado en el *CORDIAM* es el textual. En las primeras recomendaciones sobre condiciones mínimas para un corpus lingüístico *EAGLES* (Expert Advisory Group on Language Engineering Standards) señalaba la pertinencia de «una clasificación intermedia entre el corpus mayor y las ocurrencias individuales» (*EAGLES* 1996). Esta ha sido una preocupación central del *CORDIAM* desde sus inicios, por razones prácticas y por consecuencias metodológicas de la teoría del cambio lingüístico, porque uno de los objetivos en la clasificación textual de un corpus es evitar la atomización de los resultados del análisis, atomización que es casi consustancial al quehacer en gramática histórica, disciplina en que obtener generalizaciones es, como se sabe, casi un reto, en tanto que el cambio lingüístico suele operar como pequeñísimos y casi imperceptibles microquiebres, que reajustan mínimamente la esencial continuidad histórica de las lenguas. Por todo ello, era un imperativo tipologizar el corpus.

Ya en la primera etapa del *CORDIAM* se avanzó en la tipologización de los documentos de archivo (cf. Bertolotti/Company 2014) y actualmente hemos establecido ya tipos textuales para los otros dos subcorpus, de manera que cada una de las unidades hemerográficas y literarias está adscrita a un tipo textual, que puede ser seleccionado por el usuario de manera individual, por una o más de las subclases textuales, de manera general, por uno o más de los subcorpus en su totalidad, o de manera global, marcando *todo*. Cuando hablamos de *unidades* de obras literarias o de prensa, nos referimos al hecho de que hemos

Tabla 1: Tipos textuales del *CORDIAM*

Documentos	*Literatura*	*Prensa*
Administrativos	Narrativos	Comentativos
Cronísticos	Poéticos	Informativos
Jurídicos	Teatro	Publicitarios y anuncios varios
Entre particulares: cartas y otros	Textos cronísticos	
	Prosa varia	

fragmentado las obras y los periódicos que integran *CORDIAM-Literatura* y *CORDIAM-Prensa* en unidades, tanto para permitir una tipologización más fina, cuanto para evitar reproducciones completas de obras, ya que, como veremos en el apartado 3, el *CORDIAM* permite ver y bajar cualquier unidad documental completa contenida en el corpus. En el caso de *CORDIAM-Literatura*, la unidad es, por lo general, el capítulo, el acto teatral o el poema, o bien la unidad que establece la edición empleada; en el caso de *CORDIAM-Prensa*, la unidad es el fragmento hemerográfico que tiene autonomía comunicativa o informativa, que está contenido en márgenes gráficamente identificados y que es visualmente reconocible como una unidad.

La confección de las tipologías supone la resolución de un importante número de problemas teóricos y prácticos, algunos de los cuales exceden el sentido de la elaboración de una tipología para un corpus en particular. En el *CORDIAM* hemos tomado una definición operativa. Entendemos por *tipo textual* un conjunto de clases o géneros con un rasgo externo común: la finalidad comunicativa del texto en cuestión, a la cual deben sumarse ciertas regularidades internas en la recurrencia de secuencias (descriptivas, narrativas, argumentativas y dialógicas), en la temática, en el léxico, en la sintaxis y en la morfología.

La tipología del *CORDIAM* incluye 12 tipos textuales, cuatro tipos para el subcorpus de Documentos, cinco tipos para el subcorpus de Literatura y tres tipos para el subcorpus de Prensa, como se puede ver en la tabla 1[1].

Frente a otros corpus que contienen clases textuales muy superiores a 50 etiquetas, la aparente simplicidad tipológica textual del *CORDIAM* es, a nuestro modo de ver, una gran ventaja, porque permitirá al usuario comparar, con cierta

1 La tipologización del *CORDIAM* fue una decisión discutida y consensuada entre las autoras de este trabajo, directoras del proyecto y corpus *CORDIAM* a la vez que coordinadoras del subcorpus Documentos, Magdalena Coll, coordinadora de *CORDIAM-Literatura*, y Jorge Gutiérrez Reyna, coordinador de *CORDIAM-Literatura*.

facilidad, diferencias y similitudes —grafemáticas, léxicas, gramaticales o discursivas, o un conjunto de varias de ellas— entre clases. Para realizar esta tipologización, hemos vuelto a poner en foco la necesidad de capturar generalizaciones lingüísticas a partir de un corpus histórico.

La tipología propuesta en la tabla 1 surge del análisis de los datos existentes, de los géneros textuales de forma previa a ser etiquetados textualmente y de las recurrencias léxicas, gramaticales y/o discursivas en diversos niveles, y estuvo guiada en todo momento por la pregunta sobre qué y cómo buscaría un usuario en un corpus electrónico. Los procedimientos para tipologizar los tres subcorpus tienen aspectos comunes pero también algunas diferencias entre sí, porque el objetivo central ha sido hacer transparente las diferencias entre las grandes clases textuales establecidas para los tres subcorpus. El procedimiento para el establecimiento de las clases en *CORDIAM-Documentos* fue básicamente inductivo: agrupamos los documentos, por regularidades lingüísticas —la forma del documento— relacionadas con su forma de circulación —pública o privada— y por el ámbito social y estilo comunicativo —familiar-cercano vs. distante—. En el caso de *CORDIAM-Prensa* y de *CORDIAM-Literatura*, también a través de un proceso inductivo, nos pareció razonable tipologizar recuperando, en la medida de lo posible, las tradiciones de los estudiosos de la literatura —en parte, arriba comentadas— y del periodismo, ya que es razonable suponer que los usuarios del *CORDIAM* las conozcan y las puedan emplear intuitivamente. Presentamos a continuación en forma muy sucinta los diversos tipos textuales, primero los de *CORDIAM-Documentos*, luego los de *CORDIAM-Literatura* y, finalmente los de *CORDIAM-Prensa*.

Los *documentos administrativos* tienen como rasgo común ordenar, registrar, disponer y regular la interacción no privada entre personas. Contienen descripciones y lineamientos sumamente detallados de las diversas facetas de la vida cotidiana. Dan también cuenta de bienes materiales, de vivos y de difuntos; dan testimonio de la genealogía de los individuos. Son documentos de circulación pública. En (7) se ejemplifica esta clase textual.

(7) antes que vos a el llegueys el dicho contreras deve yr primero por le asegurar diziendole que vos vays A el por le ver e conosçer e tener con el Amistad. porque yendo vos con mucha gente podria ser que tomase Reçelo e se pornia A yr por los montes e herrariedes la presa. (República Dominicana, año 1494, documento administrativo, *CORDIAM*).

Los *documentos cronísticos* describen paisajes, relatan sucesos raros o curiosos, describen acciones propias de ciertos grupos humanos, sus creencias, sus costumbres, sus festividades y sus comportamientos. Tienen, muchas veces, una

ordenación temporal. Se acercan a un texto literario sin tener, sin embargo, una finalidad estética. Predominan en ellos descripciones y narraciones. Son documentos de concepción pública por lo regular. Aparecen ejemplificados en (8).

(8) vnos salieron de qüebas, los otros de çerros, /[25] y otros de fuentes, y otros de lagunas y otros de pies de árboles,/y otros desatinos desta manera; y que por auer salido y enpeçado/a muntiplicar destos lugares y auer sido de allí el prinçipio/de su linaje, hizieron guacas y adoratorios estos lugares/en memoria del primero de su linaje que de allí proçedió; /[30] y así cada nación se uiste y trae el traje con que a su guaca/uestían. (Perú, año 1600, documento cronístico, *CORDIAM*).

Los *documentos jurídicos*, producidos en el mundo legal, se acercan, en algunos aspectos, a los documentos administrativos. A diferencia de estos, sin embargo, son textos complejos y heterogéneos, ya que contienen clases dentro de otras clases (denuncias, querellas, postulación de preguntas, interrogatorios y respuestas, sentencias, segundas instancias de juicios, traslados, etc.). Suelen ser identificados como una unidad, no obstante, pueden contener otras unidades. Algunas de ellas suelen ser altamente dialógicas y se aproximan, por ello, a la oralidad. Contienen pasajes concebidos para su circulación pública y otros que no son públicos. Aparece un ejemplo en (9).

(9) Y habiendo ido a dicha/cassa reconosió ser la bolsa como se le abía notisiado, y que/preguntó a la dicha mulata si era verdad que ella traía /[20] aquella bolsa, y dixo que sí, que la traía buenamente y sin/ninguna malisia, que sólo sí sabía de una llerba que en/la bolsa traía ser buena para sanar de sus picadas de ormigas,/la qual dicha bolsa exibe ante el señor comisario. (México, año 1704, documento jurídico, *CORDIAM*).

El cuarto tipo, los *documentos entre particulares: cartas y otros*, es producto de la comunicación entre dos particulares y llegan a los archivos, en general, por azar, como documento probatorio de algún tipo de proceso. Suelen tocar tópicos personales, de carácter íntimo y afectivo con frecuencia, como se aprecia en (10). Muestran una elevada *inmediatez comunicativa* por tratarse de cartas, notas, recados, etc. en las que el interlocutor está implícitamente presente. Son el único tipo de documento que se atreve a escribir quien no domina las técnicas escriturales.

(10) Mi muy querida M*adr*e y estimada Hermanita la/Saludo con La paz de n*uestr*o S*eñ*or la que deseo avite /[5] de firme en su corazon como asi lo creo yo me alegrare/que se mantenga con Salud en compañía de mi/Padre D*o*n Josef M*aría* y desde el dia que nos ablamos/no ayga tenido noveda en la Salud. (Uruguay, año 1810, documentos entre particulares, *CORDIAM*).

Por su parte, las obras de *CORDIAM-Literatura* se adscriben a categorías tomando en consideración la forma y la «finalidad estética», con todos los problemas conocidos en la discusión centenaria acerca de los géneros. Entendemos en el *CORDIAM* que una obra *narrativa* es un texto en prosa cuya finalidad es contar una serie de acciones realizadas por un determinado número de personajes. El contenido de la narración es ficcional. Este rubro se compone por los escasos ejemplos de la novela virreinal, la novela decimonónica, la novela corta y el cuento moderno, surgido después de la mitad del siglo XIX. Se consideran dentro de este tipo narrativo las narraciones ejemplares como la fábula neoclásica. Las obras del tipo *poético* están escritas en verso, a saber, son unidades de sentido que atienden a los aspectos fundamentales de la métrica española: medida, ritmo y rima. Este rubro incluye la poesía lírica —vehículo de expresión de la subjetividad del poeta— y la épica —vehículo para narrar una serie de acontecimientos—. El tercero de los tipos, la *prosa varia*, incluye aquellas obras que son escritas con una plena conciencia estética pero que, por una razón u otra, no encajan del todo en las divisiones genéricas actuales de la literatura. Suele tratarse de textos con finalidad didáctica. Este rubro incluye la literatura homilética (sermones y homilías), diálogos, sátiras, arcos triunfales y manifestaciones tempranas del ensayo moderno. El tipo *teatro* incluye obras escritas en prosa o en verso para la representación escénica. Este tipo se diferencia de otros géneros que contienen diálogos por el hecho de que la clase textual *teatro* sí está concebida para su representación. Por fin, las obras *cronísticas* son textos en prosa que buscan dar cuenta de acontecimientos ocurridos, o no, en orden primordialmente cronológico. Suelen tener como finalidad probar la historicidad o veracidad de los hechos relatados. En este rubro entran las crónicas de conquista, las crónicas de órdenes religiosas, de conventos o de provincias, las relaciones, las memorias y los textos hagio-biográficos (vidas de monjas y frailes venerables, por ejemplo). Sin duda, los límites entre documentos cronísticos y obras cronísticas no es nítido; si el autor y la obra en cuestión forman parte del canon de la literatura hispanoamericana, la obra fue incluida como crónica de *CORDIAM-Literatura*. No ejemplificamos los subtipos textuales de este segundo subcorpus porque apenas están siendo procesadas las primeras obras para su tratamiento informático.

Por último, la tipología textual de los textos hemerográficos es, tal como las tipologías anteriores, operativa con respecto al usuario y se basa en la intención del escritor del texto que debe ser tipologizado. Como señalábamos antes, una tipología para la prensa supone combinar las regularidades textuales con la existencia de géneros «periodísticos» en ese medio de comunicación de masas (y, por cierto, con los cambios que estos han tenido). Pongamos por ejemplo

el género editorial, esto es, un texto que da cuenta de la opinión del medio, en general, sobre un acontecimiento noticioso y la carta al redactor. Ambos textos, a pesar de pertenecer a clases o a géneros[2] diferentes, van a estar juntos en la tipología de *CORDIAM-Prensa* dentro del tipo *comentativos*, ya que comparten buena parte de sus características textuales, gramaticales y léxicas.

Como puede verse en la tabla 1 arriba, la tipología con la que clasificamos los textos de *CORDIAM-Prensa* tiene tres categorías: *comentativos*, *informativos*, y *publicitarios y anuncios varios*. Incluimos como *comentativos* aquellos textos publicados en la prensa periódica cuya intención es hacer pública la opinión del periódico o de un autor particular. El fin último de estos textos es persuasivo, como se aprecia en (11), ya que el autor espera que el lector comparta su opinión, o espera influir en su opinión. En este medida, tienen una cuota de subjetividad, explícita o implícita. Los textos comentativos suelen tomar la forma de editoriales, artículos de opinión, discursos públicos, cartas al periódico, polémicas, entre otros.

(11) ¿Quiénes serán entonces los modelos de la finura, las normas del bello escribir, los espejos de la exquisita erudiccion, los luminares de las Ciencias, los censores de los Papeles públicos, los Criticos de por vida? ¡Qué catastrofe tan funesto! ¿Adonde iremos que nos corrijan nuestros errores? Pero ¿para qué viene aquí la funebre Melpomene á llenarnos de su entusiasmo doloroso? Ea, dexame Musa: déxame: vuelvete á tu Parnaso en hora buena, que hoy solo es dia de hablar sencillamente la verdad. (Colombia, año 1791, texto comentativo, *CORDIAM*).

Consideramos textos *informativos* aquellos cuya intención principal es la de aportar novedades al público, sean aquellas datos o noticias, de índole natural, social, cultural, científica, tecnológica o política, como se muestra en (12). Los textos informativos suelen tomar la forma de noticias, crónicas, artículos costumbristas, *laudatios*, efemérides, divulgación de descubrimientos, entre otros.

(12) Asimismo 14 leguas al Hueste, del dicho cabo, de dicha Costa vinieron dos Indios a bordo, cada vno con su embarcacioncilla, ó por mejor dezir de tres palos amarrados, pero no se les pudo entender, ni ellos á los nuestros. Todos son bien agestados, y corpulentos, sin rayas, ni otra cosa en todo su cuerpo. (México, año 1722, texto informativo, *CORDIAM*).

2 El establecimiento de etiquetas que sean descriptivas y no resulten anacrónicas no es siempre una tarea fácil. Por ejemplo, la *carta del lector* como comentario a un artículo ya publicado, no aparece en el siglo XVIII; es más, siquiera aparece en el siglo XVIII el concepto de lector, tal como lo conocemos ahora.

Los textos que caben bajo la rúbrica *publicitarios y anuncios varios* son aquellos cuya intención es hacer pública la existencia de un bien físico, cultural o un servicio cuyo conocimiento puede llevar a una acción por parte del lector, como comprar, rematar, suscribirse, contratar… Estos textos suelen tener la forma de avisos, anuncios, ofrecimientos, como en (13).

(13) Esperamos que nuestros amigos, y en general los literatos del pais nos favorezcan con inscribirse en el núm. de nuestros corresponsales. De este modo podrémos, sin temer el paralelo, ocupar un puesto digno entre los coescritores nuestros que están iluminando á la patria con el fanal de sus luces. (Argentina, año 1821, texto publicitario, *CORDIAM*).

Estas tres tipologías así establecidas son una marca distintiva de este corpus y son, como ya dijimos, a nuestro parecer, un valor agregado importante del *CORDIAM*. Otros corpus históricos, como el *CORDE*, el *CHARTA* o el *Corpus Davies*, no ofrecen una tipología propiamente dicha, sino que ordenan en clases o géneros muy atomizados y por temas o entornos socio-discursivos, sin definir qué entienden por unidad o tipo textual. El *CORDE*, por ejemplo, clasifica sus textos de acuerdo con: *a*) la *procedencia*: *libros, periódicos, revistas, miscelánea y orales,* y *b*) el tema, que tiene numerosas divisiones internas*: lírica* (5 *subclases*), *narrativa* (7), *teatro* (7), *didáctica* (9), *ciencia y técnica* (58), *sociedad* (36), *religión* (8), *prensa, historia y documentos* (10), *derecho* (5), *culto* (3), *dramático* (4)… etc. Este conjunto de subdivisiones termina ofreciendo al usuario más de 140 posibilidades para elegir y es el usuario quien tiene que agrupar para buscar similitudes y diferencias y obtener generalizaciones de evidencia empírica robusta. Por su parte, el *CHARTA* tipologiza por grandes clases textuales que agrupan subclases: actas y declaraciones —con 13 subclases internas—, cartas de compraventa y contratos —con 11 subclases internas—, cartas privadas —con 3 subclases internas—, certificaciones, estatutos, informes y relaciones —con 8 subclases internas—, notas breves, otros —con 6 subclases internas—, testamentos e inventarios —con 9 subclases internas— y textos legislativos —con 5 subclases internas—. Esto es, trabaja con más de cincuenta subclases agrupadas en 7 clases. El *Corpus del Español* de Davies, si bien contiene muchos datos históricos, sus agrupamientos son solo aplicables en su totalidad al siglo XX, ya que son *oral*, *ficción*, *periodístico* y *académico*.

La tipología que hemos establecido, como casi cualquier tipología, no escapa al problema de los límites que pueden resultar insuficientemente nítidos para algunos ejemplares, que se ubican, entonces, parcialmente en un tipo y parcialmente en otro. En otras palabras, en los textos incluidos en el *CORDIAM* existen ejemplos prototípicos y casos que se encuentran en los márgenes. Tomando

como ejemplo la prensa, mostramos a continuación dos ejemplos de casos que están en los límites y damos cuenta de cómo resolvimos la adscripción tipológica.

El primer ejemplo, (14), tomado de la publicación argentina *La Camelia* (4 de mayo de 1852) es un comunicado de las redactoras del periódico sobre su medio de comunicación.

(14) {p. 2} Nos es grato anunciar á nuestras suscriptoras que con motivo de poseer nuevos elementos, la redaccion de la Camelia ofrecerá inmediatamente las mejoras indispensables que comprendemos bien que necesita— // (…) La nueva organizacion que tiene que recibir nuestro periódico se irá desenvolviendo progresivamente (…) // No nos apartarémos, sin embargo de la senda que hasta hoy hemos seguido, respecto de nuestro secso—Y las correspondencias que insertémos serán generalmente las que no encierren ofensa personal en ningun sentido, ni materias inmorales bajo ningun aspecto. (Argentina, año 1852, texto informativo, *CORDIAM*).

¿Qué aspectos de este texto nos pueden llevar a dudar? Si bien por un lado informan sobre la nueva estructura del periódico, el texto tiene un componente publicitario, ya que está anunciando mejoras en el periódico y por lo tanto se puede suponer que está persuadiendo a los lectores, tal como es la función primordial de la publicidad. Tiene, por otra parte, un componente comentativo: hay una valoración y una subjetividad explícita, que se expresa en el último párrafo. Sin embargo, decidimos tipologizarlo como *informativo*, ya que entendemos que prioritariamente está informando sobre una reestructura de las secciones del diario.

El segundo de los casos de límites textuales borrosos aparece en (15). Es un texto que hemos nombrado como *artículo de opinión*, publicado en 1894 en el diario *Plus-Café* de Bogotá. La cuestión en este caso es también de límites pero por razones distintas que en el ejemplo anterior.

(15) {p. 2} \\\EL CÓLERA MORBO EN BOGOTÁ\\\ // Varios casos de este formidable azote han ocurrido en Bogotá i nos proponemos referirlos á nuestros parroquianos para que por medio de la eficaz receta que publicamos, se pongan á salvo ántes i con ántes de la aparicion de esa cruel i esterminadora enfermedad. El procedimiento es sumamente sencillo i con el favor de Dios, pueden lograrse los mejores resultados. Es tiempo ya de reseñar los casos a que nos contraemos. ¡Atencion! //El primero ocurrió en un estimable compatriota nuestro que fué removido de su empleo segun pública vos i fama por ser notoriamente contrario á la administracion del 7 de marzo; pero el pasiente no apeló al antídoto tipos de que otros han hecho un uso tan inmoderado que han llamado ácia sí el anatema de su archivamiento por lo ménos en el cuatrienio que dió principio el 1. de abril. Entre estos últimos merecen especial mencion tres jueces de la Corte de cuentas que dominados de lleno por el cólera han emprendido la desabrida tarea de una larga i cansada frotacion polémica (…)// \\\

> RECETA ANTI-COLERICA.\\\ // El empleado amovible por el Poder Ejecutivo que se sienta con síntomas i señales de ser removido de su destino, cálese sus anteojos, si los necesitare, tome un pedaso de papel i corra sobre él una pluma que forme los caractéres precisos para RENUNCIAR. (Colombia, año 1894, texto informativo, *CORDIAM*).

El texto tiene la forma de un texto informativo, sin embargo su intención es fuertemente subjetiva; ofrece un procedimiento de renuncia y, por lo tanto, no busca meramente informar sino también crear opinión respecto de alguna cuestión. Aunque tiene la apariencia de ser un texto informativo, su sentido es comentar, opinar, sobre una situación concreta. Pese a ello, en este caso, optamos por asignarlo al tipo *informativo*, ya que su sentido comentativo es pragmático, puesto que surge del hecho de poner la forma del texto en relación con alguna circunstancia política —que además ignoramos—. Si bien es un texto «falsamente» informativo, privilegiamos esta categoría porque genéricamente parece tal y las selecciones textuales, léxicas y gramaticales son las propias de un texto informativo. En otras palabras, ante la contradicción entre la forma y el sentido, optamos por la forma, ya que, como el lector recordará, la tipología está concebida, entre otras razones, para facilitar al usuario la captura de regularidades o rutinas lingüísticas.

De acuerdo con lo que hemos presentado en este apartado, *CORDIAM* captura parte de la variación inherente a las lenguas naturales y a los discursos que en ellas y con ellas se construyen, y permite superar, entendemos, la así llamada *paradoja de Enrique* (Kabatek 2016: 5). Enrique-Arias señala que:

> Una paradoja de la composición de los corpus diacrónicos es que, por un 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) (Enrique-Arias 2012: 96, en Kabatek 2016: 5).

La forma en que concebimos el *CORDIAM*, y su traducción informática permite que cada usuario pueda construir tantos corpus como desee, basándose para ello en las diferentes variables, y sus valores, identificadas a través de los metadatos. En otras palabras, el *CORDIAM* no es un corpus plano y homogéneo, en la medida en que ofrece la posibilidad de hacer selecciones multiangulares al servicio del objeto bajo investigación.

Por ejemplo, en la pantalla de *CORDIAM* capturada en la ilustración 1 a continuación, podemos ver una búsqueda realizada con las restricciones en algunos de los campos de búsqueda referidos en los párrafos anteriores. La ilustración 1 muestra la búsqueda de cualquier palabra comenzada con *barc*, escrita en el siglo

Ilustración 1: Búsqueda *barc** con restricciones de siglo, de país y de subcorpus

XVI en Nicaragua, Panamá, Puerto Rico, República Dominicana, El Salvador o Venezuela y que estuviera escrita en un texto de *CORDIAM-Documentos*.

3 Características informáticas del *CORDIAM*

El *CORDIAM* proporciona una interfaz concebida desde el proceso de investigación, con decisiones teóricas y metodológicas surgidas de la experiencia de investigación en lingüística histórica e historia de la lengua. Cuenta con siete características principales, que son prestaciones del *CORDIAM*: 1. una extensa plantilla de metadatos visibles en cada concordancia, y buscables; 2. la visualización en una ventana lateral de un contexto de 100 palabras en torno a la concordancia; 3. el acceso al documento completo y la posibilidad de guardarlo e imprimirlo; 4. la posibilidad de generar automáticamente una base de datos con la búsqueda realizada; 5. la información cuantitativa instantánea comparable con universos de palabras totales y parciales; 6. el ordenamiento variado de la información obtenida, y 7. búsquedas lematizadas, simples y complejas, continuas y discontinuas —en este momento, el *CORDIAM* está lematizado en un 70 % aproximadamente—. Comentaremos con detenimiento aquellas propiedades que son prestaciones del *CORDIAM*.

3.1 Plantilla de metadatos

Cada uno de los textos del *CORDIAM* cuenta con una plantilla de metadatos. Si comparamos los tres tipos de plantillas que corresponden a los tres subcorpus,

notaremos ligeras divergencias, las cuales permiten dar cuenta cabalmente de las especificidades del subcorpus en cuestión.

Todos los metadatos proveen información relevante para la contextualización del documento y algunos de ellos, como ya vimos, reflejan variables sociales o textuales. La plantilla de *CORDIAM-Documentos* contiene 16 metadatos, siete de los cuales permiten búsquedas. Los ejemplificamos a continuación y resaltamos en versalitas aquellos metadatos con los cuales el usuario puede acotar sus búsquedas.

1. Nombre: Corresponde o bien al nombre del documento en el corpus fuente (publicado o inédito) o bien al nombre asignado por el equipo de filología del *CORDIAM*.
2. SIGLO: Es el siglo en que el documento fue redactado o copiado.
3. AÑO: Es el año en que fue redactado o copiado el documento. Si no se tiene certeza sobre este dato, se incluye ca., esto es, circa.
4. AUTOR (datos étnicos): Indica los datos étnicos del autor del documento, cuando están disponibles. Esta etiqueta admite ocho posibilidades: indio, mestizo, español, criollo, extranjero no hispanohablante, negro, varios y s/d. Varios significa más de un autor y s/d significa que no se cuenta con la información.
5. AUTOR (hombre o mujer): Esta etiqueta admite cuatro posibles opciones: hombre, mujer, varios y s/d.
6. PAÍS ACTUAL: Corresponde a la denominación geopolítica actual del lugar físico en donde se escribió el documento. Abarca los actuales 19 países americanos hispanohablantes más el sur y oeste de Estados Unidos, Jamaica, Haití y Guyana (antiguos territorios de la corona española). El país se consigna abreviado en mayúsculas siguiendo el uso más común.
7. Topónimo actual: Es la denominación geopolítica actual del lugar físico en donde se redactó el documento (provincia, departamento, ciudad, pueblo, etc.).
8. Topónimo histórico: Es la denominación histórica del lugar en donde fue escrito el documento.
9. Adscripción histórica: Esta etiqueta indica a qué virreinato o adscripción político-administrativa correspondía el documento en el periodo virreinal/colonial. Cuando se posee la información, se añade el dato inmediato inferior (por ejemplo, audiencia, capitanía, provincia, gobernación).
10. Archivo: Corresponde al nombre del fondo documental en donde se encuentra el original y se consigna la ubicación dentro del archivo, de acuerdo con los datos proporcionados por los investigadores.
11. TIPO TEXTUAL: Corresponde a la clasificación tipológica textual del documento.
12. Número de folios: Indica la cantidad de folios que tiene el documento.
13. Número de palabras: Indica el número aproximado de palabras del documento.
14. Créditos: Consigna la referencia bibliográfica correspondiente a cada uno de los corpus que conforman *CORDIAM-Documentos*.
15. Facsimilar disponible: Indica si el facsímil está o no disponible para su futura incorporación.
16. Síntesis: Proporciona una breve descripción del contenido del documento.

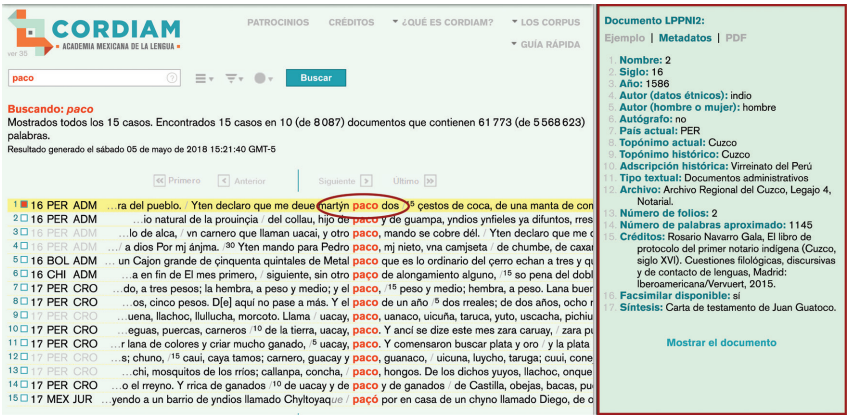


Ilustración 2: Metadatos en la ventana lateral de la primera concordancia de *paco*

Un ejemplo de una plantilla de metadatos de *CORDIAM-Documentos* completa se puede observar en la primera captura de pantalla de la ilustración 2, donde se pueden observar los metadatos desplegados de la quinta concordancia de la búsqueda *paco*. La observación del desplegado de los metadatos en la ventana lateral, permite la toma de decisiones de investigación, por ejemplo, cuáles concordancias conservar y cuáles eliminar, así como contextualizar mejor el documento, obra o unidad hemerográfica en donde se encuentra la concordancia.

La plantilla de *CORDIAM-Literatura* contiene 11 metadatos. Cinco de ellos permiten búsquedas. Los ejemplificamos a continuación resaltando en versalitas los metadatos con los cuales el usuario puede acotar sus búsquedas.

1. Nombre: 12
2. SIGLO: 17
3. AÑO: 1691
4. AUTOR (nombre): Sor Juana Inés de la Cruz
5. Primera edición: Puebla, Diego Fernández de León, 1691
6. PAÍS ACTUAL: MEX
7. TIPO TEXTUAL: Poesía
8. Número de palabras aproximado: 2013
9. Créditos: Jorge Gutiérrez Reyna, *Los villancicos de sor Juana: edición crítica, introducción y notas*, México: Universidad Nacional Autónoma de México, 2016. Tesis de maestría inédita.
10. Facsimilar disponible: sí
11. Síntesis: Villancicos con que se solemnizaron en la santa iglesia y primera catedral de la ciudad de Antequera, valle de Oaxaca, los maitines de la gloriosa mártir santa Catarina este año de mil seiscientos y noventa y uno.

Cabe realizar algunas aclaraciones. En el caso de *nombre*, este corresponde a la unidad documental. En el caso de obras conocidas o editadas, se sigue la denominación más usual. De ser inéditas, se le asigna un número. En el campo *año*, se busca incluir el año de la redacción de la obra; este dato se complementa con el de *primera edición*, ya que puede existir una gran distancia entre la fecha de redacción y de publicación.

La plantilla de *CORDIAM-Prensa* asociada a cada texto contiene 14 metadatos con información de diversa naturaleza. Cinco de ellos permiten búsquedas. Ejemplificamos a continuación una plantilla resaltando en versalitas los datos buscables. Las aclaraciones pertinentes en este subcorpus corresponden al rubro *nombre*, que designa la unidad textual y que orienta al usuario sobre el género textual y el contenido, acompañado de un código de edición que remite al periódico, país y siglo.

1. Nombre: Noticias económicas. GDMiii56-MEX18
2. Periódico: Gazeta de México
3. AUTOR (nombre): anónimo
4. SIGLO: 18
5. AÑO: 1784
6. Fecha: 14-01-1784
7. PAÍS ACTUAL: MEX
8. Topónimo actual: Ciudad de México
9. TIPO TEXTUAL: Documentos informativos
10. Fondo documental (Archivo): Hemeroteca Digital Nacional de México.
11. Número de palabras aproximado: 267
12. Créditos: http://www.hndm.unam.mx/.
13. Facsimilar disponible: sí
14. Síntesis: Informe de la Real casa de moneda y sobre las tarifas de algunos productos.

3.2 Generación automática en una base de datos

Una vez seleccionado un conjunto de concordancias, se puede generar automáticamente una base de datos, que es posible guardar, si se desea, en cualquier dispositivo personal y que, una vez descargada, puede ser personalizada de múltiples formas por el investigador, de acuerdo con las necesidades de este.

En la ilustración 3 abajo se muestra el resultado de la búsqueda *amor*, la selección de aquellas concordancias en que *amor* está ligado a lo divino y se ve el cursor apoyado en *guardar marcados*.

En la siguiente pantalla (ilustración 4), se muestra cómo, una vez guardadas las concordancias deseadas, se debe situar el cursor y abrir con el botón derecho del *mouse* en *Mostrar en carpeta* para que la base se abra con el programa

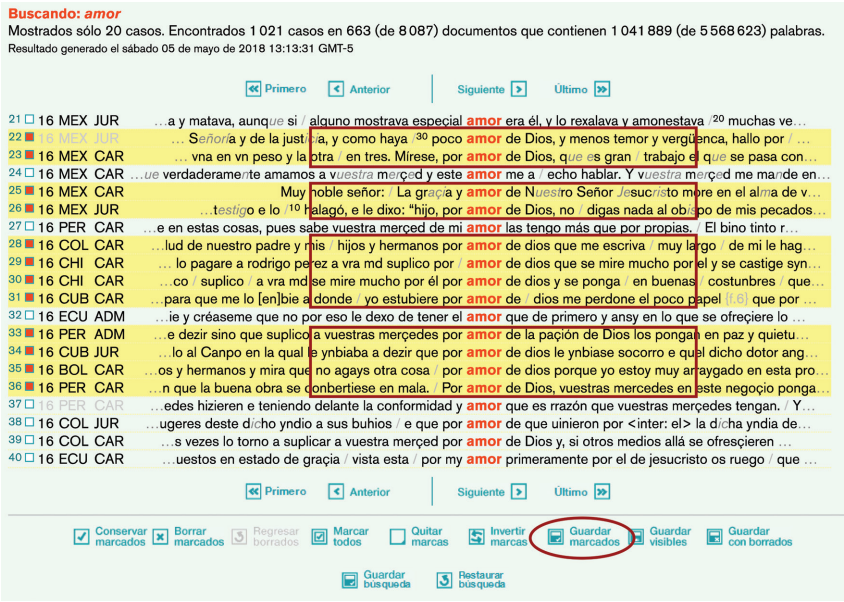


Ilustración 3: Marcado de concordancias y guardado de marcados

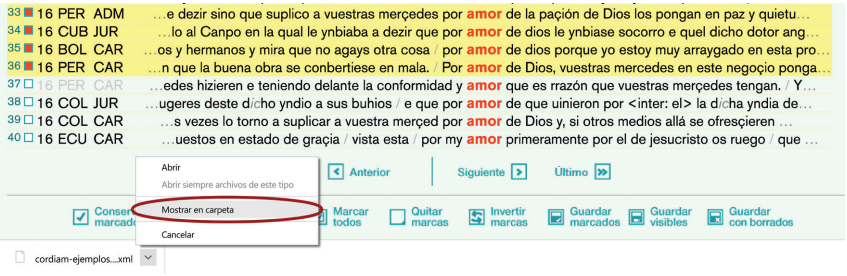
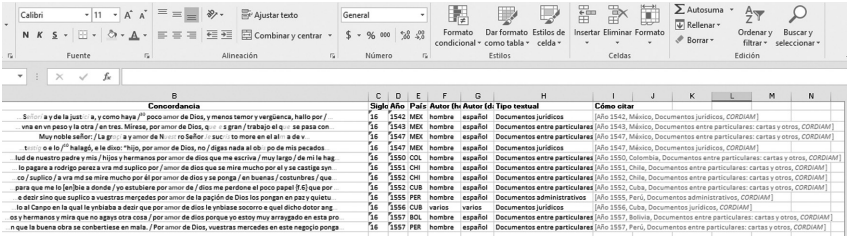


Ilustración 4: Cómo abrir la base de datos de las concordancias guardadas

del usuario. Esto se debe a que la programación se adecua a cualquier tipo de paquete informático que tenga el dispositivo del usuario de *CORDIAM*.

La ilustración 5 abajo muestra el resultado en la base de datos *Excel* —que podría ser también la de otro paquete informático—, modificable por el usuario, quien puede agregar columnas, cambiar la denominación, etcétera.



Concordancia	Siglo	Año	País	Autor (ho	Autor (dat	Tipo textual	Cómo citar
S[illegible] a y de la just[illegible] a, y como haya /[30] poco amor de Dios, y menos temor y vergüenca, hallo por /	16	1542	MEX	hombre	español	Documentos jurídicos	[Año 1542, México, Documentos jurídicos, *CORDIAM*]
vna en vn peso y la otra / en tres. Mírese, por amor de Dios, q[illegible] s gran / trabajo el q[illegible] se pasa con	16	1543	MEX	hombre	español	Documentos entre particulares	[Año 1543, México, Documentos entre particulares: cartas y otros, *CORDIAM*]
Muy noble señor: / La gr[illegible] a y amor de N[illegible] ro Señor [illegible] suc[illegible] to more en el al[illegible] a de v	16	1547	MEX	hombre	español	Documentos entre particulares	[Año 1547, México, Documentos entre particulares: cartas y otros, *CORDIAM*]
t[illegible] o e lo /[40] halagó, e le dixo: "hijo, por amor de Dios, no / digas nada al ob[illegible] po de mis pecados	16	1547	MEX	hombre	español	Documentos jurídicos	[Año 1547, México, Documentos jurídicos, *CORDIAM*]
lud de nuestro padre y mis / hijos y hermanos por amor de dios que me escriva / muy largo / de mi le hag	16	1550	COL	hombre	español	Documentos entre particulares	[Año 1550, Colombia, Documentos entre particulares: cartas y otros, *CORDIAM*]
lo pagare a rodrigo perez a vra md suplico por / amor de dios que se mire mucho por el y se castige syn	16	1551	CHI	hombre	español	Documentos entre particulares	[Año 1551, Chile, Documentos entre particulares: cartas y otros, *CORDIAM*]
co / suplico / a vra md se mire mucho por él por amor de dios y se ponga / en buenas / costunbres / que	16	1552	CHI	hombre	español	Documentos entre particulares	[Año 1552, Chile, Documentos entre particulares: cartas y otros, *CORDIAM*]
para que me lo [en]bie a donde / yo estubiere por amor de / dios me perdone el poco papel [f.6] que por	16	1552	CUB	hombre	español	Documentos entre particulares	[Año 1552, Cuba, Documentos entre particulares: cartas y otros, *CORDIAM*]
e dexir sino que suplico a vuestras mercedes por amor de la pación de Dios los pongan en pas y quietu	16	1555	PER	hombre	español	Documentos administrativos	[Año 1555, Perú, Documentos administrativos, *CORDIAM*]
lo al Canpo en la qual le ynbiaba a dezir que por amor de dios le ynbiase socorro e quel dicho dotor ang	16	1556	CUB	varios	varios	Documentos jurídicos	[Año 1556, Cuba, Documentos jurídicos, *CORDIAM*]
os y hermanos y mira que no agays otra cosa / por amor de dios porque yo estoy muy arraygado en esta pro	16	1557	BOL	hombre	español	Documentos entre particulares	[Año 1557, Bolivia, Documentos entre particulares: cartas y otros, *CORDIAM*]
n que la buena obra se conbertiese en mala. / Por amor de Dios, vuestras mercedes en este negocio ponga	16	1557	PER	hombre	español	Documentos entre particulares	[Año 1557, Perú, Documentos entre particulares: cartas y otros, *CORDIAM*]

Ilustración 5: Base de datos resultado de las concordancias guardadas

M	Concordancia	Siglo	Año	País	Autor (hor	Autor (dat	Tipo textual	Cómo citar
	v [illegible] a, y como haya /[30] poco amor de Dios, y menos temor y [illegible]	16	1542	MEX	hombre	español	Documentos jurídicos	[Año 1542, Méxic
	v tra / en tres. Mírese, por amor de Dios, q[illegible] s gran / trab[illegible]	16	1543	MEX	hombre	español	Documentos entre particulares:	[Año 1543, Méxic
	v ɔr: / La gr[illegible] a y amor de N[illegible] ro Señor [illegible] suc[illegible] to more e[illegible]	16	1547	MEX	hombre	español	Documentos entre particulares:	[Año 1547, Méxic
	v gó, e le dixo: "hijo, por amor de Dios, no / digas nada al ob[illegible]	16	1547	MEX	hombre	español	Documentos jurídicos	[Año 1547, Méxic
	v mis / hijos y hermanos por amor de dios que me escriva / [illegible]	16	1550	COL	hombre	español	Documentos entre particulares:	[Año 1550, Colom

Ilustración 6: Nombre de los campos de la base de datos automáticamente generada

La ilustración 6 muestra solo una parte del *Excel* anterior para que el lector pueda acceder fácilmente a la lectura de los campos generados automáticamente.

3.3 Visualización del contexto en la ventana lateral

Así como es posible ver el conjunto de la plantilla de metadatos en la ventana lateral, también es posible la visualización del contexto de la concordancia en esa misma ventana. Es casi innecesario explicar cómo un mayor contexto puede ser necesario en muchos casos para la mejor comprensión de una concordancia. En la ilustración 7, se puede ver un ejemplo de una de las concordancias de la búsqueda *amor*.

3.4 Otras propiedades informáticas

En esta sección, nos referiremos a la organización de la información, a las formas de buscar esta, a la posibilidad de realizar búsquedas simples y complejas y a la lematización.

3.4.1 Acceso amigable al documento completo en ventana lateral

Además de ver el documento completo en la ventana lateral, este puede ser bajado e impreso. También en la ilustración 7 es posible observar, debajo del texto que da contexto a la concordancia, la leyenda «Mostrar el documento».

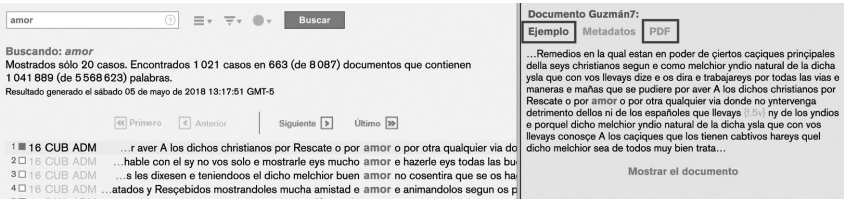


Ilustración 7: Ejemplo de una concordancia desplegada en ventana lateral

Si se pone el cursor sobre ella, se puede acceder al documento completo. Una vez abierto el documento completo, este se puede guardar, así como también imprimir. Otro camino de acceso al documento es a través de la pestaña PDF, en la parte superior de la ventana lateral, como se marca en la ilustración 7 anterior, en el segundo de los recuadros rectangulares.

3.4.2 Cuantificación de la información

El *CORDIAM* provee información cuantitativa instantánea sobre el número de casos encontrados, el número de documentos en el que se encontraron tales casos, el número de documentos en los que se buscó —esto es, la totalidad de documentos que cumplen con las características o restricciones de la búsqueda solicitada—, así como también la totalidad de palabras contenida en los documentos en los que se buscó y la cantidad de palabras contenida en la totalidad de documentos que cumplen con las condiciones de búsqueda. Esta cuantificación permite la comparabilidad de universos, por ejemplo, entre dos cortes diacrónicos. Con los datos ofrecidos por el *CORDIAM*, el usuario sabe explícitamente el tamaño de los universos de palabras o de documentos con los cuales está trabajando.

Si volvemos a la ilustración 7, repetida parcialmente debajo como ilustración 8, podemos ver que la búsqueda de la palabra *amor* sin restricción alguna arrojó como resultado 1021 concordancias, tomadas de entre 663 documentos de los 8087 que integraban el *CORDIAM* en el momento de la búsqueda. Estos 663 documentos contienen 1 041 889 palabras. El conjunto de los documentos del corpus contiene, en el momento de escribir este artículo, 5 millones 568 623 palabras.

3.4.3 Organización de la información

La visualización que proporciona el *CORDIAM* por defecto es el siglo —16, 17, 18, 19—, el país —ARG, BOL, …VEN…— y el tipo textual de las concordancias —ADM,

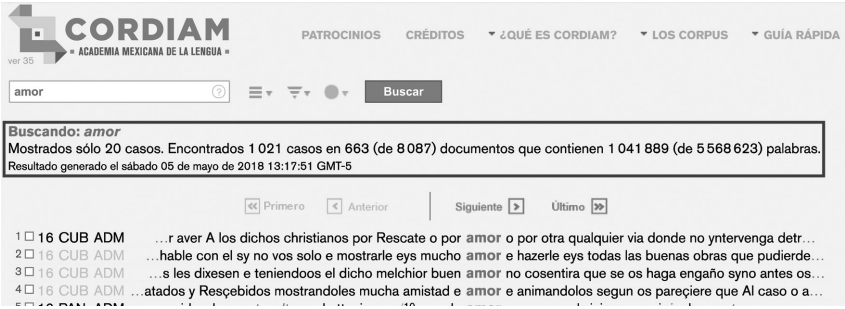


Ilustración 8: Cuantificación del universo de búsqueda

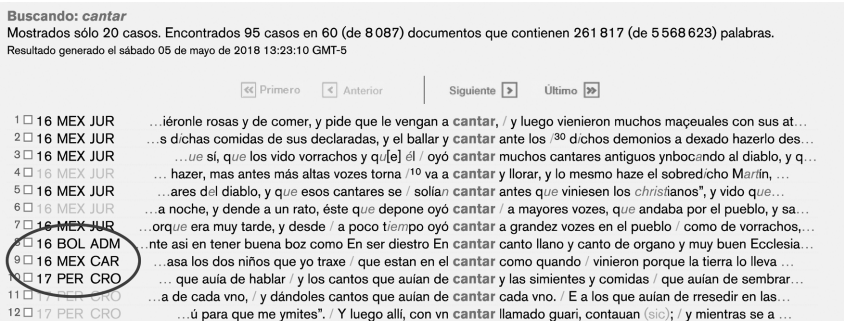


Ilustración 9: Ordenamiento por metadatos siglo, país y tipo textual

CRO, COM—. En la ilustración 9 vemos las concordancias de la búsqueda *cantar*: la octava concordancia es de un documento del siglo XVI escrito en la actual Bolivia y de tipo administrativo; la novena es de un documento también del siglo XVI pero escrito en el actual México y de tipo *cartas entre particulares y otros*, y la décima es un documento escrito en el siglo XVII en el actual Perú y es del tipo cronístico.

Este orden de presentación de concordancias puede ser personalizado, no obstante, de diversos modos: por año, por país, por tipo textual o por orden alfabético. Asimismo, es posible la presentación de resultados en diversas modalidades aleatorias; tal como se puede ver en la ilustración 10 a continuación, el orden puede ser cambiado de acuerdo con las necesidades del usuario.

Ilustración 10: Otras modalidades de ordenamiento

3.4.4 Búsquedas simples y complejas

Además de búsquedas simples, como hemos visto hasta ahora en todos los ejemplos, el *CORDIAM* permite búsquedas complejas, esto es, de más de una forma o más de un lema, tanto continuas como discontinuas, por precedencia y por subsecuencia. Supongamos que queremos buscar construcciones del verbo *ir* más *a* más infinitivo. Como se puede ver en la ilustración 11 abajo, se deberá poner el verbo *ir* en la ventana de búsqueda, ir luego al tercero de los botones de búsqueda al lado de la ventana principal y elegir allí que la búsqueda de *ir* sea como lema. Luego se indicará que se busque este verbo seguido de *a* y de cualquier palabra terminada en *–ar*, *–er* o *–ir*, o sea, un infinitivo. En este caso elegimos pedir, además, al motor de búsqueda que entre el verbo y la construcción siguiente pueda haber hasta dos palabras. Recuérdese que también podría pedirse que la construcción fuera precedente —un sinsentido en esta búsqueda específica— o que fuera tanto precedente como consecuente.

En la ilustración 12, mostramos una pequeña parte de los más de mil resultados de la búsqueda.

3.4.5 Lematización

En su estado actual el *CORDIAM* está lematizado en 70 % aproximadamente y permite búsquedas con consideración o no de diacríticos, como se puede ver en la primera línea del cuadro de diálogo de la ilustración 11.

Las siete propiedades anteriores en un mismo corpus son exclusivas del *CORDIAM*. A estas siete propiedades, se suma el rasgo definitorio de su americanidad

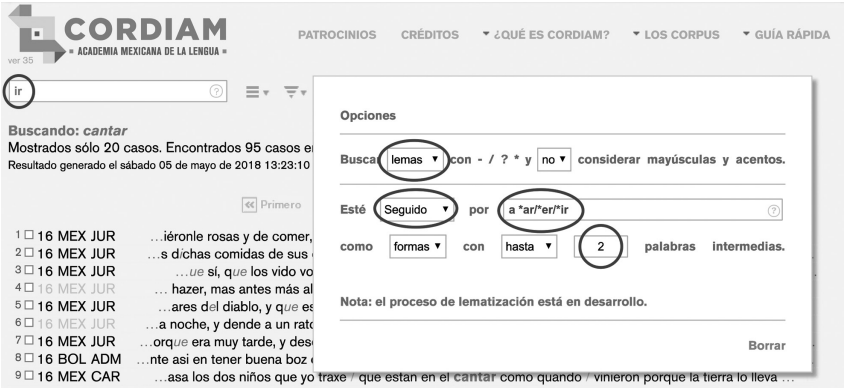


Ilustración 11: Búsqueda lematizada y discontinua

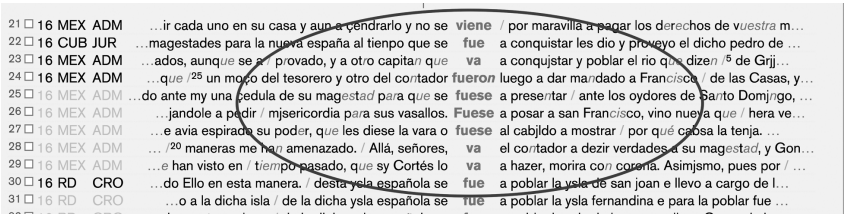


Ilustración 12: Parte de los resultados de la búsqueda del lema *ir* con *a* seguido de cualquier infinitivo con hasta dos palabras de distancia

y el hecho de haber sido concebido conjugando la reflexión sobre formas de investigar con un correlato informático amigable.

4 Proceso de creación de los tres subcorpus

En Bertolotti/Company (2014) se exponen las principales características del *CORDIAM* y las razones para su creación. Cabe señalar que muchas de las cuestiones allí afirmadas en forma de tiempo futuro ya se han concretado y, en la actualidad, el *CORDIAM* se encuentra en una etapa, que, sintéticamente, se caracteriza por su diversificación textual, como ya hemos comentado. Esta diversificación textual debe interpretarse como una superación del «giro archivístico» de las décadas de los 80 y 90 que llevó a centrarse en la búsqueda de textos de la *inmediatez comunicativa* (Oesterreicher 1996) a través de documentos recogidos

directamente de archivo, que permitieran escapar de los moldes de los textos y documentos *cultos*.

En su primera fase, 2012–2015, el *CORDIAM* se centró en la construcción de un corpus de documentos no impresos, el actual subcorpus *CORDIAM-Documentos*. Se nutrió en esa fase del trabajo previo de más de 40 investigadores que han construido corpus nacionales o regionales de documentos escritos en América, tomados directamente de archivo y editados críticamente[3]. Como ya

3 Edward Baranowski (California State University Sacramento, EEUU), Virginia Bertolotti (Universidad de la República, Uruguay), María Elvira Buelna Serrano (Universidad Autónoma Metropolitana, México), Micaela Carrera de la Red (Universidad de Valladolid, España), Belem Clark de Lara (Universidad Nacional Autónoma de México), Magdalena Coll (Universidad de la República, Uruguay), Concepción Company Company (Universidad Nacional Autónoma de México), Manuel Contreras Seitz (Universidad Austral de Chile), Jerry Craddock (University of California Berkeley, EEUU), Ana María Díaz Collazos (College of Wooster, EEUU), Lucinda Díaz de Martínez (Universidad Nacional de Jujuy), María Cristina Egido Fernández (Universidad de León, España), Adolfo Elizaincín (Universidad de la República, Uruguay), José María Enguita (Universidad de Zaragoza, España), Marta Fernández Alcaide (Universidad de Sevilla, España), Ana María Fernández Lávaque (Universidad Nacional de Salta, Argentina), Marta Guzmán (Ludwig-Maximilians Universität München, Alemania), Ofelia Huamanchumo de la Cuba (Ludwig-Maximilians Universität München, Alemania), Marisa Malcuori (Universidad de la República, Uruguay), Bárbara de Marco (University of California Berkeley, EEUU), María del Carmen Martínez Martínez (Universidad de Valladolid, España), Mariela Masih (Universidad Nacional de Córdoba, Argentina), Chantal Melis (Universidad Nacional Autónoma de México), José G. Mendoza (Universidad Mayor de San Andrés, Bolivia), Rosario Navarro Gala (Universidad de Zaragoza, España), Enrique Obediente Sosa (Universidad de los Andes, Venezuela), Fanny Osán de Pérez Sáez (†) (Universidad Nacional de Salta, Argentina), Claudia Parodi (†) (University of California, Los Ángeles, EEUU), Vicente Pérez Sáez (Universidad Nacional de Salta, Argentina), Ana Clara Polakof (Universidad de la República, Uruguay), Ana María Postigo de De Bedia (†) (Universidad Nacional de Jujuy, Argentina), Miguel Ángel Quesada Pacheco (Universidad de Bergen, Noruega), José Luis Ramírez (Universidad de Querétaro, México), Pedro Ramírez Quintana (Universidad Autónoma de Campeche, México), Néstor Fabián Ruiz Vásquez (Instituto Caro y Cuervo, Colombia), Paloma Reyna Vázquez (Universidad Panamericana, México), José Luis Rivarola (†) (Universitá di Padova, Italia), Viridiana Rivera Álvarez (Instituto Politécnico Nacional, México), Agustín Rivero Franyutti (Universidad Autónoma del Estado de Morelos, México), Elena Rojas (Universidad Nacional de Tucumán, Argentina), Juan Justino da Rosa (Academia Nacional de Letras, Uruguay), Israel Sanz-Sánchez (West Chester University, Pennsylvania, EEUU), Luciana de Stefano (†) (Universidad Central de

señalamos, *CORDIAM-Documentos* solo incluye materiales extraídos directamente de archivo, paleografiados directamente por especialistas en lengua y editados con criterios ecdóticos explícitos. Estos documentos incorporados al *CORDIAM* en esa primera fase, siempre activa y en crecimiento, provienen de 58 archivos o repositorios documentales diferentes. En otras palabras *CORDIAM-Documentos* es una compilación de corpus ya existentes, basada en la colaboración internacional. Cada uno de los actuales 4178 documentos de archivo cuenta con sus metadatos y se adscribe a uno de los cuatro tipos textuales ya presentados.

A partir de 2016 los esfuerzos y tareas del *CORDIAM* se centraron en construir otros dos subcorpus: *CORDIAM-Literatura* y *CORDIAM-Prensa*. Este está constituido por unidades textuales tomadas de la prensa periódica americana de los siglos XVIII y XIX ya disponibles en repositorios web, aunque también incluirá el trabajo de colaboradores que ya han investigado y editado prensa. Una vez seleccionado un documento y luego de haber pasado un filtro de exclusiones[4], este pasa por un proceso de conversión digital, cotejo, edición y confección de metadatos. Los criterios para los metadatos y para la edición son compatibles con los de los otros dos subcorpus, como vimos más arriba. Cada uno de los actuales 1865 documentos de prensa cuenta con sus metadatos y, entre ellos, está adscripto a uno de los tres tipos textuales ya presentados.

Por su parte, *CORDIAM-Literatura* está todavía en una fase inicial en cuanto a la disponibilidad de los materiales, aunque la programación de la interfaz, las decisiones, el relevamiento básico y la tipología de textos literarios ya ha sido realizada. Se han seleccionado ediciones críticas ecdóticamente conservadoras así como también obras inéditas. Para la selección se toma en cuenta el canon literario del país, como ya señalamos. Las obras pasan por un procedimiento de adecuación informática y edición. Cada documento lleva una plantilla de metadatos, ya descrita, y se adscribe a uno de los cinco tipos textuales presentados antes.

La estructura del *CORDIAM* en tres subcorpus, con tipologías internas diferenciadas pero apoyadas en criterios comunes y en la evidencia empírica de regularidades textuales y lingüísticas, supone una ventaja para el usuario, que puede

---

Venezuela), Ana Isabel Tsutsumi (Universidad Nacional Autónoma de México), Josefina Tejera (Universidad Central de Venezuela), Nerea Zabalegui (Universidad Central de Venezuela).

4 Recordemos que excluimos los textos escritos fuera de América, como las corresponsalías de extranjeros o reproducciones de textos escritos fuera de América, los textos de humor y juegos de palabras y los textos literarios (que serán incluidos en *CORDIAM-Literatura*).

encontrar la información transparentemente compartimentada en subcorpus o tenerla holísticamente. La decisión, como corresponde, será del investigador.

En el momento de escribir este trabajo, los documentos disponibles en www.cordiam.org es de 4 498 773 palabras (octubre de 2017) en 6147 textos: 4178 documentos, 104 textos de literatura y 1865 de prensa.

5 Consideraciones finales y tareas futuras

Hemos mostrado en este trabajo los criterios y el proceso de construcción de un corpus electrónico, de naturaleza diacrónica, cuya finalidad esencial es conocer las trayectorias de la lengua española en América, así como las culturas en ella sustentadas, además de contribuir a un mejor conocimiento del español general, en tanto que más del 90 % de hispanohablantes nativos es nacido en el continente americano. Este corpus, el *CORDIAM*, se erige como el corpus de referencia para el español de América.

Hemos descrito tanto las propiedades diatópicas, diacrónicas, autorales y textuales, cuanto las propiedades informáticas, así como una breve historia de la construcción del *CORDIAM* y sus etapas.

Es mucho, aún, lo que queda por realizar, además de seguir creciendo cuantitativamente en los tres subcorpus, y cualitativamente cuando nuevos datos y necesidades de investigación soliciten ajustes y mejoras de programación. Dos son las tareas futuras centrales: en un futuro próximo será necesaria la subida de facsimilares en *CORDIAM-Documentos* y en *CORDIAM-Prensa*, y elaborar un programa *ad hoc* que acople la búsqueda con su ocurrencia en el facsimilar. También es muy deseable una lematización total.

El *CORDIAM* sienta las bases para una tarea general urgente: crear un Corpus de Español Atlántico, al que ojalá se unan pronto, mediante los respectivos corpus electrónicos históricos, el español de Canarias y el del sur y suroeste de España. El camino ya se ha iniciado y está muy avanzado.

Referencias bibliográficas

Bertolotti, Virginia/Concepción Company Company (2014): «El *Corpus Diacrónico y Diatópico del Español de América* (*CORDIAM*). Una propuesta de tipología textual», *Cuadernos de la ALFAL* 6 (número especial: Claudia Parodi y Micaela Carrera de la Red (eds.), *El español en América. Corpus y textos*), 130–148. Disponible en <http://www.mundoalfal.org/sites/default/files/revista/06_cuaderno_011.pdf> [último acceso: 27/10/2017].

EAGLES (1996): *Preliminary recommendations on Corpus Typology*. Disponible en <http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html> [último acceso: 27/10/2017].

Enrique-Arias, Andrés (2012): «Dos problemas en el uso de corpus diacrónicos del español: perspectiva y comparabilidad», *Scriptum Digital* 1, 85–106.

Kabatek, Johannes (2016): «Un nuevo capítulo en la lingüística histórica iberorrománica: el trabajo crítico con los corpus», en Johannes Kabatek (ed.) con la colaboración de Carlota de Benito Moreno, *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín/Boston: De Gruyter, 1–17.

Maturana, Humberto A. (1996): *La realidad ¿objetiva o construida?*, vol. 2: *Fundamentos biológicos del conocimiento*. México: Universidad Iberoamericana-Anthropos-Iteso.

Oesterreicher, Wulf (1996): «Lo hablado en lo escrito. Reflexiones metodológicas y aproximación a una tipología», en Thomas Kotschi, Wulf Oesterreicher y Klaus Zimmermann (eds.), *El español hablado y la cultura oral en España e Hispanoamérica*. Madrid/Fráncfort: Iberoamericana/ Vervuert, 317–340.

Esther Hernández

Tesoro léxico de los americanismos contenidos en los vocabularios hispano-amerindios coloniales (1550–1800) [*TELEAM*][1]

Resumen: El proyecto *TELEAM* consiste en la elaboración de un tesoro con los *americanismos* contenidos en los vocabularios bilingües del español con las lenguas indígenas, que fueron producidos por los misioneros durante el periodo colonial y que abarcan gran parte de América. Sus lemas comprenden una selección de las palabras originadas en América que se hallan en el texto castellano de estos vocabularios, y van acompañados de su significado, el contexto en que aparecen, la descripción de los procedimientos de adaptación fonética, morfológica y semántica con que se integra el neologismo, otras documentaciones coetáneas, una explicación de su origen y una síntesis de la historia de su trayectoria.

Palabras clave: Léxico histórico, Español de América, Lexicografía, Variación y cambio lingüístico

Abstract: The project encompasses the elaboration of a thesaurus of the Americanisms found in the bilingual vocabularies of Spanish and indigenous languages. These vocabularies were produced by missionaries during the Colonial period, covering the majority of the American territories. The entries will comprise a selection of words of American origin found in the Castillian text of the vocabularies, in alphabetical order, and will be accompanied by their meaning, the context in which they appear, the description of the processes of phonetic, morphological and semantic adaptation incorporated in the neologism, other contemporary documentation, an explanation of their origin and a synthesis of the history of their trajectory.

Keywords: Historical Lexicology, American Spanish, Lexicography, Language variation and change

1 Introducción

El proyecto de investigación *TELEAM* se enfoca en el análisis y el estudio histórico de los americanismos léxicos que aparecen en el texto castellano de los

1 Proyecto FFI2016-78810-P: «Tesoro léxico de americanismos en los vocabularios hispano-amerindios coloniales (1550–1850) [*TELEAM*]».

vocabularios bilingües del español con las lenguas indígenas americanas, que fueron producidos por misioneros entre 1550 y 1800[2]. Todos los diccionarios, en general, permiten el acceso a una intrincada red de información sobre el lenguaje y sobre las lenguas (Durkin 2016: 3). En el caso de los vocabularios hispano-amerindios coloniales, además de ser fuentes útiles para conocer las lenguas indígenas, lo son para el estudio del español americano, ya que contienen información valiosa sobre la historia de las palabras de América o *americanismos*, de acuerdo con las acepciones 5 y 6 del *DLE*.

Debido al enfoque histórico de esta investigación, el concepto de *americanismo* se restringe a la palabra o expresión originada en América, bien en forma de acuñaciones absolutas, como son los indigenismos, bien como nuevas palabras creadas mediante los recursos formativos del español, o las nuevas acepciones o innovaciones semánticas sobre una base léxica hispánica. Es decir, se ocupa de *americanismos de origen*, no de *americanismos de uso*. Por tanto, escapan al interés las palabras que, en la actualidad, o en algún momento de la historia, son o han sido de uso general o de algún lugar o lugares de América. De este modo, son solo objeto de análisis las palabras o expresiones nativas allí creadas; es decir, las palabras que pueden fechar su nacimiento en el Nuevo Mundo, con mayor o menor precisión o certeza, y que son distintivamente características o peculiares de América. En este sentido, se procura hacer un diccionario o tesoro diferencial, con una dimensión contrastiva y dentro de un periodo cronológico bien establecido.

TELEAM abarca desde el inicio de la lexicografía bilingüe del español con las lenguas amerindias, que tuvo lugar a mediados del siglo XVI, hasta la etapa de las Independencias. El fin de esta tradición lexicográfica cabe definirla con claridad a mediados del siglo XIX, cuando surge un tipo de diccionario diferente con las lenguas indígenas americanas, que fue promovido fundamentalmente

2 En investigaciones previas, se ha acometido el estudio historiográfico de estos vocabularios y sus resultados se han plasmado en la publicación de un catálogo que contiene información bibliográfica y un análisis crítico de muchos de ellos (Hernández 2018). Este catálogo descriptivo agrupa de modo ordenado y completo el conjunto de vocabularios que fueron compilados por los misioneros de las distintas órdenes religiosas durante ese periodo. En él, se da cuenta de la localización actual de las obras, si bien muchas están perdidas o en paradero desconocido. En general, estos diccionarios comparten propósito y técnica de elaboración, y se constituyen como una tradición dentro de la historia de la Lexicografía bilingüe hispánica. La lista comprende más de ciento cincuenta vocabularios, de los cuales, la mayoría son hispano-amerindios; pero, también, los hay con las palabras-entrada en lengua indígena.

por la *Smithsonian Institution*, creada en 1846, al tiempo que emerge la ciencia de la Lingüística. Por esta razón, pero también por la distinta cronología en los procesos de independencia de los países americanos, se ha fijado el límite de la investigación al concluir el siglo XVIII, aunque existan algunos vocabularios decimonónicos de corte semejante a los coloniales, que son, por otro lado, escasos y poco significativos para nuestros propósitos léxicos.

Este trabajo se enmarca en una línea de investigación sobre la historia del léxico del español en América, cuyos hitos bibliográficos fundamentales resumo brevemente a continuación, que arranca de los inicios de la dialectología hispanoamericana. En efecto, la *americanización* del léxico fue un tema al que se prestó atención especial por parte de los primeros especialistas: Cuervo (1901a, 1901b), Lenz (1904–1910), Henríquez Ureña (1935, 1938, 1944), Morínigo (1964), Rosenblat (1973). Sin duda, el diccionario histórico de americanismos de Friederici (1960), que combina sus *Americanisches Wörterbuch* (1947) e *Hilfswörterbuch für den Amerikanisten* (1926), es la obra lexicográfica pionera, que aún hoy es de referencia imprescindible. Asimismo, es fundamental el corpus del *Léxico hispanoamericano* de Boyd-Bowman (*cfr.* Glessgen 1997), que afortunadamente está en internet desde 2015, así como los estudios léxicos de él derivados: Grace (1976), Mejías (1980) y Zamora Munné (1976, 1982). Importantes son también los trabajos de Lerner (1974), Lope Blanch (1969), López Morales (1974) y Lüdtke (1994, 2014), entre otros. Pero, de manera especial, destacan las investigaciones de Manuel Alvar, quien realizó diversos estudios sobre americanismos (reunidos en Alvar 1975); por otro lado, los indigenismos analizados en varias de las tesis doctorales por él dirigidas fueron dispuestos en forma de diccionario (Alvar Ezquerra 1997). Hay otros antecedentes importantes, como el trabajo de Buesa (1965), después revisado en Buesa/Enguita (1992), Alvar Ezquerra (1987), o el trabajo de colaboración de Sala/Munteanu/Neagu/Sandru-Oleanu (1977), los estudios de Rivarola (1985, 1990), Franco Figueroa (1992), Martinell Grife (1993), así como los estudios de conjunto de Enguita Utrilla (2004) y Torres Torres (2004). De interés resultan las investigaciones que manejan los diferentes géneros textuales, las llamadas tradiciones discursivas, y, en especial, el concepto de hablante «semiculto» (Oesterreicher/Stoll/Wesch 1998). Para nuestros propósitos, son referente filológico ineludible los trabajos sobre el español americano de Frago Gracia (1992, 1998–1999, 1999, 2010), en particular, sus estudios específicos sobre algunos elementos léxicos (2003). Por último y más importante, y no solo como un antecedente, sino como un referente específico fundamental por su contenido y por su planta, el proyecto *TELEAM* tiene en cuenta el *Tesoro léxico canario-americano* de Corrales/Corbella (*TLCA*

2010), repertorio que reúne información documental y contrastiva de las hablas de Canarias y América.

De un tiempo a esta parte, la investigación sobre el léxico histórico del español ha avanzado exponencialmente gracias a los avances digitales. Rojo señala que «la lingüística de corpus ha cambiado de manera radical la forma de acometer el estudio de la evolución del español» (2012: 433). Son muchos los corpus que hoy permiten rescatar, contextualizar y analizar hechos léxicos del español en su diacronía; pero además, se han ampliado las tipologías textuales que sirven de material de base para los grandes proyectos léxicos del español. Sin embargo, los corpus históricos, *CORDE*, *CORPES XXI*, *CORDIAM*, o el diccionario histórico *NDHLE*, no incluyen los vocabularios bilingües coloniales como fuentes para su confección[3].

Entendemos que el valor de estos productos lexicográficos, en comparación con textos americanos de otros géneros, lo determinan los americanismos que se registran como entrada; en efecto, constituye un tipo de texto en el que cobra un interés especial el uso de palabras o de estructuras morfosintácticas divergentes del español peninsular, tanto las empleadas en las entradas de los diccionarios, si la lengua de partida es la española, como las que aparecen incrustadas en las definiciones, si se trata de la lengua de llegada. Hemos constatado en trabajos realizados con anterioridad que los vocabularios coloniales tienen un valor añadido respecto de otros géneros textuales, pues proporcionan pruebas novedosas del origen, difusión y vitalidad de las palabras de América en su contexto; en ellos encontramos información etimológica relevante y registros tempranos fiables de los préstamos que el español tomó de las lenguas aborígenes, los indigenismos, así como nuevas documentaciones de los otros neologismos de base léxica patrimonial o las nuevas acepciones creadas en América. A continuación, doy cuenta de las razones específicas de su valor e interés.

En el caso de los indigenismos, la relevancia de estas fuentes es obvia porque, en muchas ocasiones, los vocabularios bilingües presentan el neologismo

3 Cuando empezamos a trabajar con estos materiales, el plan inicial consistía precisamente en hacer el inventario de los vocabularios bilingües coloniales, en gran medida ignorados, y, con la pretensión de analizar históricamente el léxico que contenían, a continuación, fundir alfabéticamente todos los vocablos del español que en ellos se empleaban, como se había realizado con el vocabulario de *Molina 1571* (Hernández 1996) y teniendo presente el *Léxico hispanoamericano* de Boyd-Bowman (2015). Sin embargo, se trataba de un trabajo excesivamente arduo y, en cierto modo, poco productivo, dado que bastantes diccionarios repiten las entradas de los vocabularios que les preceden en el tiempo.

en relación directa con su étimo. Además, permiten documentar variantes ortofonéticas y conocer los procesos de acomodación del étimo al español. En el vocabulario de *Febres 1765* de la lengua chilena, encontramos las variantes fonéticas del étimo: «*Ulpu, ulpud, ó ullpud*. bebida simple de harina tostada con agua: *ulpudu, ulpudtun* tomar este *ulpo*». En este caso, vemos cómo penetra el mapuchismo *ulpo* en el español, con la variante *-o*, que es vocal más frecuente que la *-u* etimológica en el final de palabra; tal documentación sirve, además, para determinar su étimo, que no aparece en el *DLE*. Caben, además, precisiones de otra índole al étimo, como en el caso del utensilio de cocina mexicano, el *molcajete*, cuyo étimo sería preferible transcribirlo con la variante *mulcaxitl* y no *mulcazitl* (además de por motivos de acomodación fonética, por razones documentales, *cfr. Molina 1571*: «*Mulcaxitl*. escudilla», fol. 61v, y *Sahagún 1585*: «y los platos que se llaman *molcaxitl*», «unas escudillas que se llaman *molcaxitl*», *Códice florentino,* lib. 6, fol. 109; lib. 8, fol. 25). Estos registros pueden ser útiles también para proponer otro tipo de enmiendas al diccionario académico: por ejemplo, la de eliminar la doble entrada en el americanismo *biznaga* 2, que se emplea en México para designar varios cactos espinosos (*Echinocactus*) y al que el *DLE* asigna un origen nahua (Hernández 2011b: 122, n. 16). Sobre un americanismo, *camiseta*, que en su momento calificamos de inadvertido (Hernández 2011a), hemos podido observar que la edición vigente del *DLE* ha modificado el orden de sus acepciones respecto de anteriores ediciones, quedando más acorde con el uso actual, que con la tradición lexicográfica anterior que lo motivaba. Estas aportaciones que se acaban de mencionar pueden ser consideradas como antecedentes inmediatos del proyecto *TELEAM*. De modo especial, algunos estudios previos (Hernández 1996, 1998, 2000; Carriazo Ruiz 2014) constituyen puntos de partida para tratar de completar el origen o la etimología de las palabras originadas en América.

Una atención especial merecen los americanismos que se documentan en estos diccionarios bilingües por vez primera. Debido a su carácter novedoso, suelen aparecer en contextos explicativos con su definición incluida, lo que ayuda a conocer el significado del concepto al que aluden y, en muchas ocasiones, a comprender el procedimiento para su creación. Por ejemplo, es esclarecedor el contexto en que se inserta el americanismo *vainilla* en una entrada del vocabulario español-tzeldtal de *Ara c1571*, a modo de «definición dentro de la definición», según la expresión de Seco (2003: 35). En esta obra encontramos no solo un testimonio que permite adelantar su datación, sino también la explicación de la

palabra de la que deriva (*vaina*): «*bainillas* especia aromatica son largas y negras ha hechura de vaina. *tzitzib ha*».

Aunque hoy conocemos mejor la historia del español americano, sigue habiendo lagunas en el discernimiento de los procesos de cambio léxico que tienen lugar en América, particularmente en lo que se refiere a cómo y por qué se formaron las nuevas palabras, y sobre cuál fue su origen o su motivación; pero también, acerca de los procesos de variación fonética, morfológica y semántica que precedieron al cambio léxico-semántico. Pues bien, estos vocabularios son unas fuentes óptimas para estudiar cómo empiezan a penetrar los americanismos en el español: facilitan su datación en el tiempo y permiten localizarlos mejor en el espacio. Por tanto, el análisis de las documentaciones halladas en los vocabularios, y el estudio histórico de sus distintas formas y acepciones permite contribuir a remediar inexactitudes en las definiciones de los americanismos en las obras lexicográficas de referencia, así como a favorecer la comprensión de su difusión geográfica y de su productividad semántica, en cuanto a formación de derivados y frases hechas, entre otras cuestiones[4].

2 Las fuentes y los objetivos del proyecto *TELEAM*

Nuestra hipótesis es que se puede avanzar en el conocimiento de la historia del español americano empleando los materiales léxicos que contienen los vocabularios bilingües, producidos por los misioneros durante la época colonial y que proceden de diversas zonas de América. Subsidiariamente, ese conocimiento puede propiciar un mejor análisis de los procesos de variación fonética, morfológica y semántica consustanciales al contacto entre el español y las lenguas y culturas autóctonas americanas.

Así las cosas, el objetivo primordial de *TELEAM* consiste en elaborar y publicar un tesoro, un tesoro léxico en el sentido tradicional de la lexicografía española, el que le dio Gili Gaya (1947), esto es, un «diccionario de los diccionarios bilingües». Al mismo tiempo, nos proponemos realizar trabajos de síntesis sobre aspectos generales de la variación y el cambio lingüístico, en el marco de la dialectología y la lingüística histórica.

4 Vid. los comentarios al *Diccionario de americanismos* en Cerrón-Palomino (2010) y Frago Gracia (2013); o, sobre otras obras sincrónicas de la lexicografía hispánica, las críticas de Lara (2004).

2.1 El interés de los vocabularios como nuevas fuentes para el conocimiento del léxico histórico

Hasta ahora, las fuentes para el análisis diacrónico de los americanismos han sido las crónicas y otros tipos documentales. El corpus de los vocabularios (y de las gramáticas) compuestos por los misioneros en la época colonial ha sido un fondo poco estudiado por la filología y la lingüística españolas. Es cierto que, en los estudios hispánicos, Menéndez Pidal consideró los vocabularios como materiales para la investigación de la dialectología en América (1944: xiv). Por su parte, Amado Alonso afirmaba: «vamos a reeditar gramáticas y vocabularios antiguos (época de conquista) de las lenguas indias. *Mucho nos dirán sobre el español importado*» (*apud* Guitarte 1996: 82 [el subrayado es nuestro]). Sin embargo, no existe, que sepamos, ningún proyecto que actualmente esté usando o haya usado los vocabularios misioneros para dar documentaciones o trazar etimologías. A continuación, se insiste en una exposición de los motivos del interés de los vocabularios coloniales.

Se podría aducir que una caracterización diacrónica del léxico del español americano debe contener datos lexicológicos, no lexicográficos. Sin embargo, los vocabularios bilingües coloniales no son diccionarios léxicos, sino que poseen un objetivo concreto, basado en las necesidades de un usuario modelo; es decir, tienen «finalidad genuina» (Jacinto García 2016: 148). De hecho, constituían herramientas para que los misioneros aprendieran las lenguas indígenas y, por ello, presentaban materiales léxicos correspondientes al uso que conocían y practicaban; podríamos decir, *mutatis mutandi*, que no estaban «contaminados por el normativismo inherente a muchas obras lexicográficas» (Lara 2004: 305). Por ejemplo, *Febres 1765* señalaba en su *Arte y vocabulario*: «Todas las reglas de este *Arte* son ciertas, seguras y conforme a lo que al presente se usa; y no pondré otra cosa que lo que haya oído y usado o no sepa de cierto» (*Prólogo al estudioso*).

De acuerdo con este presupuesto, el léxico que contienen es representativo del impacto que sobre el español tienen las lenguas y culturas de América, y revelador también de hechos diatópicos o dialectales, producidos por los distintos procesos de aclimatación del español al Nuevo Mundo. En este sentido, el valor que presentan los contenidos léxicos de los vocabularios se explica por las siguientes razones: (1) proporcionan información lingüística correspondiente a un nivel diastrático que podríamos considerar alto, ya que los misioneros son normalmente hombres cultos y se relacionan con las élites indígenas; (2) tienen la ventaja de que están fechados, o su datación se puede conjeturar con ciertas garantías, lo que permite datar con fiabilidad los cambios; (3) están localizados

geográficamente; y (4) los manuscritos, o sus copias, pueden aportar información valiosa desde el plano fonográfico.

Respecto de otros géneros textuales, los vocabularios coloniales tienen la ventaja de que los neologismos americanos aparecen contextualizados, y no solo los que se encuentran en el texto corrido de las equivalencias o definiciones, sino también en el caso de los americanismos contenidos en las entradas de estos diccionarios, que suelen ser pluriverbales (como se puede apreciar en el ejemplo de *vainilla* del vocabulario de *Ara c1571*, v. *supra*). De este modo, el contorno del diccionario suministra datos lingüísticos e interlingüísticos adicionales de las palabras y expresiones nuevas. En este sentido, cobra especial relieve la contigüidad con que a menudo se documentan los préstamos con respecto de su étimo. Por ejemplo, en el diccionario de la lengua chilena de *Febres 1765* hallamos esta praxis de relacionar el mapuchismo con su étimo en el artículo siguiente: «*chuchoca*, o *cunarquen*. la *chuchoca*, esto es, maiz tostado, o cozido, para secar y guardar»[5]. Según se puede observar, el indigenismo está morfológicamente adaptado al español mediante la anteposición del artículo *la*.

Pero además, es un corpus que puede permitir conocer el recorrido de un vocablo normalmente desde su creación, dando cuenta de la vida de la palabra, ya sea efímera o prolongada, extensa o reducida. Decía Rona (1969: 146) que «la palabra *chocolate* no se originó en América, sino en cierta parte de América». Y no le faltaba razón, porque se tiende a concebir América en bloque, como si fuera un solo dialecto. La etimología de *chocolate* puede ser objeto de discusión por parte de los especialistas, que debaten su origen maya o náhuatl (Dakin/Wichmann 2000); pero hoy sabemos que el préstamo se empieza a emplear en la Nueva España, concretamente en lo que hoy es Guatemala y en el México central, es decir, «no se originó en América». En nuestro corpus, hallamos las primeras documentaciones en el diccionario del cakchiquel de *Vico c1555*:

> salirse el fuego, o calor como la xícara que se sale el *chocolate* (fol. 155v).
> mesclar, o batir *chocolate* (fol. 164r).
> es la sangre y el ule que los indios lo derriten con ocote ensendido, y lo hechan enel *chocolate* que las calienta y dispone para haserse preñadas (fol. 168).
> pedaçitos de pan o *chocolate* (fol. 266v).

Pues bien, a través de estos vocabularios podemos conocer el itinerario que siguen las palabras que se originan en un lugar determinado de América, las

5 El *DLE* no presenta marca etimológica: «*chuchoca*: 1. f. Arg. y Chile. Especie de frangollo o maíz cocido y seco, que se usa como condimento. 2. f. coloq. Chile. lío (‖ barullo, desorden)».

cuales, en gran número, suelen continuar en uso en su territorio de origen. Si se han convertido en panamericanas, panhispánicas o internacionales como *chocolate* interesa, como es lógico, pero no es lo esencial. Lo que importa en *TELEAM* es el proceso de su creación léxico-semántica, fundamentalmente.

2.2 Tesoro léxico del español americano

El objetivo principal es reunir los neologismos americanos contenidos en los vocabularios misioneros en forma de repertorio. Conviene insistir en que no se trata de un diccionario sincrónico, ni interesa si la palabra se emplea en la actualidad o no. La elaboración de este tesoro o repertorio de *americanismos* se hace a partir de datos elaborados, históricos y etimológicos, una vez realizado el expurgo léxico en los vocabularios (previa selección de los mismos, v. 3.2 *Nómina de vocabulario*s).

Es importante señalar que el análisis lingüístico de cada lema incluye un estudio de los procesos de asimilación (fonética y morfológica), si se trata de un préstamo de las lenguas indígenas; o de los recursos lingüísticos (morfológicos o semánticos) que intervienen en los cambios formales y de significado, si se trata de una palabra del fondo léxico patrimonial.

El diccionario con las palabras propias de América, lematizadas y ordenadas alfabéticamente, ha de contener los siguientes campos de información (véanse los criterios lexicográficos que se siguen para la elaboración del tesoro en 3.4):

- → las variantes ortográficas, fonéticas o morfológicas;
- → el significado y sus acepciones;
- → el origen o etimología;
- → el análisis documental, con las citas de la palabra en los vocabularios u otros textos misioneros adicionales junto a su referencia abreviada, así como otras citas, preferentemente coetáneas, para comprobar su vitalidad en la época;
- → el análisis lingüístico, con la información de la categoría gramatical, las variantes morfológicas, derivados, compuestos o frases, los indigenismos, peninsularismos o canarismos equivalentes o sinónimos, y, por último, información sobre el campo léxico;
- → un breve comentario sintético final;
- → la bibliografía de las fuentes secundarias.

3 Metodología

La metodología sigue los pasos se exponen a continuación.

3.1 Elaboración de un corpus bibliográfico de fuentes secundarias

Para las comprobaciones que conducen a confirmar la condición americana de una palabra, un significado o una expresión, se maneja un corpus bibliográfico que reúne los estudios de sesgo histórico en torno a los americanismos, con particular atención a los problemas etimológicos. Incluye bibliografía en torno a la lexicografía y la lingüística de corpus, sin olvidar los estudios de dialectología y de historia de la lengua. Naturalmente, se emplean bancos de datos ya existentes, como el *Portal del léxico hispánico*, y otras bibliografías y estudios relacionados con la lexicología y la lexicografía del español. En particular, tiene en cuenta los tesoros lexicográficos de ámbito regional[6] y los distintos repertorios dialectales.

Este corpus, que incorpora los avances de las descripciones y las investigaciones ya hechas sobre el léxico del español americano, contiene una revisión crítica de la bibliografía y está, lógicamente, en construcción permanente.

3.2 Nómina de vocabularios: selección de los más originales. Estudio de un caso: *Diccionario de Motul* c1600

La lista de los vocabularios comprende más de cien diccionarios de una gran variedad de lenguas, que abarcan una vasta extensión geográfica. Se tiene experiencia práctica de los contenidos léxicos de todos ellos por la descripción efectuada para su catalogación; hay que exceptuar, no obstante, los diccionarios catalogados porque los mencionan las fuentes antiguas, pero que están perdidos o en paradero desconocido (Hernández 2018). Como se ha referido antes, durante el curso de la investigación, se decidió buscar en ellos el léxico diferencial, específicamente, las palabras de América.

La mayoría de los diccionarios han sido leídos cuidadosamente y los que tienen como punto de partida la lengua española tienen marcadas las peculiaridades léxicas que los caracterizan.

De algunos diccionarios, ya se ha hecho el expurgo de los indigenismos y su estudio correspondiente (p. ej. del vocabulario náhuatl de *Molina 1571*, del cachiquel de *Vico c1550* y del mapuche de *Febres 1765*). Por lo tanto, se tiene una idea bastante aproximada de la riqueza léxica de gran parte de los vocabularios, aunque no se hayan observado de modo exhaustivo, ni, en consecuencia, analizado con detalle sus particularidades léxicas.

La originalidad de un repertorio la marca la distancia que tenga respecto de las fuentes lexicográficas precedentes. También, como es natural, cuanto mayor es el número de artículos, cabe esperar una mayor presencia de elementos léxicamente

6 Vid. el capítulo de Corbella en este volumen.

significativos para nuestros objetivos. Si bien este corpus es homogéneo en cuanto al tipo de autor, finalidad, concepción y propósitos de los vocabularios, es lógico que sean desiguales en cuanto a su número de entradas o extensión.

Por ejemplo, uno de los repertorios extensos, que contiene materia léxica relevante, es el *Diccionario de Motul*, atribuido a fray Antonio de Ciudad Real (1551–1617). De él hay una copia manuscrita en la John Carter Brown Library, con la signatura *Codex indicus 8*. Este vocabulario tiene las entradas en la lengua maya yucateca y las definiciones en castellano. Es el más antiguo que se conserva del maya yucateco clásico y se considera una de las fuentes más valiosas para los estudios mayas[7].

Una lectura detallada del texto castellano del *Diccionario de Motul* revela la presencia de numerosas bases léxicas que corresponden a palabras propias de América. En algunos casos, incluso, se trata de la primera aparición encontrada hasta la fecha. Por ejemplo, recogemos una primera documentación de la palabra *mercachifle* y, con ella, podemos proponer su condición de *americanismo*, a juzgar por las documentaciones disponibles y por su difusión. En efecto, de acuerdo con las distintas fuentes históricas, no existe ninguna documentación anterior, ni encontramos arraigo documental de esta voz en el español peninsular hasta el siglo XIX.

El *Diccionario de Motul* presenta dos testimonios: «*ah con xohom*. bohonero, *mercachifles*», fol. 11v, y «*ah chocom conol*. bohonero y *mercachifles*», fol. 15r. La palabra habría surgido como producto de un proceso espontáneo, con los recursos morfológicos propios de la lengua. Compuesta de una base verbal y un nombre en plural (< *merca* 'vende' + *chifle(s)* 'silbato'), sigue un procedimiento habitual de creación de las hablas populares, en el que actúa el mecanismo de la expresividad (*cfr*. *vendepeines*). Aparece definida mediante sinonimia con la palabra *buhonero*, que se empleaba en el español medieval con la acepción de 'mercader ambulante'. Este procedimiento de yuxtaponer dos palabras sinónimas para aclarar el significado de los neologismos es muy frecuente en el español de la época[8]. Hay otros americanismos, como *gallo de papada* 'guajolote', o *higuerilla del infierno* para la planta que el Dioscórides llamó 'ricino'. Y también

7 Ha sido editado por Martínez Hernández (1929), Acuña (1984, 2001) y Arzápalo Marín (1995); está digitalizado en la web de la biblioteca JCB y, además, se puede consultar en Open Library.

8 El *DLE* define *mercachifle* sin proporcionar marca diatópica o pragmática: «De *mercar* y *chifle*. 1. m. y f. despect. Mercader de poca importancia. 2. m. buhonero (‖ hombre que vende buhonerías)». La historia lexicográfica nos muestra que esta palabra está desde *Covarrubias 1611*, sin muchos cambios en su definición. En el *CDH* se encuentra la primera documentación en Perú. A partir del corpus del *LHA* y del *DECH*, podemos conocer la historia de sus documentaciones.

encontramos una documentación temprana del americanismo *armadillo*, acompañado de una explicación de la palabra de la que deriva: «*ibach. armadillo*, animalejo armado de ciertas conchas» (fol. 221v)[9].

El vocablo *vainilla* (v. *supra*) presenta en este repertorio variación semántica, indicativa de que el proceso de cambio no ha culminado. Según se aprecia en los siguientes contextos, solo en la segunda cita podría haberse producido ya la lexicalización y, consecuentemente, designar ya la *Vanilla planifolia*:

> *booxel*. caxcara de calabaças xicaras, platanos, y *vainilla* de legumbres corteza de arbol yerua o matas que no se despide del tronco ni se haze della correas o cuerdas (fol. 56v).
> cijzbic. *bainillas*, olorosas (fol. 102r).
> *ppal. ah. ab*. desgranar o descascarar y mondar frisoles y cosas que tienen baynillas abriendolas & ppal. cosa desgranada (fol. 386v).

En fin, se puede observar que tanto en el vocablo *vainilla*, como en *armadillo* opera la metáfora como mecanismo del cambio semántico para la creación de las nuevas palabras (y por un proceso morfológico de derivación), ambas neologismos necesarios para denominar las nuevas especies de la flora y de la fauna americanas.

Por otro lado, como cabe esperar, hay en el texto del *Diccionario de Motul* muchos nahuatlismos y antillanismos, bajo las distintas formas fonéticas y morfológicas: entre otros, *aguacate*, *chocolate*, *cocoyoles*, *paco*, *chicoçapote*, *ayate*, *cuzcas*, *nances*, *naguatear*, *chilmol*, *piciete*, *zonote*, *poçol*, *tuchumite*, *gualipil*, *guaya*, *capulí*, *anoma*, *pozol*, *pochote*, *pitahaya*, *tupil*. Muy interesantes son los mayismos que figuran en el texto castellano: *cenote*, *macal*, *kaanes*, *balines*, *tancabal*, *chaya*, *choyo* y *xul*. Es discutida la procedencia maya o antillana de *pati*[10], y de *henequén* (Hernández 1999). Y hallamos la primera documentación del mayismo *macal* 'tubérculo semejante a la patata', en un contexto próximo a su étimo, y cuya acepción se encuentra ya en el mismo[11], como se puede observar en los siguientes artículos del diccionario:

---

9 La primera documentación la recoge Friederici (1960), con la variante morfológica *armado*: «animales que se dicen *armados*», de 1541; la siguiente ya presenta la variante con el sufijo *-illo* que triunfó. Corresponde a Acosta y es de 1590: «Lo que defiende a las dantas la fuerza del cuero, defiende a los que llaman *armadillos* la multitud de conchas que abren y cierran como quieren a modo de corazas» (*CDH*).

10 Agradezco al mayista Mario Humberto Ruz hacerme ver que existe esta discusión.

11 *DLE*: «macal[2]. Voz maya. 1. m. Méx. malanga (‖ planta). 2. m. Méx. malanga (‖ tubérculo)». El *DA* no lo recoge, pero sí lo hallamos en el *Diccionario de mexicanismos* de la Academia mexicana: «macal. m. Cierta planta tropical (*Xanthosoma sagitifolium*), y su tubérculo comestible». En el corpus del *NDHLE*, hay dos apariciones del siglo XX de una misma obra.

Hacaacnac. cosa que se va resbalando y deslizando como culebra o anguilla ettz. & *Hacaacac haa tu le macal* deslizando se va el agua por las ojas del *macal* (fol. 170r).
Numay.[…] *Numay haa tu le macal*. passa de presto y no se detiene el agua en las hojas del *macal* (fol. 335v).

Así pues, este vocabulario es uno de los seleccionados para su inclusión en el tesoro, tanto por sus dimensiones y su riqueza léxica, como por lo temprano de su elaboración. Según se acaba de intentar mostrar con algunos ejemplos, el criterio selectivo se justifica por su interés léxico.

La nómina o mapa de diccionarios y otros textos misioneros será, inicialmente, de unos cincuenta repertorios o documentos[12]. Y la selección definitiva

12 Entre ellos, figuran los siguientes: *Olmos c1547 = Vocabulario de verbos nahuas*, atribuido a fray Andrés de Olmos; *Vico c1555 = Vocabulario cakchiquel, con quiché y zutujil, y castellano*, atribuido a fray Domingo de Vico; *Ara tzeldal-hispano c1560 = Vocabulario tzeldal-hispano,* atribuido a fray Domingo de Ara; *Ara hispano-tzeldal c1560 = Vocabulario hispano-tzeldal*, atribuido a fray Domingo de Ara; *Mayathan 15- = Vocabulario de Mayathan*; *Ciudad Real c1600 = Diccionario maya-hispano/Diccionario de Motul I*, atribuido a fray Antonio de Ciudad Real; *Molina 1571 = Vocabulario de la lengua castellana y mexicana, y mexicana y castellana*; *Gilberti 1559 = Vocabulario de la lengua tarasca y castellana, y castellana y tarasca*, de fray Maturino Gilberti; *Córdova 1578 = Vocabulario de la lengua castellana y zapoteca*, de fray Juan de Córdova; *Anónimo [Antonio Ricardo impresor] 1586 =* Anónimo, *Vocabulario de la lengua castellana y de la lengua quichua*; *Alvarado 1593 = Vocabulario de la lengua castellana y mixteca*, de fray Francisco de Alvarado; *Basalenque 1642 = Vocabulario castellano-matlaltzinga*, de fray Diego Basalenque; *Basalenque 1642bis = Vocabulario matlaltzinga-castellano*, de fray Diego Basalenque; *Valdivia 1606 = Arte y gramática general enla lengva qve corre en todo el Reyno de Chile, con un vocabulario*, del P. Luis de Valdivia; *Valdivia 1607 = Arte, gramatica, vocabulario, catecismo y confesionario en lengua Chilena, y en las dos lenguas Allentiac y Milcocayac*, del P. Luis de Valdivia; *González Holguín 1608 = Vocabvlario de la lengva general de todo el Perv llamada lengua qquichua, o del Inca,* de Diego González Holguín; *Arenas 1611 = Vocabulario manual de las lenguas castellana y mexicana*, de Pedro de Arenas; *Bertonio 1612 = Vocabulario de la lengva aymara*, del P. Ludovico Bertonio; *Torres Rubio 1616 = Vocabulario aimara en Arte de la lengua aymara*, del P. Diego de Torres Rubio; *Torres Rubio 1619 = Vocabulario quichua en Arte de la lengua quichua*, del P. Diego de Torres Rubio; *Ruiz [de Montoya] 1639 = Tesoro de la lengva gvaraní*, del P. Antonio Ruiz de Montoya; *Ruiz de Montoya 1640 = Arte y bocabvlario de la lengva gvaraní*; *Tauste* 1680 = *Arte, y Bocabvlario de la lengva de los indios chaymas, cvmanagotos, cores, parias, y otros diversos de la provincia de Cvmana, o Nveva Andalvcia*; *Ximénez c1700 = Vocabulario quiché-español/ Tesoro de las lenguas Cakchiquel, Quiche y Tzutuhil*, de fray Francisco Ximénez*; Coto c1650 = Vocabulario de la lengua cakchiquel y guatimalteca,* de fray Thomas de Coto; *Zúñiga c1608 = Diccionario pocomchi-castellano y castellano-pocomchi de San Cristobal*

se realizará procurando, en la medida de lo posible, el equilibrio entre áreas geográficas y periodos cronológicos.

3.3 Establecimiento del lemario o lexicón de americanismos

Tras la lectura detenida del texto español de los vocabularios, se hará una lista alfabética provisional de los americanismos inventariados.

Para la identificación de la palabra como nativa de América se contrastará su documentación en las distintas fuentes históricas, léxicas y lexicográficas: en el *Tesoro léxico canario-americano* de Corrales/Corbella (*TLCA*), *LHA*,

---

Cahcoh, de fray Dionisio de Zúñiga; *Morán 1720 = Vocabulario castellano-pokoman de Amatitan,* de fray Pedro Morán; *Hidalgo 1735 = Vocabulario breve tzotzil,* de Manuel Hidalgo; *Bernardo de Quirós 1711 = Arte y vocabulario del idioma huasteco,* de Severino Bernardo de Quirós; *Pretovio 1729 = Vocabulario de lengua guaraní,* del P. Blas Pretovio; S.J. *Arce 17- = Vocabulario de la lengua chiquita; Anónimo 1751= Vocabulario de los Indios de los ríos Putumayo y Caquetá*; *Neira y Ribero 1762 = Arte y vocabulario de la lengua achagua*; *Anónimo andaquí 17- = Vocabulario andaquí-español*; *Castillo 17- = Idioma de la provincia de Páez*, de Eugenio del Castillo, *y pliego de algunos términos del idioma de la nación murciélaga, o huaque*; *Anónimo 1765 = Vocabulario lengua aruaca*; *Taradell 1774 = Vocabulario de español a caribe*; *Anónimo 1788 ceona = Vocabulario en lengua ceona*; *Anónimo guama 1788 = Voces del idioma guama*; *Anónimo guarauno 17- = Voces del idioma guarauno*; *Anónimo pariagoto 17- = Compendio de lengua pariagoto*; *Anónimo mosca-chibcha 17- = Vocabulario en lengua mosca-chibcha*, atribuido al P. Josef Dadey; *Anónimo mosco 17- [1612] = Vocabulario mosco; Alfaro 1788 = Vocabulario castellano y lengua motilona*, de Francisco Javier de Alfaro; *Lucena 1788 = Vocabularios en lengua otomaca, taparita y yarura*, de fray Jerónimo José de Lucena; *Figueredo 1700 = Arte de la lengva qvichva, el Vocabulario añadido, y otro vocabulario de la lengua Chinchaisuyo*, de P. Diego de Torres Rubio [P. Juan de Figueredo]; *Marbán 1702 = Arte de la lengua moxa, con su vocabulario y catecismo*, del P. Pedro Marbán; *Machoni 1732 = Arte y vocabulario de la lengua lule y toconoté*, del P. Antonio Machoni de Cerdeña; *Ortega 1732 = Vocabulario en lengua castellana y cora*, del P. José de Ortega; *Rinaldini 1743 = Arte de la lengua tepeguana con vocabulario, confessionario y cathecismo*, del P. Benito Rinaldini; *Beltrán de Santa Rosa 1746 = Arte del idioma Maya, y semilexicon Yucateco*, de Pedro Beltrán de Rosa; *Febres 1756 = Arte de la lengua general del reyno de Chile… Vocabulario Hispano-Chileno y un calepino Chileno-Hispano mas copioso*, del P. Andrés Febres; *Cortés y Zedeño 1765 = Arte, vocabulario y confesionario en el idioma mexicano*, de Fr. D. Gerónimo Thomas de Aquino Cortés y Zedeño; *Neve y Molina 1767 = Reglas de la Orthographia, Diccionario, y Arte del idioma Othomi; Tapia Zenteno 1767 = Noticia de la lengua huasteca y copioso Diccionario para facilitar su inteligencia*, de Carlos Tapia Zenteno (véase información bibliográfica completa de estas obras en Hernández 2018).

CORDE, *CORPES XXI*, *CORDIAM*, *Corpus del español*, *Fichero general*, *NDLE*; en los diccionarios de provincialismos hispanoamericanos, diccionarios de americanismos (Morínigo, Neves, Malaret, etc.), en los atlas lingüísticos (los derivados del *Atlas Lingüístico de Hispanoamérica* [*ALH*] de Alvar/Quilis, pero también los peninsulares y de las Canarias), y en otros estudios dialectales y diacrónicos.

Se manejan cifras para los indigenismos, pero son más difusas las de las palabras del fondo léxico patrimonial, que están adaptadas morfológica o semánticamente a la realidad americana, cuya identificación es más costosa. Para los indigenismos, pueden servir como orientación los resultados obtenidos anteriormente. Así, en el caso del vocabulario de *Molina 1571* y en el *Calepino* de Sahagún, se codificaron alrededor de 80 palabras de procedencia indígena en cada documento; en cuatro vocabularios sudamericanos del siglo XVIII se identificaron 28 antillanismos, 15 nahuatlismos, 3 guaranismos, 20 mapuchismos, 59 quechuismos y 19 vocablos de origen discutido (Hernández, en prensa).

3.4 Criterios lexicográficos del tesoro de americanismos

A continuación, se exponen de manera sintética los criterios lexicográficos más relevantes que en principio se siguen en la elaboración del tesoro. Se tienen en cuenta, como punto de partida, los planteamientos del *NDHLE* (Pascual/García 2007, Garcés Gómez 2008, entre otros).

Para la lematización, se siguen criterios cuantitativos, seleccionando y definiendo los vocablos que tengan un mayor número de documentaciones. Los préstamos de otras lenguas no adaptados a los patrones gráfico-fonológicos del español, o extranjerismos crudos, se incluirán en el lemario en cursiva. Cuando se documenten suficientemente en el uso las formas crudas y las formas adaptadas de un mismo préstamo, se registrarán las dos en el tesoro, definiendo la cruda por remisión a la adaptada.

Un criterio de garantía para la inclusión de un préstamo en el lemario es su productividad morfológica, esto es, que se haya convertido en base de derivados (*encolehuado* < map. *coligüe*, *guanaquear* < qch. *guanaco*, ambos documentados en *Febres 1765*).

Para la documentación de la palabra o de la acepción se considera el criterio de frecuencia en su tiempo. Un rastreo en fuentes coetáneas es imprescindible para comprobar hasta qué punto la palabra estaba en el habla de la época. Se proporcionan las documentaciones más ilustrativas en este sentido, ordenadas cronológicamente y por áreas lingüísticas.

En principio, se siguen las áreas lingüísticas que aparecen en las obras académicas –como la *NGLE*–, esto es: Chile, Río de la Plata, Área andina, Caribe continental, México y Centroamérica y Antillas.

Se emplea un sistema de abreviaturas de los vocabularios y otras fuentes conforme a la tradición lexicográfica española.

Cada vocablo se contrasta en las distintas fuentes documentales almacenadas en los corpus, diccionarios y otros estudios, o incluso en las aplicaciones de *Google* (*Google Scholar* o *Google Books*)[13], y, por último, si fuera el caso, también en las fuentes del español europeo o de otras lenguas; en definitiva, se revisan las fuentes que revelan mejor su trayectoria histórica y geográfica. Por ejemplo, para documentar palabras de los campos de la flora y de la fauna, los datos se contrastan con el *Lexicón* de Malaret; para este campo semántico, también se prevé la utilización del fondo digitalizado de la biblioteca del *Real Jardín Botánico*, que presenta una lista de autores y obras muy adecuadas a ese fin.

La definición es breve y lo más ajustada posible al significado de las citas de los vocabularios, aunque también se tiene en cuenta la definición que proporcionan los diccionarios de provincialismos o los sincrónicos de referencia.

Las transcripciones de los extranjerismos crudos se hacen conforme a las normas de escritura reconocidas de las distintas lenguas indígenas. En cuanto al origen o étimo de la palabra, se proporciona, si es el caso, el étimo con la forma con la que aparece en el vocabulario.

Como todo trabajo léxico, el repertorio no será sino un primer acercamiento que podrá irse completando en el futuro con nuevos datos aportados por otros documentos. Los criterios técnicos, lógicamente, se irán actualizando a medida que avance la elaboración del tesoro léxico.

4 Alcances, resultados y perspectivas futuras

4.1 Alcances del estudio filológico y lingüístico de los americanismos

La creación léxico-semántica en América es un proceso ininterrumpido que dejó evidencias en los vocabularios amerindios. Por ello, el análisis filológico y lingüístico detenido de los vocablos identificados como americanismos en dichos vocabularios, y su disposición en forma de tesoro léxico, trata de explicar, fundamentalmente, aspectos etimológicos, de variación fonética, morfológica

13 Teniendo en cuenta los debates actuales de la lingüística histórica y la lexicografía sobre la fiabilidad científica de lo que se ha dado en llamar «Googleology» (Dollinger 2016).

y semántica de los mismos, y también su dispersión geográfica y su recorrido histórico.

Por consiguiente, el alcance de *TELEAM* no es otro que tratar de proporcionar una descripción individualizada de cada americanismo lo más completa posible, al servicio de las necesidades prácticas del uso lexicográfico. Para ello, se toman como modelo, según ya se ha señalado, estudios filológicos hechos con rigor y amplitud documental, estudiando la palabra a la luz de su marco dialectal y sociocultural, esto es, de sus contextos y/o citas. Con todo, se ahonda en el estudio de los procesos de cambio del léxico en el español americano y en los distintos componentes que lo han forjado. Y, como ya señaló Kany en el primer estudio importante de conjunto sobre semántica hispanoamericana: «de mayor interés que la riada de nuevas aportaciones indias en la América española son los cambios semánticos que, acelerados por el contacto de los españoles con los nuevos ambientes, experimentaron las propias palabras españolas» (Kany 1962: 5).

Los neologismos documentados podrán estudiarse desde dos perspectivas, por un lado a partir del análisis de los recursos que se emplean en la propia creación léxica o las vías de la americanización (préstamo o extranjerismo, o procedimientos de analogía, metáfora, descripción, etc.); y, por otro, en relación con los campos léxicos (de la flora, de la fauna y de otros vocabularios, como la minería o la alimentación). Se procurará hacer una tipología de los cambios léxicos combinando ambas perspectivas.

4.2 Resultados de *TELEAM*

El proyecto aporta pruebas documentales decisivas y ofrece los correspondientes análisis lingüísticos que aseguran la etimología y el origen de muchos americanismos que siguen sin estar bien establecidos. Según se ha tratado de mostrar, por ser los vocabularios textos bilingües, el trasvase de elementos léxicos nuevos de una lengua a otra, o de una cultura a otra, es mucho más directo y más visible. En el propio acto de la traducción (García Yebra 1985), los neologismos surgen más fácilmente. Y en este caso lo hacen dentro del texto castellano de las entradas o de las definiciones, incorporados a la lengua española mediante diversos procedimientos formales y semánticos, y presentan distintos niveles de asimilación fonética y morfológica.

Los indigenismos, en muchos casos, se documentan en contacto directo con la palabra de la cual proceden –según veíamos en *chuchoca*–, de modo que la información etimológica que aportan puede ser clave para completar los datos de los corpus o diccionarios históricos. Pero además, es un género específico de América, que dura toda la época colonial y que contiene un español de un nivel

diastrático culto. Una de las características más relevantes de este género textual es que los autores que redactan los diccionarios fueron todos misioneros de las distintas órdenes religiosas, quienes los escribían para su propio uso interno y que vivieron de manera muy directa la situación de contacto del español con las lenguas indígenas. *TELEAM* proporciona, en suma, testimonios léxicos imprescindibles para conocer mejor los distintos procesos de variación y cambio léxico del español de América en sus diversos territorios y a lo largo de distintas épocas, y dan muestras fiables de la historia compartida entre el español y las lenguas indígenas durante siglos.

4.3 Perspectivas de futuro: Estudios sobre neología, variación y cambio léxico-semántico

Como ya se ha señalado, los vocabularios suministran materiales para conocer los cambios léxicos o de significado del español americano. Pero además, a partir del texto del español inserto en los mismos, se pueden llevar a cabo análisis de las *relaciones léxicas*, dado que contienen entradas pluriverbales que facilitan la observación de los fenómenos en niveles no solo léxicos. Aparte está, como es lógico, el hecho de que las definiciones tienen texto aprovechable para los análisis gramaticales o semánticos.

Un ejemplo de este caso sería el uso metafórico del adjetivo *barbado/a* asociado a una planta que aportan los vocabularios novohispanos *Gilberti 1559*, *Córdova 1578* o *Alvarado 1593*, todos ellos a partir de *Molina 1555*. Según los corpus históricos, se trata de una relación léxica que no se establece en el español general, si bien su significado puede resultar transparente y tendría la metáfora como procedimiento de denominación. Pues bien, hoy encontramos referencias de «plantas barbadas» como voz técnica de jardinería o de horticultura en páginas web americanas. Pero acaso una explicación de la motivación de esta relación léxica, o coocurrencia, se pueda descubrir partiendo de la equivalencia de una palabra del náhuatl. Véase la siguiente cita de *Francisco Hernández 1616*: «A esta planta llaman *tlacoxiloxochit* y otros *xiloxochitl* y otros *tzonxochitl*, o *flor barbada*, y otros la llaman *tlamacazcaçotl*, y otros *tepexiloxochitl*, y *tlaxiloxochitl*, debaxo de los quales nombres se entiende una mata que tiene las ojas de *Mizquitl*, con las flores a manera de cauellos largos y rojos». En este caso, nuestra suposición es que se trata de un proceso relativamente frecuente, que está pendiente de análisis, y del que los vocabularios pueden arrojar luz.

La descripción de los fenómenos que muestran por dónde se innova la lengua en América, y de los mecanismos involucrados en tales cambios, conducirá a la elaboración de otros trabajos que describirán tales hechos y procedimientos,

dentro de una perspectiva más general o desde un planteamiento más teórico, y teniendo en cuenta los proyectos y los estudios sobre préstamos que se llevan a cabo para otras lenguas (Haspelmath 2008).

Referencias bibliográficas

Acuña, René (ed.) (1984): *Calepino maya de Motul* [de Antonio de Ciudad Real]. México: Universidad Nacional Autónoma de México.

ALH = Alvar, Manuel/Antonio Quilis (1984): *Atlas Lingüístico de Hispanoamérica. Cuestionario*, estudios introductorios de Manuel Alvar. Madrid: Instituto de Cooperación Iberoamericana.

Alvar Ezquerra, Manuel (1987): «La recepción de americanismos en los diccionarios generales de la lengua», en Humberto López Morales y María Vaquero (eds.), *Actas del I Congreso Internacional sobre el Español de América. San Juan de Puerto Rico, 4 a 9 de octubre de 198*2. San Juan: Academia Puertorriqueña de la Lengua Española, 215–218.

Alvar Ezquerra, Manuel (1997): *Vocabulario de indigenismos en las crónicas de Indias*. Madrid: Consejo Superior de Investigaciones Científicas.

Alvar, Manuel (1975): *España y América cara a cara*. Valencia: Editorial Bello.

Alvarado 1593 = Alvarado, fray Francisco (1593): *Vocabulario de la lengua castellana y mixteca*. México: Pedro Balli.

Ara c1571 = Ara, Domingo de (1986 [c1571]): *Vocabulario de lengua tzeldal segun el orden de Copanabastla*, edición de Mario Humberto Ruz. México: Universidad Nacional Autónoma de México.

Arzápalo Marín, Ramón, *et al.* (eds.) (1995): *Calepino de Motul: diccionario maya-español*. México: Universidad Nacional Autónoma de México.

Boyd-Bowman, Peter (2015): *Léxico Hispanoamericano (1493–1993)*. <https://textred.spanport.lss.wisc.edu/> [último acceso: 15/10/2017].

Buesa Oliver, Tomás (1965): *Indoamericanismos léxicos en español*. Madrid: CSIC.

Buesa Oliver, Tomás/José Mª Enguita Utrilla (1992): *Léxico del español de América: su elemento patrimonial e indígena*. Madrid: Mapfre América.

Carriazo Ruiz, José Ramón (2014): «Los indigenismos en el diccionario crítico etimológico castellano e hispánico de Joan Corominas y José Antonio Pascual», *Epos* XXX, 147–160.

CDH = *Corpus del nuevo diccionario histórico*, v. *NDHLE*.

Cerrón-Palomino, Rodolfo (2010): «Acotaciones al *Diccionario de americanismos*», *Lexis* XXXIV, 1, 161–176.

Códice florentino = Sahagún, fray Bernardino de (1979): *Códice Florentino de fray Bernardino de Sahagún*. Florencia: Talleres Casa Editorial Giunti Barberá, 3 vols.

CORDE = Real Academia Española: *Corpus diacrónico del español*. <http://www.rae.es> [último acceso: 15/10/2017].

CORDIAM = *Corpus Diacrónico y Diatópico del Español de América*. Academia Mexicana de la Lengua. <www.cordiam.org> [último acceso: 15/10/2017].

Córdova 1578 = Córdova, fray Juan (1578): *Vocabulario de la lengua castellana y zapoteca*. México: Pedro Ocharte – Antonio Ricardo.

CORPES XXI = Real Academia Española: *Banco de datos (CORPES XXI) [en línea]. Corpus del Español del Siglo XXI (CORPES)*. <http://www.rae.es> [último acceso: 15/10/2017].

Corpus del español = Davies, Mark: *Corpus del español* [en línea]. <http://www.corpusdelespanol.org> [último acceso: 15/10/2017].

Covarrubias 1611 = Covarrubias, Sebastián de (1611): *Tesoro de la lengua castellana o española*. Madrid: Luis Sánchez.

Cuervo, Rufino José (1901a): «Sabana», *Romania* XXX, 123–127.

Cuervo, Rufino José (1901b): «Canoa», *Romania* XXX, 120–122.

Dakin, Karen/Søren Wichmann (2000): «*Cacao* and *chocolate*. A Uto-Aztecan perspective», *Ancient Mesoamerica* 11, 55–75.

DECH = Corominas, Joan/José Antonio Pascual (1980–1991): *Diccionario crítico etimológico castellano e hispánico*. Madrid: Gredos.

Diccionario de Americanismos = Asociación de Academias de la Lengua Española (2010): *Diccionario de americanismos*. Lima: Santillana.

Diccionario de Motul = *Diccionario maya*, atribuido a fray Antonio de Ciudad Real. [JCBL Ms. Codex 8].

DLE = Real Academia Española/Asociación de Academias de la Lengua Española (2014[23]): *Diccionario de la lengua española*. Barcelona: Espasa Libros.

Dollinger, Stefan (2016): «Googleology as Smart Lexicography: Big Messy Data for Better Regional Labels», *Dictionaries: Journal of the Dictionary Society of North America* 37, 60–98.

Durkin, Philip (ed.) (2016): «Preface», en *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press.

Enguita Utrilla, José María (2004): *Para la historia de los americanismos léxicos*. Fráncfort: Peter Lang.

Febres 1765 = Febres, Andrés S. J. (1765): *Arte de la lengua general del reyno de Chile… Vocabulario Hispano-Chileno y un calepino Chileno-Hispano mas copioso*. Lima: [en la calle de la Encarnación].

Fichero general = *Fichero general de la Real Academia Española*, de acceso electrónico a partir del *NDHLE*.

Frago Gracia, Juan Antonio (1998–1999): «Sobre la criollización del español en América: apuntes documentales y metodológicos», B*oletín de Filología de la Universidad de Chile* XXXVII, 523–539.

Frago Gracia, Juan Antonio (1999): *Historia del español de América. Textos y contextos*. Madrid: Gredos.

Frago Gracia, Juan Antonio (2003): «*Alarife*, un americanismo léxico entre la tradición y la innovación», *Revista de Filología Española* LXXXIII, 291–298.

Frago Gracia, Juan Antonio (2010): *El español de América en la Independencia*. Santiago de Chile: Taurus.

Frago Gracia, Juan Antonio (2012): «Filología y lexicografía. Notas americanas», en Dolores Corbella *et al.* (eds.), *Lexicografía hispánica del siglo XXI: nuevos proyectos y perspectivas. Homenaje al profesor Cristóbal Corrales Zumbado*. Madrid: Arco/Libros, 251–263.

Francisco Hernández 1615 = Hernández de Toledo, Francisco (1615): *Quatro libros de la naturaleza y virtudes de las plantas y animales*. México: Viuda de Diego López Dávalos.

Franco Figueroa, Mariano (1992): *Léxico hispanoamericano de los siglos XVI y XVII en fuentes de América Central y de la Nueva España*. Sevilla: Universidad de Sevilla.

Friederici, Georg (1960): *Amerikanistisches Wörterbuch* [Hamburg: Cram, de Gruyter, 1947] *und Hilfswörterbuch für den Amerikanisten* (Auxiliary dictionary for Americanists) [Halle: Max Niemeyer, 1926]. Hamburgo: Cram, de Gruyter & Co.

Garcés Gómez, Mª Pilar (ed.) (2008): *Diccionario histórico: nuevas perspectivas lingüísticas*. Madrid/Fráncfort: Iberoamericana/Vervuert.

García Yebra, Valentín (1985): «El neologismo», *Tradução e Comunicação* 7, 21, 32.

Gilberti 1559 = Gilberti, fray Maturino (1559): *Vocabulario de la lengua tarasca y castellana, y castellana y tarasca*. México: Juan Pablos Bressano.

Glessgen, Martin-Dietrich (1997): «Prolegómenos para un *Diccionario Histórico de Americanismos* (1492–1836)», en Luis Jaime Cisneros y José Luis Rivarola (eds.), *Italica et Romanica. Festschrift für Max Pfister zum 65. Geburtstag*, t. 1. Tubinga: Niemeyer, 403–434.

Grace, Lee Ann (1976): *The Effect of Bilingualism on Sixteenth-Century Mexican Spanish*. SUNY at Buffalo [tesis doctoral inédita].

Guitarte, Guillermo L. (1996): «Una carta de Amado Alonso a Rodolfo Lenz. El proyecto de un *corpus* de estudios sobre el español extrapeninsular», *Lexis*.

Revista de Lingüística y Literatura XX, 1–2, 63–86. [Luis Jaime Cisneros y José Luis Rivarola (eds.), *Centenario de Amado Alonso (1896–1996)*: *Temas de filología hispánica*].

Haspelmath, Martin (2008): «Loanword Typology: Steps toward a systematic cross-linguistic study of lexical borrowability», en Thomas Stolz, Dik Bakker y Rosa Salas Palomo (eds.), *Aspects of Language Contact: New Theoretical, Methodological and Empirical*. Berlín: Mouton de Gruyter, 43–63.

Henríquez Ureña, Pedro (1935): «Palabras antillanas en el *Diccionario* de la Academia», *Revista de Filología Españo*la XXII, 175–186.

Henríquez Ureña, Pedro (1938): *Para la historia de los indigenismos. Papa y batata, el enigma del aje, boniato, caribe, palabras antillanas*. Buenos Aires: Editorial de la Universidad. [Anejo III de la *Biblioteca de Dialectología Hispanoamericana*].

Henríquez Ureña, Pedro (1944): «*Papa* y *batata*. Notas adicionales», *Revista de Filología Hispánica* VI, 387–394.

Hernández, Esther (1996): *Vocabulario en lengua castellana y mexicana de fray Alonso de Molina. Estudio de los indigenismos léxicos y registro de las voces españolas internas*. Madrid: CSIC.

Hernández, Esther (1999): «Revisión de los mayismos del diccionario de la Academia (21ª ed.)», *Lexis. Revista de Lingüística y Literatura* XXIII, 337–358.

Hernández, Esther (2000): «Propuestas etimológicas para palabras de origen indoamericano (*DRAE*, 21ª ed.)», *Boletín de la Real Academia Española* LXXX, CCLXXXI, 361–396.

Hernández, Esther (2008): «Indigenismos en el Vocabulario de la lengua cakchiquel atribuido a fray Domingo de Vico, ms. BNF R. 7507», *Revista de Filología Española* LXXXVIII, 1, 67–88.

Hernández, Esther (2011a): «Para la historia de *camiseta*, un americanismo inadvertido», *Nueva Revista de Filología Hispánica* 50, 2, 539–550.

Hernández, Esther (2011b): «Nahuatlismos de México con la primera documentación en la obra de Sahagún», en José Rubén Romero Galván y Pilar Máynez (coords.), *El Universo de Sahagún. Pasado y presente*. México: UNAM, 119–147.

Hernández, Esther (2018): *Lexicografía hispano-amerindia (1550–1800). Catálogo descriptivo de los vocabularios bilingües con las lenguas indígenas*. Madrid/Fráncfort: Iberoamericana/Vervuert.

Hernández, Esther (en prensa): «El impacto de las lenguas indígenas en los vocabularios indoamericanos de los jesuitas. Siglo XVIII», en Álvaro

Ezcurra, Rodolfo Cerrón-Palomino y Otto Zwartjes (eds.), Lima/Ámsterdam: Pontificia Universidad del Perú/John Benjamins.

Jacinto García, Eduardo José (2016): «La información sintagmática en la lexicografía española actual: unidades pluriverbales, ejemplos lexicográficos y otras indicaciones cotextuales», *Cuadernos Aispi* 6, 147–170.

Jardín Botánico = Biblioteca digital del Real Jardín Botánico. CSIC.

Kany, Charles E. (1962 [1960]): *Semántica hispanoamericana*. Madrid: Aguilar.

LHA = v. Boyd Bowman 2015.

Lara, Luis Fernando (2004): «Culturas nacionales y léxico contemporáneo del español», en Jens Lüdtke y Christian Schmitt (coords.), *Historia del léxico español: enfoques y aplicaciones: homenaje a Bodo Müller*. Madrid/Fráncfort: Iberoamericana/Vervuert.

Lenz, Rodolfo (1904–1910): *Diccionario etimolójico de las voces chilenas derivadas de lenguas indígenas americanas*. Santiago de Chile: Imprenta Cervantes.

Lerner, Isaías (1974): *Arcaísmos léxicos del español de América*. Madrid: Ínsula.

Lope Blanch, Juan M. (1969): *El léxico indígena en el español de México*. México: El Colegio de México.

López Morales, Humberto (1974): «Indigenismos en los textos cronísticos de Puerto Rico: índices de frecuencia y densidad», *Estudios filológicos y lingüísticos. Homenaje a Ángel Rosenblat en sus 70 años*. Caracas: Departamento de Cultura y Publicaciones del Inst. Pedagógico de Caracas, 337–346.

Lüdtke, Jens (ed.) (1994): *El español de América en el siglo XVI: Actas del Simposio del Instituto Ibero-Americano de Berlín, 23 y 24 de abril de 1992*. Fráncfort: Vervuert.

Lüdtke, Jens (2014): *Los orígenes de la lengua española en América. Los primeros cambios en las Islas Canarias, las Antillas y Castilla del Oro*. Madrid: Iberoamericana Vervuert.

Malaret, Augusto (1946): *Diccionario de americanismos*. Buenos Aires: Emecé Editores.

Malaret, Augusto (1961): *Lexicón de fauna y flora*. Bogotá: Instituto Caro y Cuervo.

Martinell Grife, Emma (1993): «La incorporación de americanismos al español y su adaptación», *The Bulletin of the International Institute for Linguistic Sciences* XV, 97–111.

Martínez Hernández, Juan (1929): *Diccionario de Motul. Maya-Español*. Mérida: Talleres de la Compañía Yucateca.

Mejías, Hugo A. (1980): *Préstamos de lenguas indígenas en el español americano del siglo XVII*. México: UNAM.

Menéndez Pidal, Ramón (1944): «Los incunables americanos», *Doctrina cristiana en lengua española y mexicana* por los religiosos de la Orden de Santo Domingo, México, Juan de Pablos, 1548 en *Colección de Incunables Americanos Siglo XVI*. Madrid: Ediciones de Cutura Hispánica, vii–xxii.

Molina 1555 = Molina, fray Alonso (1555): *Aqui empieza el vocabulario de la lengua castellana y mexicana*… México: Juan de Pablos.

Molina 1571 = Molina, fray Alonso (1571): *Vocabulario de la lengua castellana y mexicana, y mexicana y castellana*. México: Antonio de Spinosa.

Morínigo, Marcos Augusto (1964): «La penetración de los indigenismos americanos en el español», en *Presente y futuro de la lengua española*. Madrid: Oficina Internacional del Español, II, 217–226.

Morínigo, Marcos Augusto (1966): *Diccionario de Americanismos*. Buenos Aires.

NDHLE = Instituto de Investigación Rafael Lapesa de la Real Academia Española (2013–): *Nuevo diccionario histórico de la lengua española*. <http://web.frl.es/DH> [último acceso: 15/10/2017].

Neves, Alfredo N. (1973): *Diccionario de americanismos*. Buenos Aires: Sopena.

NGLE = Real Academia Española-Asociación de Academias de la Lengua Española (2010): *Nueva gramática de la lengua española. Manual*. Madrid: Espasa.

Oesterreicher, Wulf/Eva Stoll/Andreas Wesch (eds.) (1998): *Competencia escrita, tradiciones discursivas y variedades lingüísticas. Aspectos del español europeo y americano en los siglos XVI y XVII. Coloquio internacional, Friburgo en Brisgovia, 26–28 de septiembre de 1996*. Tubinga: Narr.

Pascual, José Antonio/Rafael García Pérez (2007): *Límites y horizontes en un diccionario histórico*. Salamanca: Diputación de Salamanca.

Portal del léxico hispánico = *Portal del Léxico Hispánico*, Seminario de Filología e Informática, Universidad Autónoma de Barcelona.

Rivarola, José Luis (1985): «Para la historia de los americanismos léxicos, A propósito de la nueva versión de la Relación de Pedro Pizarro», *Filología* XX, 1, 69–88.

Rivarola, José Luis (1990): «Los *baquianos* de América», en *La formación lingüística de Hispanoamérica*. Lima: Pontificia Universidad Católica del Perú, 79–90.

Rojo Sánchez, Guillermo (2012): «El papel de los corpus en el estudio de la historia del español», en E. Montero Cartelle y C. Manzano Rovira (coords.),

Actas del VIII Congreso Internacional de Historia de la Lengua Española Santiago de Compostela, 14–18 de septiembre de 2009. Meubook: Asociación de Historia de la Lengua Española, 433–444.

Rona, José Pedro (1969): «¿Qué es un americanismo?», en *Simposio de México. Actas, informes y comunicaciones*. México: UNAM, 310–321.

Rosenblat, Ángel (1973): *El castellano de España y el castellano de América. Unidad y diferenciación*. Madrid: Taurus.

Sala, Marius/Dan Munteanu/Valeria Neagu/Tudora Sandru-Oleanu (1977): *El léxico indígena del español americano. Apreciaciones sobre su vitalidad*. México: Academia mexicana.

Seco, Manuel (2003): *Estudios de lexicografía española*. Madrid: Gredos.

TLCA = Corrales, Cristóbal/Dolores Corbella (2010): *Tesoro léxico canario-americano*. Las Palmas de Gran Canaria: Casa de Colón.

Torres Torres, Antonio (2004): *Procesos de americanización del léxico hispánico*. Valencia: Universidad de Valencia.

Zamora Munné, Juan Clemente (1976): *Indigenismos en la lengua de los conquistadores*. [Río Piedras]: Editorial Universitaria, Universidad de Puerto Rico [Barcelona: Artes Gráficas Medinaceli].

Zamora Munné, Juan Clemente (1982): «Amerindian loanword in general and local varieties of American Spanish», *Word* 33 (1–2), 159–171.

Dolores Corbella

Del tesoro lexicográfico analógico al digital[1]

Resumen: El modelo de *Tesoro lexicográfico* que desarrolló Samuel Gili Gaya durante la primera mitad del siglo XX fue retomado en la década de los noventa por algunos investigadores que han querido dar cuenta de la historia del patrimonio léxico del español a través de las fuentes diccionarísticas. En este trabajo revisamos algunas de las publicaciones más relevantes de los últimos años de este tipo de memoria sistemática del acervo léxico y, dentro del contexto hispánico, registramos la aparición de los tesoros lexicográficos de ámbito geográfico restringido, tanto en el español europeo como en el americano. Al mismo tiempo, analizamos cómo la incorporación de los recursos informáticos ha influido en la planta de estas obras de referencia, tanto por las nuevas técnicas de indexación de unos materiales tan amplios como por las múltiples opciones de búsqueda relacional que ofrecen como bases de datos lexicográficas.

Palabras clave: Lexicografía española, Tesoro lingüístico, Edición digital

Abstract: The lexicographical model developed by Samuel Gili Gaya throughout the first half of the twentieth century and illustrated by his *Tesoro lexicográfico* has been revisited ever since. This has been the case with researchers having to account for the history of the Spanish lexical patrimony by means of the study of its dictionary sources. In this work we survey some of the most relevant publications released in recent years in the field of the systematic memory of the lexical wealth. As for the Hispanic context, we also record the emergence of restricted scope lexicographical thesauruses, both for European and American Spanish. Simultaneously, we analyse in what ways integrating computer resources may have altered the layout of these reference titles. The work pays attention, therefore, to new indexation techniques for a wider range of materials, as well as to the multiple choices of relational search being offered as lexicographical databases.

Keywords: Spanish lexicography, Linguistic treasure, Digital editing

1 Introducción

Entre las fuentes para el estudio histórico del léxico español no puede faltar la memoria lexicográfica de la lengua, el conjunto de repertorios que a lo largo

1 Este trabajo se enmarca en el proyecto de investigación FFI2016-76154-P, financiado por el Ministerio de Economía y Competitividad del Gobierno de España.

de los siglos y con mayor o menor acierto ha ido dando cuenta de su acervo patrimonial, de la incorporación de préstamos y neologismos o de la evolución misma del significado de cada palabra. No quiere esto decir que el diccionario histórico deba basarse en la suma de monografías diccionarísticas conocidas, ya que la lexicografía histórica debe partir del uso real de las palabras en los textos. Pero en muchas ocasiones, especialmente cuando se trata de abordar el léxico diferencial, los corpus documentales apenas aportan testimonios antiguos de voces que tuvieron una transmisión exclusivamente oral y que fueron recogidas por vez primera por los lexicógrafos, diletantes o no, que con perspicacia y fina intuición advirtieron su singularidad y anotaron su empleo.

Hacia 1920 Samuel Gili Gaya inició la ordenación de los materiales léxicos que contenían los diccionarios de Covarrubias y Nebrija, «para tener con ellos un instrumento de trabajo destinado a nuestras investigaciones lingüísticas en el Centro de Estudios Históricos» (1960: vii), a los que pronto añadiría, siguiendo la recomendación de Ramón Menéndez Pidal, la mayor parte de la lexicografía clásica a manera de «Corpus Glossariorum». Surgió así la planta de un nuevo modelo lexicográfico, un «diccionario de diccionarios» que Amado Alonso (1951: 324) llegó a calificar como la obra más importante en nuestra lengua después del *Diccionario de Autoridades*. El proyecto de aquel *Tesoro lexicográfico*, del que Gili Gaya solo podría publicar los fascículos del primer tomo entre 1947 y 1957 (Vila Rubio 2014), fue retomado en la década de los noventa del siglo XX y, gracias a las nuevas tecnologías, en los inicios del siglo XXI ha resurgido la necesidad de contar con este tipo de compilación lexicográfica, no solo en el ámbito del español sino también en otras lenguas cercanas.

2 Recepción peninsular del modelo de Gili Gaya

El *Tesoro* de Gili Gaya fue tomado como modelo por Dieter Messner en su *Dicionário dos dicionários portugueses (1562–1858)*, una obra también inacabada que el investigador de la Universidad de Salzburgo concibió con la idea de contribuir al conocimiento histórico del léxico luso y reconstruir la evolución de la lexicografía portuguesa. Para ello recopiló los principales diccionarios de esta lengua entre mediados del siglo XVI y 1858, año de la publicación de la sexta edición del «Moraes». Como novedad, incluyó, además, «algunas listas de palabras provenientes de obras no filológicas, en caso de que contengan informaciones suplementarias (sobre el origen, el significado, el registro, o una traducción a otra lengua)» (Messner 2008: 34–35). A pesar de que la obra quedó inconclusa, el objetivo de elaborar un texto de referencia que compile la memoria lexicográfica portuguesa ha tenido como continuador el *Corpus Lexicográfico do Português*

(*DICI-web*), un proyecto de la Universidad de Aveiro con la colaboración del Centro de Lingüística de la Universidad de Lisboa[2], que reúne un total de veinticinco textos de carácter lexicográfico o paralexicográfico[3]. Con una interfaz muy sencilla, *DICI-web* permite el acceso a siete millones y medio de registros, ya sean lemas o voces documentadas en el interior de las definiciones, así como a las concordancias de las búsquedas solicitadas e, incluso, a la edición on-line de cada una de las obras.

La huella de Gili Gaya parece advertirse asimismo en el *Dicionario de dicionarios* (*DdD*) elaborado para el gallego[4]. En una reseña sobre esta obra, Gutiérrez Cuadrado (2003: 370) resaltaba que «Antón Santamarina y sus colaboradores vierten el vino añejo de obras lexicográficas gallegas (no sólo diccionarios) en un odre nuevo que, entre otras virtudes está dotado de poder mágico: almacena todos los vinos sin que se mezclen y permite a los que se acercan a él catarlos cómodamente». Con esta metáfora, Gutiérrez Cuadrado ponía de relieve la confluencia del quehacer lexicográfico tradicional con las nuevas tecnologías, ya que por vez primera (antes incluso que *DICI-web*) conseguía crearse un entorno virtual que facilitara la consulta unificada de todos los lemas que los distintos textos aportaban al tesoro lexicográfico. Junto a la novedad que suponía la navegación relacional por unas obras que se habían digitalizado íntegramente, las fuentes no quedaron supeditadas a los diccionarios, sino que se enriquecieron con los aportes de trabajos dialectales y de las terminologías populares que aparecían en monografías de botánica o de zoología, incluyendo algunos manuscritos inéditos, con un marco cronológico que abarcaba desde el siglo XVIII hasta 1992 en la última versión[5]. Con ello se ha conseguido poner a disposición de

2 Según señalan Verdelho/Silvestre (2014: 300), el proyecto «Comenzó en la Universidad de Aveiro, dando continuidad a una investigación iniciada hace más de treinta años en la Sorbona (Universidad de París IV) sobre el patrimonio lexicográfico y sobre la historia de las primeras organizaciones alfabéticas del vocabulario de la lengua portuguesa».

3 «El conjunto, si bien no reúne toda la lexicografía del portugués de los siglos XVI–XVIII, por lo menos recoge seguramente los textos que mejor representan la estructuración de un vocabulario normalizado y que tuvieron mayor uso en el contexto escolar, contribuyendo a expandir el léxico del portugués con el pretexto de apoyar el estudio del latín» (Verdelho/Silvestre 2014: 308–309).

4 Coordinado por Antón Santamarina, se publicó primeramente en CD-ROM, con tres ediciones sucesivas en 2000, 2001 y 2003. La última versión actualizada puede consultarse a través de la página web <http://sli.uvigo.es/DdD>.

5 Desde la obra de fray Martín Sarmiento hasta la *Nomenclatura vernácula da flora vascular galega*, de Losada Cortiñas, J. Castro González y E. Niño Ricoi, de 1992. En total, se

los investigadores una herramienta de trabajo utilísima que integra los 392 768 lemas extraídos de los repertorios indexados y que constituirá uno de los pilares de referencia para la elaboración del futuro diccionario histórico del gallego.

El prototipo del *DdD* tuvo continuación en el *Dicionario de dicionarios do galego medieval. Corpus lexicográfico medieval da lingua galega* (*DDGM*)[6], dirigido por Ernesto González Seoane, de la Universidad de Santiago de Compostela. Su novedad consiste en que «es un multidiccionario electrónico que compila catorce glosarios y vocabularios —cuatro de ellos inéditos— elaborados a partir de textos o colecciones textuales medievales gallegas o pertenecientes a la tradición común gallego-portuguesa» (González Seoane *et al*. 2008: 385)[7]. En total, se han incorporado 25 109 lemas procedentes de todas estas obras. La heterogeneidad de los materiales recopilados llevó al equipo de investigación a redactar una plantilla de metadatos en la que se tuvieron en cuenta, aparte del lema (discriminando topónimos, antropónimos y unidades fraseológicas), la categoría gramatical, el étimo, la definición (en gallego, portugués o español), las citas textuales y las fuentes. El *DDGM* se convierte así en un verdadero diccionario relacional donde las posibilidades de búsqueda —simples o complejas— se multiplican, al poder combinarse también operadores lógicos. Sin duda, supone un paso adelante en la concepción misma del *tesoro* como herramienta digital, aparte de la utilidad que posee como fuente de análisis del léxico medieval y de la historia misma del vocabulario.

---

ofrecen 32 léxicos diferentes, algunos de ellos correspondientes a diversas ediciones de un mismo texto (como las cuatro entregas del *Diccionario galego-castelán*, de Leandro Carré Alvarellos, de 1928, 1933, 1951 y 1972); en otras ocasiones, un único registro integra diferentes obras de un mismo autor, como sucede con Fray Martín Sarmiento, Juan Sobreira Salgado, Aníbal Otero o Elixio Rivas.

6 Una primera versión se publicó el año 2006 en formato CD-ROM, como Anexo de la revista *Verba*. Con similar formato que el *DdD*, puede consultarse la misma edición en la web: <http://sli.uvigo.es/DDGM/index.html>.

7 «El corpus incluye vocabularios de textos poéticos, tanto del cancionero profano (los preparados por José Joaquim Nunes, Manuel Rodrigues Lapa y Carolina Michäelis de Vasconcelos), como del religioso (el glosario del cancionero mariano de Alfonso X elaborado por Walter Mettmann); así como vocabularios de obras en prosa de diverso tipo: historiográficas (los trabajos de Ramón Lorenzo y de Kelvin M. Parker), hagiográficas (la obra inédita de María del Carmen Barreiro) y notariales (los repertorios léxicos realizados por Fernando Tato, Montserrat de Miguel y María del Carmen Barreiro a partir de diversas colecciones documentales); y también la publicación *Sobre cronología do vocabulario galego-português* de Ramón Lorenzo» (González Seoane *et al*. 2008: 385).

Otro de los últimos macroproyectos lexicográficos que ofrece el Instituto da Lingua Galega es el *Tesouro do léxico patrimonial galego e portugués* (*TLPGP*). Esta investigación inició su andadura el año 2009 bajo la coordinación de Rosario Álvarez Blanco y puede consultarse en la web desde 2014[8]. En la actualidad colaboran con el *TLPGP* lingüistas de veinte universidades (gallegas, portuguesas y brasileñas), pues el objetivo es compilar las obras lexicográficas dialectales, los datos de los atlas lingüísticos y las monografías etnolingüísticas tanto del gallego como del portugués (europeo y brasileño), con la finalidad de construir un portal unificado del léxico patrimonial. Como se indica en los créditos del proyecto, la búsqueda puede realizarse por el lema o por las diferentes variantes, aunque también incorpora un sistema avanzado que recupera los materiales por localización, campos semánticos (temáticos) y categoría gramatical. Constituye una novedad importante la representación cartográfica de cada una de las formas integradas en el corpus (a modo de atlas lingüístico)[9], así como la posibilidad de asociar a cada lema una serie de materiales etnográficos (dibujos y fotografías) extraídos asimismo de las obras indexadas. Se trata de una investigación abierta a nuevas incorporaciones: en la actualidad está integrada por 172 obras (fechadas entre 1935 y 2017), con un total de 56 811 lemas. Debe considerarse, sin duda alguna, como uno de los proyectos más sólidos y ambiciosos del panorama gallego y portugués que ha sabido superar fronteras y aunar las posibilidades que ofrecen las nuevas tecnologías con el rescate de todo ese riquísimo fondo patrimonial que aparece descrito en monografías de difícil acceso, poco conocidas por los investigadores o incluso inéditas.

En cuanto al catalán, el hecho de contar con el *Diccionari Català-Valencià-Balear* (1930–1962) (*DCVB*), de Antoni Maria Alcover y Francesc de B. Moll, quizás ha minimizado la necesidad de elaborar corpus lexicográficos similares a los del portugués, del gallego o del español, si bien, como ha señalado Colón (2014: 119), el Alcover-Moll es una obra que no debe considerarse cerrada y que debe continuar actualizándose. Para ello se han planteado algunos proyectos importantes en lo que atañe a la digitalización de diccionarios, vocabularios o glosarios, entre los que cabe destacar el «Portal de léxicos y gramáticas dialectales del catalán del siglo XIX (*LEXDIALGRAM*)»[10], coordinado por Maria-Pilar

8 <http://ilg.usc.es/Tesouro/>.

9 La estructura de la base de datos, en realidad, es mucho más amplia, ya que contiene diez campos: variante, transcripción fonética, categoría gramatical, lema, ejemplos, remisiones, definiciones, clasificación semántica, referencia geográfica e imágenes o dibujos.

10 <http://www.ub.edu/lexdialgram/index.jsp?item=presentacio&idx=0>. Vid. Perea 2015.

Perea, y el *Tresor lexicogràfic valencià (1543–1915)* (*TLV*), dirigido por Jordi Colomina (Guardiola i Savall 2008: 107).

Uno de los objetivos del *LEXDIALGRAM* ha sido poner en red un conjunto de veintidós obras de lexicografía dialectal redactadas entre 1739 (el *Breve Diccionario Valenciano-Castellano*, de Carlos Ros) y 1909 (el *Vocabulario valenciano-castellano en secciones*, de Joaquim Martí i Gadea). En total, se han incorporado a la base de datos unos 40 000 lemas que contenían las obras de referencia: las búsquedas pueden realizarse por cada texto individualmente o por el conjunto y los resultados aparecen reflejados en mapas dialectales, según las áreas en las que se ha documentado la voz (valenciano, mallorquín, menorquín, alguerés o rosellonés). El portal ofrece, además, un novedoso apartado dedicado a «estratigrafía dialectal» en el que «Es pretén de representar l'evolució diatòpica i cronològica d'una mostra de seixanta paraules catalanes», combinando simultáneamente los ejes temporal y diatópico[11].

El *Tresor lexicogràfic valencià* (1543–1915) (*TLV*) constituye otra de las apuestas de la lexicografía dialectal, surgido en la Universidad de Alicante con el aval de la Acadèmia Valenciana de la Llengua. Se trata de un proyecto en curso, formado por un corpus de 38 repertorios léxicos valencianos, publicados o inéditos, entre los que se encuentran el *Vocabulario del humanista* (1569) de Juan Lorenzo Palmireno o el *Vocabulari monosilábich valenciá-castellá* (1915) de Joaquim Martí i Gadea. «La raó principal que ha empés a agrupar tot aquest volum de lèxic ha estat la relació de dependencia de fonts que s'estableix entre totes aquestes obres» (Guardiola i Savall 2017: 433) y su objetivo es crear una gran base de datos de la lexicografía valenciana que muestre la evolución del léxico desde sus inicios hasta la publicación del *DCVB*.

Una obra bien distinta, aunque se asemeja en su concepción al *DDGM* gallego, es el «Glossari de Glossaris»[12]. A partir de la colección «Els Nostres Clàssics»,

11 «El projecte s'inicia en el període més primerenc del català escrit i mostra la ubicació geogràfica de les paraules representades, la seva forma gràfica i la possible progressió semàntica que han experimentat amb el pas del temps fins arribar als nostres dies. A diferència dels atles sincrònics convencionals, que utilitzen un sol l'espai com a eix, aquesta metodologia permet de fer servir simultàniament dos eixos: l'espai i el temps. La distribució de cada paraula en el mapa es representa en forma de capes, les quals representen el canvi lingüístic que el mot ha experimentat tant des d'una perspectiva temporal com des d'una perspectiva espacial» (<http://www.ub.edu/lexdialgram/estratigrafia/html/pagina1.html>). Es lo que Germà Colón y Maria-Pilar Perea han denominado «visualización dinámica del léxico» o «cronoestratigrafía dialectal».

12 <http://www.glossaris.net/advSearch.php>. Vid. Colón (2014: 123–124).

creada en 1924 con la finalidad de publicar ediciones críticas de obras catalanas desde sus inicios hasta 1800, Germà Colón ha dirigido desde 2007 un portal donde se funden todos los vocabularios que acompañan a estas ediciones y en los que se da cuenta de las particularidades dialectales de los textos, de los términos de especialidad, de los arcaísmos o de las primeras documentaciones de las voces catalanas. Es una base de datos que se actualiza periódicamente con los términos que aparecen en los nuevos volúmenes de la colección, que continúa siendo uno de los referentes de la filología catalana y que contiene algo más de 150 títulos. «El resultat serà un gran repertori lèxic de la llengua antiga», se indica en la presentación de la web, con la clara pretensión de conocer con mayor profundidad el recorrido histórico del vocabulario catalán[13].

3 Tradición española

A partir del *DRAE* 1884, la Academia incorporó a la entrada *tesoro* la acepción figurada[14] de «Nombre dado á ciertos diccionarios ó catálogos de palabras, con definiciones ó noticias útiles ó curiosas». Con esta definición impropia se refería la RAE, como muy bien especificaron unos años más tarde Elías Zerolo *et al*. en su *Diccionario* (de 1895), al título que Sebastián de Covarrubias había dado a su obra en 1611: *Tesoro de la lengua castellana o española*[15], así como al contenido de este diccionario del que Francisco de Quevedo decía que era «obra grande, y de erudición desaliñada»[16]. A partir de la decimoquinta edición del *Diccionario de la lengua española* (*DRAE* 1925), la definición de la palabra *tesoro* quedaría

13 Para el aragonés, en 1999 Nagore Laín publicó el *Endize de bocables de l'aragonés seguntes os repertorios lesicos de lugars y redoladas de l'Alto Aragon*, una especie de tesoro léxico basado en diccionarios, vocabularios y glosarios. Existe asimismo un proyecto de *Tesoro léxico del caló* (Buzek 2008) que incluiría el léxico romaní de los gitanos españoles, retomando otro proyecto anterior sobre *El léxico caló documentado*, de Gómez Alfaro (1997).

14 Perdería esta marcación en la edición del *DRAE* de 2001.

15 El sentido que tiene el término en francés había sido recogido por el *Diccionario de Autoridades* (s. v. *thesauro*) y por el *DRAE* desde su edición de 1780 (s. v. *tesauro*). De ahí los títulos del diccionario de César Oudin (*Tesoro de las dos lenguas francesa y española*, de 1607) o el de Girolamo Vittori (*Tesoro de las tres lenguas francesa, italiana y española*, de 1609).

16 Como es de todos conocido, Quevedo fue implacable en sus comentarios contra su coetáneo Covarrubias y su obra: «En el origen de ella [de la lengua] han hablado algunos linajudos de vocablos, que desentierran los huessos á las voces; cosa mas entretenida que demostrada; y dicen, que averiguan lo que inventan» (1626 [1699]: 506).

modificada en parte: «Nombre dado por sus autores a ciertos diccionarios, catálogos o antologías», añadiéndole ese valor de 'antología o colección' con el que la empleó a partir de esos mismos años veinte Gili Gaya. Es ese sentido el que se ha mantenido en la lexicografía española, especialmente en las tres últimas décadas, mientras que en otras tradiciones lingüísticas el nombre de *tesoro* o *tesauro* ha tenido otras aplicaciones, tal como resumió hace unos años Gregorio Salvador:

> Es conocida la variedad de significados con que el término *tesoro* suele utilizarse en lexicografía. Así bautizó Covarrubias a nuestro primer diccionario monolingüe; hay quien entiende la voz, pensándola desde el francés, como equivalente de diccionario histórico, completo y exhaustivo, y no falta quien, calcando el significado inglés, suponga que un tesoro no es otra cosa que un diccionario ideológico o incluso de sinónimos, porque el *Thesaurus* de Roget ha delimitado, por extensión, el valor de esa voz en aquella lengua (Salvador 1992: 11).

La labor que en su día emprendió Gili Gaya ha tenido dos continuadores, el *Nuevo Tesoro Lexicográfico del español (s. XIV-1726)* (*NTLE*), de Nieto Jiménez y Alvar Ezquerra (publicado en once volúmenes, en 2007), y el *Nuevo Tesoro Lexicográfico de la Lengua Española* (*NTLLE*), de la RAE, editado en un principio en DVD (2001) e incorporado pronto a los materiales de la web de la Academia, en un primer momento solo con los diccionarios de la institución y después con todo el material recogido en los dos DVD, esto es, las ediciones académicas y los diccionarios no académicos[17].

Según señalan los autores del *NTLE*, resultaba fundamental disponer de un «*corpus glossariorum* como nunca ha tenido la lengua española» (2007: ix)[18]. Adelantaron el marco cronológico del planteamiento de Gili Gaya hasta los inicios del siglo XIV con la inclusión de los *Glosarios latino-españoles de la Edad Media*, hasta llegar al capítulo que Benito Feijoo dedicó a «Algunas observaciones sobre la introducción de voces nuevas en nuestro idioma» en su *Theatro Crítico Universal*, de 1726, obra coetánea a la publicación del primer volumen del *Diccionario*

17 La selección se impone en un trabajo de estas características ya que la proliferación de diccionarios en el último siglo ha sido constante, a lo que hay que sumar todo el registro lexicográfico de las etapas anteriores. La *Biblioteca Virtual de la Filología Española* (*BVFE*) ofrece el enlace de 4570 obras o ediciones de estos textos que se encuentran disponibles en la red, entre diccionarios, vocabularios, glosarios, plantas de diccionarios, tratados de lexicografía, etc.

18 El inventario incluye los materiales de otra obra anterior realizada también en forma de tesoro y publicada en 2002 por Nieto Jiménez, el *Tesoro lexicográfico del español marinero anterior a 1726*.

de Autoridades y, por tanto, al inicio de la tradición académica[19]. Frente a las 73 referencias que barajaba Gili Gaya en su *Tesoro* (de las que solamente 60 aparecen efectivamente citadas), el *NTLE* duplicó las fuentes hasta completar los 145 textos que forman la nómina total de obras indexadas. La fiabilidad de la investigación y, por tanto, su valor cualitativo ha sido objeto de un delicado y detallado trabajo filológico que ha llevado a los autores a optar por las primeras ediciones y por volcar íntegramente los repertorios en el centón lexicográfico, con el fin de reconocer deudas y filiaciones. Ese afán de exhaustividad también se ha logrado integrando los materiales que aportaban pequeños glosarios insertos en otro tipo de obras[20]. Por otro lado, la heterogeneidad de un corpus tan complejo se ha suplido con una microestructura muy simple, con la lematización de todas las variantes y con un sencillo sistema de referencias internas. A pesar de que su edición se haya realizado en un formato tradicional, constituye una base de datos perfectamente estructurada de toda la información metalexicográfica que compila y su organización sistemática permitiría un volcado rápido en la red y, por tanto, su conversión en un producto digital.

La edición del *NTLLE* viene a completar el marco cronológico que tanto Gili Gaya como Nieto Jiménez y Alvar Ezquerra habían establecido para sus respectivos tesoros, al integrar, aparte de los principales diccionarios preacadémicos, casi toda la lexicografía institucional hasta 1992[21] y una parte de los repertorios no académicos compuestos entre 1729 (el manuscrito inédito de Juan Francisco de Ayala Manrique) y 1918 (el *Diccionario general y técnico hispano-americano*, de Manuel Rodríguez Navas y Carrasco)[22]. En total, se ofrece la imagen digital (o

19 Insertan asimismo los materiales del *Tesoro de la Lengua Castellana* de Juan Francisco de Ayala Manrique, una obra manuscrita inacabada, que se inició en 1693 pero que todavía no estaba concluida en 1726.

20 Se han tenido en cuenta asimismo los glosarios bilingües, al menos los anteriores a 1500, con el fin de completar las escasas referencias que existen del cuatrocientos.

21 Faltan los materiales editados del segundo proyecto de *Diccionario Histórico de la Lengua Española* (*DHLE 1960–1996*), que ahora se pueden consultar en la web de la RAE, con una interfaz que permite acceder tanto a los lemas como al texto. Esa doble consulta como diccionario y como base de datos se puede realizar asimismo en la aplicación que se ha implementado en la misma web para el primer proyecto de *Diccionario Histórico* (*DHLE 1933–1936*).

22 Unos años antes, en 1998, Pedro Álvarez de Miranda había coordinado la publicación en CD-ROM de un conjunto de diccionarios fundamentales para la historiografía, con un marco cronológico que cubría desde 1495 hasta 1847: *Lexicografía Española Peninsular. Diccionarios Clásicos (I y II)*. La primera parte ofrecía la reproducción facsímil del *Vocabulario Español-Latino* de Nebrija y el llamado *Diccionario de Autoridades*; la

facsimilar) de 66 diccionarios, con una herramienta de búsqueda sencilla (con comodines y operadores lógicos) mediante la cual se accede a los lemas de todo el conjunto o de los repertorios que se seleccionen[23]. Al no haber sido transcritos los textos en formato digital, lo que se recuperan son imágenes de la columna donde figura la voz analizada, perdiendo así la dinamicidad que hubiera adquirido un corpus totalmente digitalizado. A pesar de sus limitaciones, se trata de una biblioteca lexicográfica fundamental puesta a disposición de todos, que se ha hecho imprescindible para cualquier investigación léxica. Así todo, solo es un primer paso de lo que una herramienta digital totalmente desarrollada y unos materiales debidamente etiquetados podrían significar para el análisis de la riqueza que atesoran las miles de páginas que contienen estos diccionarios[24].

4 Tesoros dialectales diferenciales

La mayor parte de los diccionarios generales contiene en alguna medida muestras del léxico diatópicamente marcado, pero esa selección que ofrecen estos repertorios no constituye sino una mínima parte del acervo léxico que conforma la lengua en la totalidad de sus variedades y que contribuye a caracterizar, junto a otros aspectos como la entonación, la fonética, los giros gramaticales o determinados rasgos pragmáticos, la modalidad de cada región o de cada nación (del español en América) y, en muchos casos, la ascendencia concreta de sus hablantes. García Márquez afirmaba que Simón Bolívar «hablaba con la cadencia y dicción de las islas Canarias, y con las formas cultas del dialecto de Madrid» (1993: 83). A través de los materiales del *Diccionario de madrileñismos* de Alvar Ezquerra (2011), del *Atlas Dialectal de Madrid* (*ADiM*) y de los estudios que se están realizando a partir del *Atlas Lingüístico Diacrónico e Interactivo de la*

segunda entrega incorporaba el *Tesoro* de Covarrubias y los diccionarios de Terreros, Salvá y Domínguez.

23 Como se indica en la presentación de la web, se han respetado las características gráficas peculiares propias de cada diccionario, sin someterlas a ningún proceso de regularización o unificación, lo que multiplica innecesariamente la búsqueda al no tener en cuenta las variantes de un mismo lema. Presenta también algunos problemas en la lematización de colocaciones y locuciones, si bien ello puede ser debido a la ordenación interna de algunos de los textos que componen el corpus.

24 Un primer desarrollo de esa herramienta la ofrece otra aplicación elaborada para el *Nuevo Diccionario Histórico del Español*, disponible asimismo en la página web de la RAE: el *Mapa de diccionarios*. El sistema de consulta permite comparar de manera simultánea los más de 100 000 lemas correspondientes a seis de las ediciones del diccionario académico (las de 1780, 1817, 1884, 1925, 1992 y 2001).

Comunidad de Madrid (*ALDICAM*)[25] será posible confirmar la validez de las palabras del gran escritor colombiano. Las voces procedentes del acervo lingüístico regional canario delatan, en un ejercicio de arqueología lingüística, el origen de otro autor del que apenas se conocen datos biográficos, fray Thomas de Padilla, humanista de la época de Felipe II y traductor de la *Historia de las cosas de Etiopía* (1557). Incluso la obra de un novelista tan cosmopolita como Pérez Galdós, que analizó y recreó tan bien en su obra el habla de los ambientes madrileños, refleja inconscientemente el bagaje léxico que formaba parte del entorno familiar de la infancia del autor[26]. La recopilación de todas estas voces singulares y su marcación constituye uno de los objetivos de la lexicografía diferencial, que también ha encontrado en la planta del *tesoro* uno de los cauces más relevantes de recopilación y difusión de esta cantidad inmensa de datos, fundamentales para la historia de la lengua, la cultura y la etnografía.

El tesoro geolectal se concibe con un sentido práctico y puramente descriptivo, como un recurso que reúne una documentación metalexicográfica excepcional que, en buena parte, no ha tenido más que una difusión puramente local (Corrales/Corbella 2009: 285). Cada una de sus entradas se convierte en una monografía con registros perfectamente datados y autorizados, con reflexiones importantes sobre los elementos que permiten catalogar una voz como diferencial, los rasgos contrastivos que avalan ese empleo particular, su origen, ámbito de uso y vitalidad. Los materiales del *tesoro* ponen al descubierto, asimismo, aparte de las deudas más o menos manifiestas de unos lexicógrafos con respecto a otros, los errores en la caracterización geolectal de un término, lapsus y erratas en la transmisión escrita, los hápax o la ausencia de muchos vocablos y acepciones que, por un empleo exclusivamente oral, por pertenecer a determinadas terminologías populares o porque fueron exclusivos de una época, nunca han sido objeto de estudio o no han entrado a formar parte de la tradición lexicográfica. Puede resultar contradictorio que el carácter exhaustivo del *tesoro* sirva para mostrar las lagunas existentes, pero quizás en ello radica uno de sus máximos logros. En realidad, todo *tesoro* es un reconocimiento a la tradición lexicográfica y a los lexicógrafos que la hicieron posible, pero también constituye, como

25 Véanse los trabajos de Torrens y de Sánchez-Prieto Borja/Vázquez Balonga en este mismo volumen.

26 Como la locución adverbial «a la pela», de la que el *Corpus diacrónico del español* (*CORDE*) recoge tres únicos ejemplos, precisamente en obras de este canario universal. Ninguno de estos testimonios se corresponde con los significados que ofrece el *Diccionario de la lengua española* (*DLE*: s. v. *pela*), sino con el registro canario de este portuguesismo.

señaló Alvar Ezquerra a propósito del *Tesoro léxico de las hablas andaluzas*, «un punto de partida para nuevas tareas, sabedores de que nos queda todavía muchísimo por reunir, aunque no es poco lo que ya tenemos» (*TLHA* 2000: 9)[27].

4.1 Tesoros de ámbito regional del español de España

El concepto de *tesoro* de ámbito restringido constituye una novedad en la lexicografía española y, como hemos visto, su desarrollo también comienza a ofrecer logros relevantes en las investigaciones gallegas y catalanas. Según Ahumada (2000: 25), el más antiguo de estos repertorios dialectales es el *Diccionario de los Bables de Asturias* (1989), de Jesús Neira Martínez y Mª del Rosario Piñeiro. Sin embargo, no se trata de un «tesoro lexicográfico», sino de un diccionario bidireccional castellano-bable y bable-castellano que ha utilizado las referencias diccionarísticas para autorizar las definiciones y localizaciones propuestas. Entre las veintinueve fuentes utilizadas, diez textos corresponden al bable central, siete al bable oriental y seis al bable occidental, a los que habría que añadir seis obras más relativas a la flora y la fauna de la región.

Exceptuando este precedente, y centrándonos en las modalidades lingüísticas del español europeo, el primer diccionario de diccionarios que se publicó fue, en realidad, el *Tesoro lexicográfico del español de Canarias* (*TLEC*), que ha conocido dos ediciones, una de 1992 y otra, aumentada, de 1996. En opinión de Ahumada (2000: 25), «es hasta ahora el más completo de los tesoros publicados», con un índice documental que abarca 293 textos. Sus datos se nutren de dos fuentes principales: los diccionarios dialectales (vocabularios, léxicos, glosarios, nomenclaturas populares) y las monografías lingüísticas sobre esas voces consideradas diferenciales, entre las que destacan el *Atlas lingüístico y etnográfico de las Islas Canarias* (*ALEICan*) y las encuestas realizadas en el archipiélago para el *Léxico de los marineros peninsulares* (*LMP*), cuyos materiales geolectales fueron ordenados alfabéticamente y volcados al catálogo en su integridad. Concebido como un paso previo y como fuente de referencia para iniciar la redacción de un diccionario de uso del habla de las islas, la cantidad de lemas recolectados (incluyendo los procedentes de algunos textos manuscritos), su ordenación cronológica y la exhaustividad de sus datos dieron entidad a la obra en sí misma, ya que permitía cubrir un espectro cronológico bastante amplio (desde 1799 hasta 1996). El *TLEC* supuso, al mismo tiempo, la puesta al día de los materiales metalexicográficos para contrastarlos con las documentaciones efectivas en textos

27 En el apartado siguiente presentamos una actualización de los datos recopilados en Corrales/Corbella (2009) y en Corbella (2016).

escritos y orales y así advertir el uso diferencial de los canarismos en la actualidad y a través del tiempo, material que conformaría el *Diccionario ejemplificado de canarismos* (*DECan* 2009) y el *Diccionario histórico del español de Canarias* (*DHECan* 2001 y 2013[2]).

A partir del *TLEC*, la edición de este tipo de inventarios del léxico dialectal se ha visto incrementada, tanto en el español europeo como en el americano (Corrales/Corbella 2009: 292–294). En lo que atañe al español peninsular, el área occidental es la que cuenta con un mayor número de recopilaciones recientes. En 1993 se publicó el *Diccionario de las hablas leonesas (León, Salamanca, Zamora)*, de Miguélez[28]: este «primer intento de recapitulación de monografías dialectales leonesas» partió del análisis de veintinueve glosarios, el primero de ellos de 1903. Se trata, no obstante, de un texto intermedio entre el tesoro y el diccionario, ya que el compilador no se limita a transcribir cada uno de los registros, sino que reelabora los materiales, elimina ambigüedades y utiliza un léxico no marcado y de nivel medio-culto en las definiciones[29]. Un año más tarde, en 1994, se editaba el *Léxico leonés* de Díez Suárez, que recogía el léxico de un total de doce monografías dialectales, reordenando todo el corpus en nueve campos onomasiológicos. Pero sin duda la obra lexicográfica más relevante de esta región es el *Léxico del leonés actual* (*LLA*) de Janick Le Men, un repertorio riquísimo, publicado en cuatro tomos entre 2002 y 2012, que tuvo como base la tesis doctoral de su autora. El objetivo era realizar «una recopilación exhaustiva de las voces incluidas en todos los estudios lexicográficos, publicados e inéditos, sobre la provincia de León» (2002: 13), con un total de 215 fuentes indexadas, de las que dieciocho permanecían inéditas, todas ellas pertenecientes a un periodo que abarca desde 1861 a 2000. Aunque ninguno de estos tres diccionarios leoneses reproduce la microestructura de un tesoro tal como lo concibió Gili Gaya, en cuanto que no es posible realizar un seguimiento tópico-cronológico de la aparición de cada voz en los vocabularios y glosarios dialectales, sí que se ofrecen, al menos en el *LLA*, las referencias concretas que han servido de testimonio para documentar lexicográficamente el uso de una palabra en una determinada localidad. Ha primado, por tanto, el criterio geolectal en detrimento del cronológico, y se han reelaborado los materiales lexicográficos (no se han limitado a reproducirlos tal como aparecen en las obras de referencia).

28 Se volvió a editar en 1998.

29 Aunque «faltan muchas palabras utilizadas en las hablas regionales», afirma Miguélez (1998[2]: xii).

La publicación del *Tesoro léxico de las hablas andaluzas* (*TLHA* 2000), de Alvar Ezquerra, supuso un avance muy significativo no solo para el vocabulario de esta región, sino también por la importancia que el conocimiento de esta modalidad puede tener en la historia de la conformación del español atlántico. Como se indica en el «Prólogo», salvo la recopilación que en 1933 había realizado Alcalá Venceslada para su *Vocabulario Andaluz*, «Las demás informaciones con que contábamos sobre nuestro vocabulario andaban desperdigadas, muchas veces en obras insospechadas, y sin sistematizar» (2000: 9). Para la elaboración del *TLHA* se empleó una base de datos muy sencilla, cuando las nuevas tecnologías apenas acababan de llegar a los estudios filológicos (vid. Alvar Ezquerra 2004: 41–52). En la versión impresa, el conjunto léxico está compuesto por 146 textos, desde 1852 hasta 1999, entre los que destacan los registros del *Atlas Lingüístico y Etnográfico de Andalucía* (*ALEA*) y los resultados de las encuestas andaluzas realizadas para el *Léxico de los Marineros peninsulares* (*LMP*)[30].

Siguiendo una ordenación cronológica de edición, el siguiente compendio correspondió al *Tesoro léxico de las hablas riojanas* (*TLHR*), de Pastor Blanco, publicado en 2004. Como señala su autor, «Hasta ahora, las principales informaciones con que contaba nuestro vocabulario se hallaban en gran medida desperdigadas, a menudo en estudios misceláneos, faltos de rigor y sin sistematizar» (2004: 11). La base documental se formó a partir de las fuentes lexicográficas ya conocidas (los registros que proporciona el *DRAE*, la información léxica sobre las hablas riojanas que ofrece el *Atlas Lingüístico y Etnográfico de Aragón, Navarra y Rioja –ALEANR–* y las compilaciones realizadas por los dialectólogos de la región o por aficionados que, con su labor, dieron cuenta de un léxico tradicional que con el tiempo se ha ido perdiendo) y los testimonios inéditos, en total 73 textos datados entre 1807 y 2004. A todo ello el autor añade su propia labor de campo «por los diferentes valles riojanos que aún quedaban por estudiar de manera exhaustiva» (2004: 13). Como sucedía con los diccionarios leoneses, no se trata de un tesoro acumulativo sino de un diccionario autorizado por las

30 Un tesoro léxico distinto, que parte de los regionalismos (en este caso, andalucismos) que aparecen recogidos en las fuentes lexicográficas generales (y no en los repertorios diferenciales), es el que publicó Moreno Moreno con el título de *Léxico histórico andaluz. I. Período clásico* (2007). Su objetivo consistió en sistematizar las referencias geolectales contenidas en repertorios como los de Nebrija, Guadix, Francisco del Rosal, Covarrubias, Minsheu o Franciosini, entre otros lexicógrafos renacentistas. Un estudio similar fue realizado por Corrales/Corbella (2015) para los dialectalismos canarios en los diccionarios no académicos (especialmente en los preacadémicos y en los generales de los siglos XIX y XX).

Tabla 1: Tesoros lexicográficos y léxicos de ámbito regional del español europeo

Ámbito	Publicación	Fecha de edición	Número de obras indexadas	Materiales procedentes de los atlas lingüísticos	Período cronológico
Canarias	*TLEC*	1992 y 1996[2]	293	Sí	1799–1996
Andalucía	*TLHA*	2000	146	Sí	1852–1999
León	*LLA*	2002–2012	215	No	1861–2000
La Rioja	*TLHR*	2004	73	Sí	1807–2011
Madrid	*Diccionario de madrileñismos*	2011 y 2011[2]	191	No	1861–2011

fuentes lexicográficas de las que se nutre, cuyas referencias aparecen ordenadas por la diatopía que abarcan, desde las más generales de toda la región hasta las provinciales y, por último, las de ámbito local[31].

En 2011 se publicó el último compendio que ha visto la luz en formato papel, el *Diccionario de madrileñismos. Voces patrimoniales y populares de la Comunidad de Madrid*, de Alvar Ezquerra. En la segunda edición de esta obra, editada apenas unos meses después en el mismo año 2011, la base de datos se formó a partir del vaciado de 191 fuentes documentales (la más antigua, de 1861, mientras que las más recientes fueron coetáneas a la elaboración del diccionario), aparte de los datos extraídos de un cuestionario específico realizado con el fin de cubrir algunos ámbitos de la actividad cotidiana o algunas zonas con escasa o nula información (Alvar Ezquerra 2011: xiv). El compendio recoge en total 7815 entradas y 11 647 acepciones. Las definiciones están tomadas, en su mayoría, de las obras que han servido de base para elaborar el repertorio léxico, aunque en otras ocasiones pertenecen al autor del *Diccionario*. En todo caso, siempre se indica la fuente que ha servido de punto de partida, así como las localidades donde se registra cada significado. El resultado final de esta investigación no es un tesoro lexicográfico dialectal propiamente dicho, sino un diccionario diferencial contrastivo basado, mayoritariamente, en fuentes lexicográficas.

31 «A partir de la información que aquí se ofrece —señala el autor—, espero haber dado una visión bastante aproximada y cabal de lo que es el léxico vivo característico de esta tierra. Confío en que, tras su publicación, este trabajo contribuya a impulsar nuevos estudios sobre el léxico riojano, pues uno de sus principales valores es mostrar cómo numerosas palabras poco conocidas en la lengua general tienen aquí notable arraigo» (2004: 15).

Cuatro proyectos más hay que sumar a estas publicaciones pioneras. El primero, el *Tesoro lexicográfico del español de Murcia*, fue planteado hacia 1995 por José Muñoz Garrigós y José Perona, con la pretensión de elaborar «un libro de conjunto que sea capaz de ofrecer todos los datos geográficos, históricos y sociales sobre las palabras que han vivido y viven en esta tierra» (1996: 98–99). El fallecimiento de Muñoz Garrigós en 1997 y el de Perona en 2009 dejó este trabajo truncado a falta de que los demás miembros del grupo de investigación retomen los ricos materiales que, sin duda, ya estaban recopilados.

El segundo proyecto parte del grupo ARALEX que actualmente elabora el *Diccionario diferencial del español de Aragón* (*DDEAR*) en la Universidad de Zaragoza. Según señalan sus autores, la recopilación se está realizando a partir de materiales que figuran dispersos en algo más de un centenar de monografías lingüísticas y repertorios lexicográficos aparecidos desde 1950 hasta la actualidad (Arnal *et al*. 2012: 82–83), entre los cuales figura el léxico de esta comunidad registrado en el *Atlas Lingüístico y Etnográfico de Aragón, Navarra y Rioja* (*ALEANR*). La edición de los materiales metalexicográficos acopiados por este equipo, más el añadido de los glosarios anteriores a la década de los cincuenta del siglo XX, sin duda redundaría en un mayor conocimiento de la historia léxica de esta región.

Pero quizá el trabajo que se encuentra más avanzado es el correspondiente al *Tesoro léxico de las hablas extremeñas*. La indexación de los materiales se inició en 2004 e incluye el vaciado de 136 obras (González Salgado 2012: 167–170), con un marco temporal de un siglo (desde 1909 hasta 2008), teniendo en cuenta las encuestas realizadas durante la elaboración del *Atlas lingüístico de la Península Ibérica* (*ALPI*)[32] y del *Corpus dialectal de Extremadura*.

Una última investigación de la que tenemos constancia es el *Tesoro léxico de la frontera hispano-portuguesa* (del grupo FRONTESPO), que incluirá los portuguesismos, castellanismos y dialectalismos propios de esa zona de confluencia del occidente peninsular. La edición será totalmente digital, aunque los objetivos marcados son similares al resto de los tesoros regionales: «rellenar un hueco en la investigación lexicográfica con la incorporación en una sola obra de los materiales que están dispersos por multitud de estudios, y conseguir con ello una visión de conjunto del léxico compartido, en una demostración más de que las palabras no conocen fronteras» (González Salgado 2017: 124).

El análisis del léxico regional contrastado de este modo aparece refrendado en estas obras de las que han surgido o se derivarán, a su vez, diccionarios

32 Vid., en el presente volumen, el capítulo de García Mouton.

sincrónicos, como los realizados para el canario, el proyectado para Aragón o el que en su día esbozó Alvar Ezquerra para Andalucía (2004: 55). Además, la catalogación de los materiales de los atlas lingüísticos, quizá los repertorios léxicos más amplios de los que dispone el dialectólogo (Corrales/Corbella 2004: 1219–1220), perfectamente datados y localizados en el amplio espacio geográfico, responde a una de las demandas clásicas de la lexicografía y facilitará la tarea de inclusión de estos registros en los diccionarios generales o su eliminación, en caso de que su extensión no esté confirmada en todo el ámbito regional. Por último, aunque no menos importante, el tesoro geolectal regional puede facilitar la labor de la lexicografía española, sincrónica o histórica, y un conocimiento más ajustado de la historia misma del léxico del español europeo.

4.2 Tesoros del español en América

El estudio de Salvador Rosa sobre el *Diccionario de Autoridades* reveló la presencia en aquel primer diccionario académico de 129 voces procedentes de América (1985: 126)[33]. Y los materiales del *NTLE* y del *NTLLE* muestran la inclusión de americanismos en los repertorios lexicográficos de Nebrija (¿1495?), Alcalá (1505), Junius (1567), Las Casas (1570), Percivale (1591), Guadix (1593), Decimator (1596), Hornkens (1599), Minsheu (1599), Del Rosal (1601), Palet (1604), Oudin (1607), Vittori (1609), Covarrubias (1611), Franciosini (1620), Mez (1670), Stevens (1706) o Bluteau (1721), entre otros[34]. Sin embargo, el *NTLE* no recoge entre sus fuentes ningún diccionario exclusivamente americano, como el que se ha considerado el primer glosario de americanismos, una lista de 18 palabras que Pedro Antonio Fernández de Castro y Andrade (VII Conde de Lemos) incluyó en la *Descripción de la provincia de los Quixos* (1608), en el actual Ecuador (Quesada Pacheco 2002: 19). Tampoco aparece en este panléxico la «Tabla para la inteligencia de algunos vocablos de esta historia», añadida por fray Pedro Simón en sus *Noticias historiales de las conquistas de Tierra Firme en las Indias occidentales* (1627) (Quesada Pacheco 2002: 20). Ni se tiene en cuenta el glosario que, a finales del siglo XVII, aparece como anexo del poema *Alteraciones del Dariel*, de Juan Francisco de Páramo y Cepeda (Haensch/Omeñaca 2004[2]: 303). Por su parte, aunque el *NTLLE* de la RAE registra algunas muestras magníficas de la lexicografía decimonónica que reivindicaban la presencia de voces del Nuevo Mundo en los diccionarios generales, no incorporó, entre otros, el *Diccionario de*

33 De ese total, 101 aparecen catalogadas como generales del continente; 13 se recogen como propias del Perú y 15 como exclusivas de México.

34 Las referencias bibliográficas y sus dataciones están tomadas del *NTLE*.

voces americanas (¿1751–1777?) atribuido al jurista panameño Manuel José de Ayala, ni el *Vocabulario de voces provinciales de América* que Antonio de Alcedo añadió como apéndice de su *Diccionario geográfico histórico de las Indias Occidentales o América* (1789), ni otros diccionarios posteriores que iniciaron con aplomo la lexicografía del otro lado del Atlántico[35] y que han tenido una amplia recepción, como el *Diccionario provincial casi razonado de Vozes y frases cubanas* de Esteban Pichardo[36].

Señalaba con razón Günther Haensch que en la lexicografía hispanoamericana no son precisos «sueños quijotescos, pero sí un trabajo de base que no sea desmesurado en sus exigencias» (1994: 54). La necesidad de compilar en un único texto el corpus de palabras o el contenido extenso de los diccionarios y vocabularios de cada nación americana[37], generalmente con la finalidad de elaborar los respectivos diccionarios académicos, ha sido una constante en el quehacer lexicográfico de los últimos cuatro lustros. Al mismo tiempo era necesario plantear una recopilación exhaustiva de las voces que aparecen en los textos bilingües con las lenguas amerindias, desde las obras de los primeros misioneros hasta el inicio de las independencias, investigación que se describe en el capítulo que Esther Hernández dedica en este libro al *Tesoro léxico de los americanismos contenidos en los vocabularios hispano-amerindios coloniales (1550–1800)* (*TELEAM*)[38].

---

35 Las primeras recopilaciones surgieron a la par que se desarrollaban las identidades nacionales en América y fueron producto, en cierta medida, de la necesidad de construir una tradición cultural propia.

36 Con el objeto de facilitar el acceso a las principales fuentes americanas, Günther Haensch publicó en 2000 un CD-ROM con los *Textos clásicos sobre la historia de la lexicografía del español en América*, en el que incluía la imagen digital de diecinueve diccionarios y vocabularios, desde la recopilación de fray Pedro Simón hasta los *Honduñerismos* de Alberto Membreño.

37 Vid. el volumen de obras publicadas en el siglo XX que recogen Haensch/Omeñaca (2004[2]: 301–326). Para las fuentes modernas, el *Diccionario de americanismos* (*DAMER* 2010: «Introducción») da cuenta de la existencia de la herramienta ARU ('lengua', en aimara), utilizada internamente por la Asociación de Academias de la Lengua Española (ASALE) para la elaboración y revisión del corpus de americanismos registrados tanto en el *DLE* como en el mismo *DAMER*. Permite la consulta de los lemas que reúnen «los casi 150 diccionarios de americanismos (generales y nacionales) publicados desde 1975 hasta la fecha, más algunos inéditos aún, todo en formato electrónico con excelentes y ricos índices que facilitan cualquier tipo de búsqueda».

38 Ya desde 1928, Amado Alonso escribía a Rodolfo Lenz que quería «publicar esos vocabularios y gramáticas, no al modo de [Julio] Platzmann, sino de manera que se dé a

El primer ejemplo de un lemario de léxicos emprendido por una de las academias americanas lo constituye el *Índice de mexicanismos registrados en 138 listas publicadas desde 1761*, que ha conocido tres ediciones (de 1997, 1998 y 2000). Se trata de un repertorio muy amplio, de 76 000 lemas, en el que «no se reproduce la mayor o menor información de cada autor para cada mexicanismo […]. Tampoco se da el número de la página, únicamente el número de la fuente (del 1 al 138) en la bibliografía» (2000: 7). La aplicación permite la ordenación de repertorios por antigüedad[39], número de registros y autor; recuperar las entradas de una única obra, así como la búsqueda por terminaciones, a modo de diccionario inverso, y el rastreo por las posibles variantes ortográficas (2000: 8). Uno de los resultados inmediatos de esta investigación fue la publicación, en 2001, del *Diccionario breve de mexicanismos*, de Gómez de Silva. También sirvió de punto de partida para la elaboración del *Diccionario de mexicanismos* de la Academia Mexicana de la Lengua (*DM* 2010)[40].

Concebida asimismo como herramienta auxiliar en la que se ofrece el término de entrada o la remisión a un vocablo tal como se encuentra en los vocabularios y glosarios regionales, la Academia Argentina inició en 1992 la digitalización del *Registro de Lexicografía Argentina* (*RLA*). La finalidad de este proyecto era disponer de un fichero lexicográfico representativo de los argentinismos recogidos en los diccionarios, así como en artículos y notas dispersas de difícil localización. A este glosario se incorporaron, además, las referencias a las fichas manuales conservadas en los ficheros lexicográficos y de autoridades de la misma Academia. En total, la versión en CD-ROM, publicada en el año 2000, recoge 92 288 registros procedentes de la indexación de 215 documentos, con una franja temporal que cubre desde 1845 hasta 1999[41]. Como se indica en el prólogo, la obra contiene

---

los lingüistas un material lo más abundante y lo menos farragoso posible» (Guitarte 1996: 82).

39 Los últimos vocabularios indexados corresponden a 1996: una breve selección de mexicanismos propios del ámbito familiar, comercial y escolar del Distrito Federal (de Esperanza Berumen de Cuesta), una recopilación de voces del Noroeste de México (Nuevo León, Coahuila y Tamaulipas, de Ricardo Elizondo) y el texto *Minucias del lenguaje* (de José G. Moreno de Alba). También se incorporó el facsímil publicado en 1997 de los *Proverbios mexicanos* (de Ignacio M. Altamirano), si bien esta obra fue compuesta originariamente en la segunda mitad del siglo XIX.

40 Concepción Company, en la «Introducción» a este último diccionario, afirma que el Índice «ha sido una fuente constante de consulta y orientación para nuestras labores» (*DM* 2000: xxii).

41 Desde los trabajos pioneros de Francisco Javier Muñiz (1845), Tobías Garzón (1910) o Lisandro Segovia (1911), hasta los diccionarios de argentinismos o de regionalismos de finales del siglo XX.

todas aquellas voces que «en algún momento algún lexicógrafo las consideró representativas de nuestro modo de hablar. De allí que un interés, y no menor, que puede hallarse en este *Registro* es su contribución al conocimiento de la historia de la lexicografía hispanoamericana, y de la argentina en particular». Las referencias del *RLA* han servido para autorizar las fuentes lexicográficas que avalan el empleo de cada uno de los lemas contenidos en el *Diccionario del habla de los argentinos* (*DiHA* 2003)[42].

Todas las academias poseen archivos lexicográficos junto a ficheros de datos léxicos, pero no siempre las han puesto a disposición del público general como ha sucedido con los registros argentinos y mexicanos. Así, por ejemplo, en el «Preámbulo» del *Diccionario del español del Uruguay* (*DEU*), de la Academia Nacional de Letras, se señala la existencia de una base de diccionarios y glosarios uruguayos (*DICUR*), cuyos datos «corresponden a la obra de los lingüistas que han trabajado con el léxico regional y a las compilaciones realizadas por la casi totalidad de lexicógrafos y aficionados a la lexicografía del Uruguay, además de monografías de similares características, realizadas por estudiantes adelantados de Lingüística o por participantes de los concursos que ha convocado la Academia sobre temas del habla» (*DEU* 2011: 15)[43].

En el año 2005 se publicó el que debe considerarse el modelo de tesoro para América, el *Tesoro lexicográfico del español de Puerto Rico* (*Tesoro.PR*), editado primero en edición en papel y, desde 2016, de acceso libre en la página web de la Academia Puertorriqueña de la Lengua Española. Elaborado inicialmente por María Vaquero y Amparo Morales, ofrece ordenados de manera cronológica los materiales íntegros de 19 333 lemas procedentes de 60 léxicos y vocabularios de la isla antillana, desde la obra de Íñigo Abbad y Lasierra, *Historia geográfica, civil y natural de la isla de San Juan Bautista de Puerto Rico* (de 1788), hasta la *Revisión del léxico gallístico* que Edwin Figueroa Berríos realizó en 2003, pasando por los textos clásicos de la lexicografía puertorriqueña, como las recopilaciones de Augusto Malaret, Tomás Navarro Tomás, Manuel Álvarez Nazario o las *Palabras de Puerto Rico* que María Vaquero había publicado en 1995 y en las que ofrecía todos los datos léxicos de las encuestas geolingüísticas. En la introducción del *Tesoro*, las autoras expusieron de forma muy clara su propósito:

42 El *DiHA* tuvo una segunda edición en 2008 y para 2017 estaba anunciada una tercera entrega.

43 Para un panorama de la periodización de la lexicografía en Uruguay, véase Coll (2013 y 2017).

> Ha sido lograr, en lo posible, que el usuario tenga la mínima dificultad a la hora de hacer las consultas y que pueda acercarse, mediante constantes remisiones cruzadas y dentro del formato alfabético que impone un diccionario, a la cantidad de relaciones diversas que puede haber entre las unidades léxicas recogidas. […] Esta enumeración histórica permite, además, rastrear la vida de cada vocablo: hay palabras constantes que mantienen su imperio a través del tiempo y las hay que se gastan o se retiran; otras cambian su significado, sin cambiar de piel, y muchas amplían o modifican sus valores semánticos con rasgos nuevos. Y si la enumeración histórica de los registros nos ayuda a datar olvidos paulatinos, también será útil para constatar las novedades. Un *Tesoro Lexicográfico* permite asediar el vocabulario desde muchos puntos de vista, y esta es la razón de que sea útil en distintos campos de trabajo (*Tesoro.PR* 2005: 13).

Como señalamos en Corrales/Corbella (2009: 294), un repertorio planteado de esta manera se convierte en un instrumento que facilita en buena medida la labor del lexicógrafo, los futuros trabajos contrastivos entre las distintas regiones del español meridional[44] y la elaboración de cualquier monografía léxica. No obstante, María Vaquero y Amparo Morales también fueron conscientes de las limitaciones de todo tesoro, sometido en su edición en papel a la catalogación alfabética, así como a la heterogeneidad de una obra de esta envergadura que necesitaría la aplicación de unos criterios uniformes en el tratamiento de las entradas y en la identificación de las relaciones mutuas. Decían Haensch/Omeñaca (2004²: 28) que todo diccionario (y el *tesoro* es, al fin y al cabo, un diccionario) suele ser «el resultado de una serie de compromisos entre el máximum deseable y el óptimum realizable». Esas posibilidades a las que apuntaban las autoras del *Tesoro* de Puerto Rico se han visto notablemente acrecentadas con el tratamiento informático de los materiales, tarea que se ha realizado en la Academia Puertorriqueña bajo la dirección de Maia Sherwood Droz:

> Ahora el *Tesoro* se encuentra con la lexicografía electrónica: con la página tesoro.pr, Puerto Rico da su primer paso en esta disciplina que se va constituyendo como el presente y en definitiva el futuro de los diccionarios. […] Las nuevas plataformas son el destino natural de los diccionarios: no solo salvan la limitación de espacio de la página impresa, sino que permiten establecer redes y relaciones entre palabras, significados y otros datos, y explorarlas de maneras dinámicas y noveles. Justamente, la base de datos y herramienta de consulta de tesoro.pr han sido construidas para permitir el máximo aprovechamiento de la información. Paralelamente, hemos dotado a la herramienta de

44 El cotejo de estas y otras recopilaciones léxicas hizo posible la edición del *Tesoro léxico canario-americano* (*TLCA* 2010), donde se analizan los viajes de ida y vuelta de muchas de las voces compartidas y la simbiosis surgida de las relaciones históricas del archipiélago canario con toda América, especialmente con el Caribe.

una interfaz sencilla, de manera que el usuario logre realizar búsquedas productivas de manera casi intuitiva (<https://tesoro.pr>).

Con un programa muy simple, aparte de la consulta de cada lema como en un diccionario tradicional (o como en un diccionario inverso, si lo que se pretende es la agrupación según las terminaciones), la búsqueda sencilla permite pasar de un elemento a otro por medio de hipervínculos. Pero al ser concebido como una base de datos relacional, el análisis de los metadatos (categoría gramatical, origen de la palabra, fuente, año, campo temático)[45] desvela también de manera inmediata múltiples conexiones entre las palabras. Además, al introducir todas las definiciones, el mismo tesoro se ha convertido en un corpus documental: «podemos buscar palabras presentes en ellas, lo cual amplía las búsquedas formales y abre las puertas a diversos análisis, incluidos estudios del diccionario como género textual». A todo ello se une la amplia recepción que la difusión en la red ha supuesto tanto para el patrimonio lingüístico puertorriqueño como para el estudio histórico del léxico americano en su conjunto.

En un futuro próximo, el español cubano y el español dominicano contarán con sus respectivos tesoros, lo que significará un avance cualitativo en el conocimiento de todo el léxico caribeño. Los trabajos para la realización del primero ya están concluidos, a falta de que sus resultados puedan consultarse en la web —por ahora el acceso es interno, restringido a los investigadores del Instituto de Literatura y Lingüística «José Antonio Portuondo Valdor» (ILL), con sede en La Habana—. De su desarrollo ha dado debida cuenta en diversas publicaciones Camacho Barreiro (2009–2013 y 2010). Señala esta investigadora que el *TLEX-Cuba* reúne un total de catorce diccionarios, desde el de Esteban Pichardo (1836), pasando por el *Diccionario cubano, etimológico, crítico, razonado y comprensivo* de José Miguel Macías (1885) hasta el *Diccionario del español de Cuba*, de Reinhold Werner, Antonia Mª Tristá y Gisela Cárdenas (2000)[46]. A la base de datos se ha incorporado únicamente la información lingüística que aportan estos repertorios (se ha prescindido de la información enciclopédica), con la finalidad de «convertirse en un depósito de las palabras y expresiones de nuestra lengua en Cuba. Se podrá rastrear la vida de nuestras palabras y su historia. Permitirá, en definitiva, un mayor y mejor dominio de la variante cubana del español, a través

45 De los 19 333 lemas incorporados, 11 812 constituyen lexías simples y 7521 son unidades fraseológicas. En total, se han indexado 36 088 acepciones, distribuidas en 42 campos temáticos (desde la agricultura o la comunicación hasta el tiempo atmosférico, el tiempo cronológico, la fauna, la flora o las formas de tratamiento).

46 De las obras que han tenido varias ediciones, como el texto de Pichardo, solamente se ha tenido en cuenta la última.

Tabla 2: Tesoros lexicográficos, léxicos o índices de ámbito regional del español americano

Ámbito	Publicación	Fecha de edición	Número de obras indexadas	Período cronológico
México	*Índice mex.*	1997[1], 1998[2], 2000[3]	138	1761–1996
Argentina	*Registro arg.*	2000	215	1845–1999
Uruguay	*DICUR*	Inédito	¿?	¿?
Puerto Rico	*Tesoro.PR*	2005, 2016[2]	60	1788–2003
Cuba	*TLEX-Cuba*	Inédito	14	1836–2000

del tiempo, y a través de lo que los diccionarios reflejaron» (Camacho Barreiro 2010: 2). En cuanto al español dominicano, recientemente ha sido presentado en la Academia Dominicana de la Lengua y en la Fundación Guzmán Ariza el proyecto de recopilación del *Tesoro lexicográfico de la República Dominicana*, del que será coordinadora la investigadora Mª José Rincón. Se trata de un reto que, de manera similar, será necesario acometer en otros países con una amplia y arraigada tradición lexicográfica.

5 Conclusiones

Sea como «diccionario de diccionarios», «léxico de léxicos», «multidiccionario», «memoria sistemática del acervo lexicográfico», «tesoro», «thesaurus» o como «granero del idioma» (siguiendo el símil con el que Neruda se refería a los diccionarios), la lexicografía de los últimos años ha dado muestras de que aquel proyecto de Gili Gaya no resulta inalcanzable y que las nuevas herramientas informáticas pueden favorecer la proliferación de este tipo de «diccionario de segunda generación», reuniendo en un único texto un número ilimitado de vocabularios distintos que, a la vez que mantengan su autonomía e identidad, puedan consultarse como si fueran uno solo. El contraste permitirá advertir mejor lo diferencial y la originalidad, el cambio y la variación, pero esa innovación será más apreciable en tanto el conjunto vaya adquiriendo un volumen importante en calidad y en cantidad.

Hay que tener en cuenta que las nuevas tecnologías no han hecho sino iniciar un proceso irreversible en que el lexicógrafo debe ampliar sus expectativas, ya que el horizonte de una base de datos relacional bien diseñada supera con creces los límites de la edición en papel. La selección de buenas ediciones como punto de partida y el tratamiento informático de los datos redundarán en una mejor catalogación de las informaciones heterogéneas que contiene cualquier tesoro (lingüísticas, etnográficas, culturales, literarias, históricas…), al tiempo

que permitirá generar, a partir de él, nuevos tipos de repertorios específicos atendiendo a criterios como la organización y ordenación de las entradas (superando el problema que suponía la lematización de las expresiones fijas), las lenguas de procedencia, las categorías gramaticales o las distintas marcas diasistemáticas (incluyendo las de especialidad). Los tesoros así concebidos serán imprescindibles para trazar con datos objetivos la tradición lexicográfica pero también como fuente de referencia para reconstruir la historia de las palabras y, con ella, la historia de la cultura.

Referencias bibliográficas

Academia Mexicana de la Lengua (2000[3]): *Índice de mexicanismos registrados en 138 listas publicadas desde 1761*. México: Academia Mexicana de la Lengua-Consejo Nacional para la Cultura y las Artes-Fondo de Cultura Económica.

ADiM = García Mouton, Pilar/Isabel Molina Martos (2015): *Atlas Dialectal de Madrid*. <http://adim.cchs.csic.es/> [último acceso: 20/10/2017].

Ahumada, Ignacio (2000): «Nuevos horizontes de la lexicografía regional», en Stefan Ruhstaller/Josefina Prado Aragonés (eds.), *Tendencias en la investigación lexicográfica del español. El diccionario como objeto de estudio lingüístico y didáctico*. Huelva: Universidad de Huelva, 15–35.

Alcalá Venceslada, Antonio (1998): *Vocabulario andaluz*. Estudio preliminar y edición de Ignacio Ahumada, edición facsímil de la impresa por la Real Academia Española en el año 1951. Jaén: Universidad de Jaén y Caja Sur. [La primera edición es de 1933].

Alcedo, Antonio de (1789 [1967]): «Vocabulario de las voces provinciales de América usadas en el Diccionario Geográfico-Histórico de ella; y de los nombres propios de plantas, aves y animales», en el *Diccionario geográfico histórico de las Indias Occidentales o América*, tomo IV. Edición y estudio preliminar por Ciriaco Pérez-Bustamante. Madrid: Editorial Atlas (Biblioteca de Autores Españoles).

ALDICAM = *Atlas lingüístico diacrónico e interactivo de la Comunidad de Madrid*. <http://aldicam.blogspot.com.es> [último acceso: 15/11/2017].

ALEA = Alvar, Manuel, con la colaboración de Antonio Llorente y Gregorio Salvador (1991[2][1961–1973]): *Atlas Lingüístico y Etnográfico de Andalucía*. Madrid: Arco/Libros.

ALEANR = Alvar, Manuel, con la colaboración de Antonio Llorente, Tomás Buesa y Elena Alvar (1979–1983): *Atlas Lingüístico y Etnográfico de Aragón, Navarra y Rioja*. Madrid-Zaragoza: CSIC-Institución Fernando el Católico.

ALEICan = Alvar, Manuel (1975–1978): *Atlas Lingüístico y Etnográfico de las Islas Canarias*. Las Palmas de Gran Canaria: Cabildo Insular de Gran Canaria.

Alonso, Amado (1951): «Reseña del *Tesoro lexicográfico, 1492–1726*», *Nueva Revista de Filología Hispánica* V/3, 324–328.

ALPI = Navarro Tomás, Tomás (dir.)/Aurelio M. Espinosa hijo/Luís F. Lindley Cintra/Francesc de Borja Moll/Armando Nobre de Gusmão/Aníbal Otero/Lorenzo Rodríguez Castellano/Manuel Sanchis Guarner (1962): *Atlas Lingüístico de la Península Ibérica,* I, *Fonética*. Madrid: CSIC. [Vid. *web ALPI* = García Mouton, Pilar (coord.)/Inés Fernández-Ordóñez/David Heap/Maria Pilar Perea/João Saramago/Xulio Sousa (2016–): *ALPI-CSIC* <www.alpi.csic.es> [último acceso: 20/10/2017], edición digital de Navarro Tomás, Tomás (dir.): *Atlas Lingüístico de la Península Ibérica*. Madrid: CSIC].

Alvar Ezquerra, Manuel (2004): «Evocación y sucesos del *Tesoro léxico de las hablas andaluzas*», en Ignacio Ahumada (ed.), *Lexicografía regional del español*. Jaén: Universidad de Jaén, 37–55.

Alvar Ezquerra, Manuel (2011[2]): *Diccionario de madrileñismos. Voces patrimoniales y populares de la Comunidad de Madrid*. Madrid: Ediciones La Librería.

Álvarez de Miranda, Pedro (1998): *Lexicografía Española Peninsular. Diccionarios Clásicos*, edición en CD-ROM. Madrid: Fundación Histórica Tavera.

Arnal, Mª Luisa/Rosa María Castañer/José María Enguita/Vicente Lagüéns (2012): «La información diatópica en el *Diccionario diferencial del español de Aragón* (DDEAR)», en Dolores Corbella *et al.* (eds.), *Lexicografía hispánica del siglo XXI: nuevos proyectos y perspectivas. Homenaje al Profesor Cristóbal Corrales Zumbado*. Madrid: Arco/Libros, 81–96.

Ayala, Manuel José de (¿1751–1777? [1995]): *Diccionario de voces americanas*, [manuscrito conservado en la Biblioteca del Palacio Real], presentación y edición de Miguel Ángel Quesada Pacheco. Madrid: Arco/Libros.

Buzek, Ivo (2008): «Anotaciones del proyecto "Tesoro léxico del caló"», en Dolores Azorín Fernández *et al.* (coords.), *El diccionario como puente entre las lenguas y culturas del mundo: Actas del II Congreso Internacional de Lexicografía Hispánica*. Alicante: Biblioteca Virtual Miguel de Cervantes, 500–505.

BVFE = Alvar Ezquerra, Manuel (2017): *Biblioteca Virtual de la Filología Española (BVFE): directorio bibliográfico de gramáticas, diccionarios, obras de ortografía, ortología, prosodia, métrica, diálogos e historia de la lengua*. <www.bvfe.es> [último acceso: 20/10/2017].

Camacho Barreiro, Aurora (2009–2013): «Tesoro lexicográfico de Cuba: apuntes sobre un proyecto de rescate», *Anuario L/L, Estudios Lingüísticos* (Instituto de Literatura y Lingüística, La Habana) 7–21.

Camacho Barreiro, Aurora (2010): «Tesoro lexicográfico de Cuba: un recorrido a través de la historia de las palabras (siglos XIX–XXI)», *Revista digital Librínsula*, 1–3. <http://librinsula.bnjm.cu/secciones/300/nombrar/300_nombrar_2.html> [último acceso: 20/10/2017].

Coll, Magdalena (2013): «Prácticas lexicográficas del siglo XIX en territorio uruguayo: de la nominación al registro de piezas de museo», *Revista Argentina de Historiografía Lingüística* 2, 115–136.

Coll, Magdalena (2017): «Hacia una periodización de la lexicografía en territorio uruguayo», *Lingüística* 33–1, 47–71.

Colón, Germà (2014): «La lexicografía histórica catalana a partir de fines del siglo XIX», en Félix Córdoba Rodríguez *et al.* (eds.), *Lexicografía de las lenguas románicas. Perspectiva histórica*. Berlín/Boston: De Gruyter, 115–124.

Corbella, Dolores (2016): «Mundo digital y diccionarios de ámbito regional en el español europeo: la mirada dialectal», en Ignacio Sariego López, Juan Gutiérrez Cuadrado y Cecilio Garriga Escribano (eds.), *El diccionario en la encrucijada: de la sintaxis y la cultura al desafío digital*. Santander: Escuela Universitaria de Turismo Altamira-Asociación Española de Lexicografía Hispánica, 91–110.

CORDE = Real Academia Española: *Corpus Diacrónico del Español*. <www.rae.es> [último acceso: 20/10/2017].

Corrales, Cristóbal/Dolores Corbella (2004): «El *ALEICan* en los diccionarios», *Archivo de Filología Aragonesa* 59–60, 1203–1222.

Corrales, Cristóbal/Dolores Corbella (2009): «Avances y perspectivas de la lexicografía histórico-diferencial», en Dolores Corbella y Josefa Dorta (eds.), *La investigación dialectológica en la actualidad*. Santa Cruz de Tenerife: Agencia Canaria de Investigación, Innovación y Sociedad de la Información, 281–306.

Corrales, Cristóbal/Dolores Corbella (2015): «Recepción de los canarismos en la lexicografía no académica», en María del Pilar Garcés Gómez (coord.), *Léxico, historia y diccionarios*. La Coruña: Universidade da Coruña, 59–76.

Covarrubias, Sebastián (1611 [1977]): *Tesoro de la lengua castellana o española*. Madrid: Ediciones Turner.

DAMER = Asociación de Academias de la Lengua Española (2010): *Diccionario de americanismos*. Lima: Santillana Ediciones Generales. <http://www.asale.org> [último acceso: 20/10/2017].

DCVB = Alcover, Antoni Maria/Francesc de Borja Moll (1930–1962): *Diccionari català-valencià.balear*. Palma: Moll. <http://dcvb.iecat.net> [último acceso: 20/10/2017].

DdD = Santamarina, Antón (coord.) (2006–2013): *Dicionario de dicionarios. Corpus lexicográfico da lingua galega*. Santiago de Compostela: Instituto da Lingua Galega. <http://sli.uvigo.es/DdD> [último acceso: 20/10/2017]. [Se había publicado previamente en CD-ROM en tres ocasiones: 2000, 2001[2] y 2003[3]].

DDGM = González Seoane, Ernesto (dir.)/María Álvarez de la Granja/Ana Isabel Boullón Agrelo (2006): *Dicionario de dicionarios do galego medieval*. Santiago de Compostela: Servicio de Publicacións e Intercambio Científico-Universidad de Santiago de Compostela (*Verba*, anexo 57). <http://sli.uvigo.es/DDGM/index.html> [último acceso: 20/10/2017].

DECan = Corrales, Cristóbal/Dolores Corbella (2010): *Diccionario ejemplificado de canarismos*. La Laguna: Instituto de Estudios Canarios.

DEU = Academia Nacional de Letras (2011): *Diccionario del español del Uruguay*. Montevideo: Ediciones de la Banda Oriental.

DHECan = Corrales, Cristóbal/Dolores Corbella (2013[2] [2001]): *Diccionario histórico del español de Canarias*. La Laguna: Instituto de Estudios Canarios. <www.frl.es> [último acceso: 20/10/2017].

DHLE (1933–1936) = Real Academia Española (1933–1936): *Diccionario histórico de la lengua española*. Madrid. [Tomos I y II: a-cevilla]. <http://web.frl.es/DH1936.html> [último acceso: 20/10/2017].

DHLE (1960–1996) = Real Academia Española (1960–1996): *Diccionario histórico de la lengua española*. Madrid. [Desde el fascículo 1º al segundo del tercer tomo y primero del tomo cuarto]. <http://web.frl.es/DH.html> [último acceso: 20/10/2017].

Diccionario de Autoridades = Real Academia Española (1726–1739): *Diccionario de la lengua castellana, en que se explica el verdadero sentido de las voces, su naturaleza y calidad…* (conocido como *Diccionario de Autoridades*). Madrid. <http://web.frl.es/DA.html> [último acceso: 20/10/2017].

DICI-web = Universidade de Aveiro/Centro de Linguística da Universidade de Lisboa: *Corpus Lexicográfico do Português*. <clp.dlc.ua.pt/DICIweb> [último acceso: 20/10/2017].

Díez Suárez, Soledad (1994): *Léxico leonés*. León: Secretariado de Publicaciones de la Universidad de León.

DiHA = Academia Argentina de Letras (2008[2] [2003]): *Diccionario del habla de los argentinos*. Buenos Aires: Espasa.

DLE = Real Academia Española (2014[23]): *Diccionario de la lengua española*. Barcelona: Espasa Libros. [Denominado tradicionalmente como *DRAE*].

DM = Academia mexicana de la lengua (2010): *Diccionario de Mexicanismos*. México, D. F.: Siglo XXI editores.

DRAE = Real Academia Española (2001[22]): *Diccionario de la lengua española*. Madrid: Espasa. [Si se hace referencia a alguna de las ediciones anteriores se precisa en cada caso. A partir de 2014 se conoce con la sigla *DLE*].

FRONTESPO = *Frontera hispano-portuguesa*. <http://www.frontespo.org/es> [último acceso: 20/10/2017].

García Márquez, Gabriel (1993): *El General en su Laberinto*. Barcelona: RBA.

Gili Gaya, Samuel (1960): *Tesoro lexicográfico (1492–1726)*. Tomo I (A–E). Madrid: CSIC. [Se publicó primero en fascículos entre 1947 y 1957].

Glossari de Glossaris = Colón Domènech, Germà (2007): *Glossari de Glossaris*. Barcelona: Editorial Barcino. <http://www.glossaris.net/advSearch.php> [último acceso: 20/10/2017].

Gómez Alfaro, Antonio (1997): «Diccionarios de la lengua romaní», *Interface* 28, 3–7.

Gómez de Silva, Guido (2008[2] [2001]): *Diccionario breve de mexicanismos*. México: Fondo de Cultura Económica-Academia Mexicana de la Lengua.

González Salgado, José Antonio (2012): «Aspectos teóricos y metodológicos del *Tesoro léxico de las hablas extremeñas*», en Antoni Nomdedeu Rull, Esther Forgas Berdet y Maria Bargalló Escrivà (eds.), *Avances en lexicografía hispánica*. Tomo I. Tarragona: Universidad Rovira i Virgili, 155–170.

González Salgado, José Antonio (2017): «El léxico portugués en las hablas dialectales de las comarcas rayanas españolas», en Dolores Corbella y Alejandro Fajardo (eds.), *Español y portugués en contacto. Préstamos léxicos e interferencias*. Berlín/Boston: De Gruyter, 105–127.

González Seoane, Ernesto *et al.* (2008): «El *Dicionario de dicionarios do galego medieval*», en Janet Ann DeCesaris y Elisenda Bernal (coords.), *Proceedings of the XIII EURALEX International Congress (Barcelona, 15–19 July 2008)*. Barcelona: Institut Universitari de Lingüística Aplicada-Universitat Pompeu Fabra, 385–389.

Guardiola i Savall, Mª Isabel (2008): «El "Tresor lexicogràfic valencià" (1543–1915) (*TLV*)», en Dolores Azorín Fernández *et al.* (coords.), *El diccionario como puente entre las lenguas y culturas del mundo: Actas del II Congreso Internacional de Lexicografía Hispánica*. Alicante: Biblioteca Virtual Miguel de Cervantes, 106–111.

Guardiola i Savall, Mª Isabel (2017): «Diccionaris de diccionaris o tresors lexicogràfics», *Estudis romànics* 39, 427–436.

Guitarte, Guillermo L. (1996): «Una carta de Amado Alonso a Rodolfo Lenz. El proyecto de un *corpus* de estudios sobre el español extrapeninsular», *Lexis* XX, 63–86.

Gutiérrez Cuadrado, Juan (2003): «*Diccionario de Diccionarios*, vino añejo en odres nuevos», *Estudis Romànics* XXVIII, 370–377.

Haensch, Günther (1994): «Dos siglos de lexicografía del español de América: Lo que se ha hecho y lo que queda por hacer», en Gerd Wotjak y Klaus Zimmermann (eds.), *Unidad y variación léxicas del español de América*. Madrid/Fráncfort: Iberoamericana/Vervuert, 39–82.

Haensch, Günther (2000): *Textos Clásicos sobre la Historia de la Lexicografía del Español de América*, edición en CD-ROM. Madrid: Fundación Histórica Tavera.

Haensch, Günther/Carlos Omeñaca (2004²): *Los diccionarios del español en el siglo XXI*. Salamanca: Ediciones Universidad de Salamanca.

LEXDIALGRAM = Perea, Maria-Pilar (coord.) (2014): *Lexdialgram. Portal de lèxics i gramàtiques dialectals del català del segle XIX*. Barcelona: Universitat de Barcelona. <http://www.ub.edu/lexdialgram/index.jsp?item=presentacio&idx=0> [último acceso: 20/10/2017].

LLA = Le Men, Janick (2002–2012): *Léxico del leonés actual*. León: Caja España-Archivo Histórico Diocesano.

LMP = Alvar, Manuel (1985–1989): *Léxico de los marineros peninsulares*. Madrid: Arco/Libros.

Mapa de diccionarios = Instituto de Investigación Rafael Lapesa de la Real Academia Española (2013): *Mapa de diccionarios*. <http://web.frl.es/ntllet/SrvltGUILoginNtlletPub> [último acceso: 20/10/2017].

Messner, Dieter (1994–2007): *Dicionário dos dicionários portugueses*. Salzburgo: Universidad.

Messner, Dieter (2008): «El *Dicionário dos dicionários* portugueses», en Dolores Azorín Fernández *et al*. (coords.), *El diccionario como puente entre las lenguas y culturas del mundo: Actas del II Congreso Internacional de Lexicografía Hispánica*. Alicante: Biblioteca Virtual Miguel de Cervantes, 33–38.

Miguélez, Eugenio (1998² [1993]): *Diccionario de las hablas leonesas (León, Salamanca, Zamora)*. León: Ediciones Monte Casino.

Moreno Moreno, Mª Águeda (2007): *Léxico histórico andaluz. 1. Período clásico*. Jaén: Universidad de Jaén.

Muñoz Garrigós, José/José Perona (1996): «Los vocabularios murcianos», en Ignacio Ahumada (coord.), *Vocabularios dialectales. Revisión crítica y perspectivas*. Jaén: Universidad de Jaén, 83–100.

Nagore Laín, Francho (dir.) (1999): *Endize de bocables de l'aragonés, seguntes os repertorios lésicos de lugars y redoladas de l'Alto Aragón*. Huesca: IEA.

Neira Martínez, Jesús/Mª del Rosario Piñeiro (1989): *Diccionario de los bables de Asturias*. Oviedo: Real Instituto de Estudios Asturianos.

Nieto Jiménez, Lidio (2002): *Tesoro lexicográfico del español marinero anterior a 1726*. Madrid: Arco/Libros.

NTLE = Nieto Jiménez, Lidio/Manuel Alvar Ezquerra (2007): *Nuevo Tesoro Lexicográfico del Español (s. XIV-1726)*. Madrid: Arco/Libros.

NTLLE = Real Academia Española (2001): *Nuevo Tesoro Lexicográfico de la Lengua Española*, edición en DVD. Madrid: Espasa Calpe. <www.rae.es> [último acceso: 20/10/2017].

Oudin, César (1607): *Tesoro de las dos lenguas francesa y española/Thresor des deux langues françoise et espagnolle*. París: Marc Orry. [Se cita a través del *NTLLE*].

Padilla, fray Thomas de (1557): *Historia de las cosas de Etiopia*…, de Francisco Aluarez, Capellan del Rey Don Manuel de Portugal. Agora nueuamente traduzido de Portugues en Castellano, por… Amberes. [Seguimos la edición anotada de Manuel de Paz Sánchez, inédita].

Perea, Maria Pilar (2015): «"*Lexdialgram*": un portal de diccionarios y gramáticas dialectales del catalán del siglo XIX», en María del Pilar Garcés Gómez (coord.), *Léxico, historia y diccionarios*. La Coruña: Universidade da Coruña, 309–322.

Pichardo y Tapia, Esteban (1985[6] [1836]): *Diccionario provincial casi razonado de vozes y frases cubanas*. La Habana: Editorial de Ciencias Sociales [Se publicó por vez primera con el título de *Diccionario provincial de voces cubanas*].

Quesada Pacheco, Miguel Ángel (2002): *El español de América*. Costa Rica: Ed. Tecnológica de CR.

Quevedo, Francisco de (1626 [1699]): *Cuento de cuentos*, Alicante, Biblioteca Miguel de Cervantes. Reproducción digital a partir de *Obras de Francisco de Quevedo Villegas… [tomo primero]*. Amberes: Henrico y Cornelio Verdussen, 506–514.

RLA = Academia Argentina de Letras (2000): *Registro de Lexicografía Argentina*. Buenos Aires: Departamento de Investigaciones Filológicas de la Academia Argentina de Letras.

Salvador, Gregorio (1992): «Prólogo» del *Tesoro lexicográfico del español de Canarias* (*TLEC*). Madrid-Canarias: Real Academia Española y Consejería de Educación, Cultura y Deportes del Gobierno de Canarias, 9–11.

Salvador Rosa, Aurora (1985): «Las localizaciones geográficas en el *Diccionario de Autoridades*», *Lingüística Española Actual* 7, 103–139.

Tesoro.PR = María Vaquero/Amparo Morales (2005): *Tesoro lexicográfico del español de Puerto Rico*. San Juan: Academia Puertorriqueña de la Lengua Española. <https://tesoro.pr> [último acceso: 20/10/2017].

TLCA = Corrales, Cristóbal/Dolores Corbella (2010): *Tesoro léxico canario-americano*. Las Palmas de Gran Canaria: Casa Colón y Cabildo de Gran Canaria.

TLEC = Corrales, Cristóbal/Dolores Corbella/Mª Ángeles Álvarez (1996[2] [1992]): *Tesoro lexicográfico del español de Canarias*. Madrid-Canarias: Real Academia Española y Consejería de Educación, Cultura y Deportes del Gobierno de Canarias.

TLHA = Alvar Ezquerra, Manuel (2000): *Tesoro léxico de las hablas andaluzas*. Madrid: Arco/Libros.

TLHR = Pastor Blanco, José María (2004): *Tesoro léxico de las hablas riojanas*. Logroño: Universidad de La Rioja.

TLPGP = Álvarez, Rosario (coord.): *Tesouro do léxico patrimonial galego e portugués*. Santiago de Compostela: Instituto da Lingua Galega. <http://ilg.usc.es/Tesouro/> [último acceso: 20/10/2017].

Verdelho, Telmo/João Paulo Silvestre (2014): «El *Corpus Lexicográfico do Português*: la memoria de los diccionarios en la historia de la lengua y de la cultura», en Félix Córdoba Rodríguez *et al.* (eds.), *Lexicografía de las lenguas románicas. Perspectiva histórica*. Berlín/Boston: De Gruyter, 297–311.

Vila Rubio, Mª Nieves (2014): «El "Tesoro Lexicográfico" de Samuel Gili Gaya: contexto, recepción y destino de un diccionario inacabado», en María Bargalló Escrivà, María del Pilar Garcés Gómez y Cecilio Garriga Escribano (eds.), *«Llaneza»: estudios dedicados al profesor Juan Gutiérrez Cuadrado*. La Coruña: Servicio de Publicaciones de la Universidad de La Coruña, 371–394.

Vittori, Girolamo (1609): *Tesoro de las tres lenguas francesa, italiana y española. Thresor des trois langues françoise, italienne et espagnolle*. Ginebra: Philippe Albert & Alexandre Pernet. [Se cita a través del *NTLLE*].

Zerolo, Elías/Miguel de Toro y Gómez/Emiliano Isaza y otros escritores españoles y americanos (1895): *Diccionario Enciclopédico de la Lengua Castellana*. París: Garnier Hermanos. [Se cita a través del *NTLLE*].

Rafael Arnold, Stefan Serafin, Anna-Susan Franke y
Jutta Langenbacher-Liebgott

Una nueva fuente para la historia del léxico español: el *DEMel*

Resumen: El objetivo del proyecto *Diccionario del Español Medieval electrónico* (*DEMel*), que empezó en el mes de noviembre de 2016 con el apoyo financiero de la Fundación Alemana para la Investigación Científica (DFG), es la digitalización del fichero del *Diccionario del español medieval* (*DEM*). A finales de 2019 deberán estar disponibles en la red cerca de 33 000 lemas con las respectivas documentaciones, que superarán el millón de ejemplos. En el capítulo, presentaremos, por un lado, la amplia base documental del repertorio lexicográfico del *DEM* y las características principales de este archivo lexicográfico y expondremos, por otro lado, las etapas de la digitalización del fichero, así como las posibilidades que el material, una vez digitalizado, podría ofrecer al usuario.

Palabras clave: Lexicografía histórica, Lexicografía digital, Español medieval, *Diccionario del Español Medieval electrónico* (*DEMel*)

Abstract: The research project *Diccionario del Español Medieval electrónico* (*DEMel*), which is funded by the German Research Foundation (DFG) and began in November 2016, has the objective of digitalizing the paper slip collection of the *Diccionario del español medieval* (*DEM*), the *DEM* file. At the end of 2019, about 33,000 lemmas with their lexical items and corresponding documentation, which far exceed a total of one million, will be made available online. This article will present, on the one hand, the wealth of material constituted by the lexicographic repertory of the *DEM* and, on the other, the steps relating to the digitization of the *DEM* file. Finally, it will be shown which possibilities the digitalized material can offer to the user.

Keywords: Historical lexicography, e-Lexicography, Medieval Spanish, *Diccionario del Español Medieval electrónico* (*DEMel*)

1 Introducción

Si caracterizamos el *Diccionario del Español Medieval electrónico* (*DEMel*)[1] como nueva fuente para la historia del léxico español, hay que precisar, primero, que lo

1 Dedicamos respetuosamente este artículo al profesor Max Pfister (fallecido en octubre de 2017), uno de los grandes maestros de la Lingüística y de la Filología Románicas y

nuevo concierne a la puesta a disposición, en versión electrónica, de una fuente que ya existía antes, el fichero del *Diccionario del español medieval* (*DEM*) de Bodo Müller (Universidad de Heidelberg); segundo, que a través de la digitalización de los originales físicos, o sea de las fichas lexicográficas, fundamento esencial del *DEM*, se crea una nueva fuente lexicográfica para el conocimiento de la historia del léxico español, el *DEMel*, que se encuentra actualmente en la fase de elaboración. Por eso, en las páginas siguientes, nos proponemos presentar estas dos fuentes, también en comparación con otros recursos digitales de orientación histórica que, de cualquier forma, toman en consideración el léxico del español medieval, para ubicar el proyecto *DEMel* en el contexto de los corpus informatizados, de bases de datos y especialmente de la lexicografía digital. Por supuesto, revisaremos, muy brevemente, la base material del *DEM* y expondremos en forma sucinta la estructura de su fichero para poder explicar con más detenimiento las etapas de la digitalización del material léxico. Pero antes de entrar en los detalles de estas cuestiones, nos parece interesante prestar atención especial a otro tipo de fuente que data del año 1972, una carta, escrita a máquina por Rafael Lapesa, que tiene como destinatario a Bodo Müller, autor del *DEM*, porque este documento nos permite formarnos una idea sobre el diálogo y las visiones de ambos eruditos en cuanto a la lexicografía histórica del español en el último cuarto del siglo XX.

2 Carta de Rafael Lapesa a Bodo Müller (1972), unas breves pinceladas históricas recordando los primeros momentos del *DEM*

La carta de Rafael Lapesa, que era en ese entonces el Director del Seminario de Lexicografía de la Real Academia Española, se dirige al «Distinguido colega» (Lapesa 1972: 1) y amigo Bodo Müller, y el contenido nos muestra que se trata de la respuesta a una carta enviada, probablemente poco antes, por el colega alemán. Por desgracia, no conocemos la carta que escribió Bodo Müller a Rafael Lapesa, pero podemos entrever a través de las palabras de Lapesa las reflexiones de su colega de Heidelberg. De todos modos, la carta nos desvela claramente las ideas de los dos eruditos respecto a una pasión que compartieron, la historia de la lengua española y la historia de su léxico.

presidente de la Comisión del *DEM* (2001–2007), que nos alentó con gran entusiasmo en la realización del proyecto *DEMel*.

Sin duda alguna, en su carta Müller había anunciado el proyecto del *DEM* porque empieza la respuesta de Lapesa (1972: 1) con las siguientes palabras:

> Recibo con grata sorpresa su atenta carta, en la que me comunica Vd. el nacimiento de obra tan necesaria y estimable como la preparación de un *Diccionario del Español Medieval*. El enfoque que Vd. da a dicha tarea, extendiéndola a todos los dominios de la lengua, es un acierto indudable, y augura muy valiosos resultados.

Por falta de espacio no podemos comentar todos los aspectos tratados y discutidos en la carta de Lapesa. Por ejemplo, refiriéndose a una lista de 98 términos que Müller había añadido a su carta, términos que comienzan con la letra A- y que no se encuentran en el *Diccionario histórico de la lengua española* (*DHLE*), Lapesa (1972: 1) explica estas «lagunas» con los «distintos criterios respecto a la determinación de las fuentes, a su utilización y al modo de analizar y presentar el material». Por supuesto, Lapesa comenta una serie de grafías y sus posibles interpretaciones, también diversos manuscritos y vocabularios o el tratamiento de los diminutivos y participios en el *DHLE*. Pero aún más impresionante en su argumentación a favor de la lexicografía histórica son su perspicacia, la calidad de su juicio científico y su pensamiento estratégico, que emergen a través de sus palabras que reflejan las ideas y proposiciones de Bodo Müller: «Acojo con gran interés su proposición de establecer un intercambio de listas de términos procedentes de determinadas obras, para evitarnos un doble y paralelo esfuerzo de lectura» (Lapesa 1972: 4). Ambos pensaban, pues, en una colaboración coordinada que podría ser fructuosa para los dos proyectos, el *DHLE* y el *DEM*. Concluye Lapesa (1972: 4–5):

> […] como Vd. mismo señala, es tanta la necesidad de cubrir el vacío existente en la filología española en punto a instrumentos para la investigación histórica del léxico, que la Academia no dudó en acuciar a nuestro Seminario [de Lexicografía] para que se publicaran las primeras entregas. En relación al volumen del diccionario, este primer tomo [i. e. del *DHLE*] será mínima y reparable parte, pero habrá cumplido con *la fundamental misión de impedir que el impulso inicial se perdiese en la interminable espera de perfeccionar unos archivos que son por naturaleza imperfectibles, por la naturaleza viva del idioma, de un lado, y de otro por la constante labor de adición de manuscritos y documentos inéditos medievales* [las cursivas son nuestras].

Con estas ideas directrices de Lapesa (1972: 5), nosotros, los equipos investigadores, lingüistas y filólogos de los departamentos de Filología Románica de las Universidades de Rostock y Paderborn con la colaboración de los informáticos de la Biblioteca Universitaria de Rostock, quisiéramos realizar el *DEMel*, o sea aprovechar «el impulso inicial» del proyecto, aceptando ciertas imperfecciones del archivo *DEM*, y conseguir poner en línea todo su material lexicográfico.

3 Lingüística histórica y lexicografía digital en el ámbito hispánico: corpus y bancos de datos del español y el *DEMel*

En 1987, en su artículo célebre «Notes sur *lexicographie* et *dictionnairique*», Bernard Quemada (1987: 229–242) redefine el concepto tradicional de *lexicografía* que propone concebir como actividad meramente científica, como estudio científico de las palabras, sus significaciones, su uso morfosintáctico, su uso gráfico, etc. en un contexto dado de una lengua determinada. Por consiguiente, la finalidad de la lexicografía deja de estar ligada necesariamente a la elaboración y producción de un diccionario, tarea central de la *diccionarística* («dictionnairique»). Así, la lexicografía, por su parte, «peut très bien ne pas sortir d'un laboratoire et correspondre, par exemple, à des bases informatisées destinées aux seuls chercheurs» (Pruvost 2005: 10), de modo que «lexicógrafo y diccionarista *ya no son sinónimos*» (Campos Souto 2015: 80). Cabe destacar que es precisamente la aparición de las bases o bancos de datos lo que va a reforzar esta diferenciación esencial entre lexicografía y diccionarística, o como lo subraya Rodríguez Barcia (2016: 140):

> […] la recopilación, estructuración y análisis del léxico a partir de la realización de una base de datos es competencia de la lexicografía y de otras disciplinas, y solo en un segundo paso se procedería a la confección de un diccionario con los materiales e informaciones conseguidas.

Por supuesto, estas reflexiones teóricas plantean la cuestión de si nuestro proyecto forma parte de la lexicografía o de la diccionarística, una cuestión que está vinculada también al nombre del proyecto que mantiene la etiqueta de *diccionario*. Para dar respuesta a este problema, tratamos, en primer lugar y de manera general, de comparar y relacionar el *DEMel* con los corpus y bases de datos más relevantes en el ámbito hispánico que enfocan, lingüísticamente hablando, la dimensión histórica.

Sin duda alguna, con el banco de datos de la Real Academia Española, disponemos de una gran variedad de informaciones lingüísticas, filológicas y otras para el estudio de la lengua española, pero se evidencia que ninguno de estos recursos disponibles en la red se centra exclusivamente en el español medieval. Sin embargo, entre los corpus textuales con orientación histórica, si partimos de los criterios establecidos por Rojo (2016b: 285), que define *corpus* como:

> […] un conjunto de (fragmentos de) textos naturales, almacenados en formato electrónico, representativos en su conjunto de una variedad lingüística, en alguno de sus componentes o en su totalidad, y reunidos con el propósito de facilitar su estudio científico […].

destacan el *Corpus diacrónico del español* (*CORDE*) con textos desde el origen de la lengua hasta 1974, complementado por el *Corpus de referencia del español actual* (*CREA*) que incorpora textos escritos y orales de los años 1975 hasta 2005 y representa así, junto con el *Corpus del español del siglo XXI* (*CORPES XXI*), la perspectiva sincrónica. La imprescindibilidad del *CORDE* para el estudio de la historia del español que subraya Kabatek (2016: 7)[2] es evidente, y, a pesar de algunos inconvenientes tanto a nivel filológico como en cuanto a su aplicación (Rojo 2016a: 197), estamos totalmente de acuerdo con las conclusiones que Rojo (2010: 45 y 48) deduce de su comparación del *CORDE* con el *Corpus del español* (*CE*) de Mark Davies, otro corpus general con textos del siglo XIII al XX y «gran competidor holocrónico» (Octavio de Toledo y Huerta 2016: 61, nota 11) del *CORDE*:

> […] el *CE* está parcialmente lematizado y tiene un sistema de consultas muy ágil y brillante, capaz de devolver en décimas de segundo las frecuencias generales y relativas de las formas correspondientes a una expresión regular o los candidatos a colocaciones de una palabra y mostrar cuadros con la distribución de esas frecuencias por siglos y, en el caso del XX, también del tipo de texto […]. El *CREA* y el *CORDE*, en cambio, están basados en una codificación muy cuidadosa y detallada que permite recuperar los datos que interesan estableciendo el filtro por cualquiera de los parámetros que han intervenido en la construcción de los corpus (año, país, tipo de texto y área temática), así como autor y obra […]. Por supuesto, todos estos parámetros son combinables en una búsqueda única […]. El *CREA* y el *CORDE* […] resultan mucho más útiles que el *CE*. Por número de textos, por su codificación externa e interna, por las posibilidades de selección de textos según diferentes parámetros que brinda el sistema de búsquedas e incluso por las facilidades de exportación que ofrecen.

Esta conclusión de Rojo, que es una réplica a un artículo de Davies (2009) en el que este había subrayado las ventajas del *CE* frente al *CORDE*, tiene todavía validez, incluso después de la publicación de la nueva versión 2015–2016 del *CE* que ahora

> […] consta de dos partes (ambos ya están disponibles con una interfaz y con archivos de ayuda en español e inglés): el corpus (original y más pequeño) que permite buscar cambios históricos y variación de géneros; el corpus (nuevo y mucho más grande) que permite buscar variaciones dialectales (y tiene cien veces más datos para el español contemporáneo) (Davies 2017, refiriéndose a su artículo de 2009).

Por su parte, la Real Academia Española pone a disposición del usuario otro corpus diacrónico de gran valor, el *Corpus del diccionario histórico* (*CDH*) o, más

2 Vid. también Rojo 2014: 380–381; Rojo 2016a: 197; Rojo 2016b: 286 y 289; y Sánchez Sánchez/Domínguez Cintas 2007: 138–140 y 142–146.

precisamente, el *Corpus del Nuevo diccionario histórico del español* (*CNDHE*). Concebido sobre todo como base material para el *Nuevo Diccionario Histórico del Español* (*NDHE*), el *CNDHE* se presenta en su versión 3.1. con una nueva interfaz cuya

> […] principal novedad […] consiste en la posibilidad de efectuar consultas dinámicas sobre coapariciones (combinaciones frecuentes de las palabras) en todo el corpus, con lo que se pueden obtener con gran facilidad datos relevantes para el estudio del léxico y la gramática del español desde el siglo XIII hasta la actualidad (Real Academia Española 2017b).

El corpus integra, «en buena medida», textos «comunes al *CORDE* y al *CREA*», pero estos textos «se han sometido a un proceso semiautomático de anotación lingüística» (Real Academia Española 2017a). De este modo, el *CNDHE* representa, sobre todo en comparación con el *CORDE*, una notable mejora respecto a la calidad lingüística, filológica (Kabatek 2016: 8) e informática. En lo que concierne al *NDHE*, que fue concebido desde su origen «como un diccionario electrónico y relacional», Pascual/Campos Souto (2017) lo caracterizan también «como un diccionario perfectible, presentado por capas, en el que está previsto que los propios usuarios contribuyan a su mejora». Esa obra en desarrollo es indudablemente la obra maestra, emblemática y ejemplar de la lexicografía digital del hispanismo actual, como lo pone de manifiesto la alta calidad científica tanto a nivel lexicográfico, filológico (Pascual 2015: 7–13) como informático de los más de 1000 artículos del *NDHE* que ya están disponibles en la red (Campos Souto 2015: 71–93 y Salas Quesada/Torres Morcillo 2015: 15–69). Bajo la dirección de José Antonio Pascual y Mar Campos Souto, el equipo del *NDHE* prevé la realización de más de 1000 entradas anuales consultables en línea (Ruiz Mantilla 2017).

Por supuesto, hay otros corpus textuales y bases de datos de gran valor y utilidad para el estudio de la historia del español, aparte de los que acabamos de mencionar (v. también los artículos respectivos en este libro), como p. ej. el *Corpus Hispánico y Americano en la Red: Textos Antiguos* (*CHARTA*), pensado y diseñado para la edición y estudio lingüístico de documentos archivísticos en español de los siglos XII–XIX. Para tal fin, el corpus presenta los textos en una triple edición, facsímil, paleográfica y crítica (Enrique-Arias 2012: 87); el *Corpus de documentos españoles anteriores a 1800* (*CODEA+ 2015*), que desde el principio ha aplicado los criterios establecidos por la red *CHARTA* y que en su versión actual comprende 2500 documentos y brinda una gran versatilidad en la búsqueda; o el corpus *Biblia Medieval* «que permite consultar en paralelo versiones medievales españolas de la Biblia, compararlas con su fuente hebrea o latina y ver

imágenes digitales de los originales» (Enrique-Arias 2012: 87). Una trayectoría más larga tienen el *Archivo Digital de Manuscritos y Textos Españoles* (*ADMYTE*) y la *Biblioteca Digital de Textos del Español Antiguo*. El primero está ahora en línea, técnicamente mejorado y con una ampliación del número de los textos accesibles, después de dos versiones en CD de los años 1991 y 1992. La segunda se basa fundamentalmente en los archivos textuales del *Hispanic Seminary of Medieval Studies* (HSMS) en Madison (Wisconsin) y sus proyectos, iniciados en los años 70 del siglo XX y dirigidos en aquel entonces por Lloyd A. Kasten y John J. Nitti, dieron lugar, entre otros, al *Dictionary of the Old Spanish Language*, de donde surgieron posteriormente el *Diccionario de la prosa castellana del Rey Alfonso X* (Kasten/Nitti 2002) y el *Tentative Dictionary of Medieval Spanish* (Kasten/Cody 2001[2]). Este último se refiere al léxico de 86 textos en total, sobre todo literarios, de alrededor de 1140 hasta 1489 (Müller 2003: 392; Reinhardt 2014: 638). Además, disponemos hoy en día de prestigiosos «corpus territoriales» (Carrasco Manchado 2011: 351) y obras lexicográficas electrónicas como el *Diccionario histórico del español de Canarias* (*DHECan*) de Corrales/Corbella que, después de dos ediciones impresas (2001 y, muy aumentada, 2013[2]), está disponible en línea a través de la página web del Instituto de Investigación Rafael Lapesa de la Real Academia Española.

4 Características del *DEM* y el archivo *DEM*

Dado el gran número y la diversidad de corpus y bases de datos de que nos podemos beneficiar para el estudio de la historia de la lengua española y que, evidentemente, son también útiles, de una manera u otra, para el análisis del léxico del español medieval, hay que precisar las ventajas del futuro *DEMel* y su base material, el fichero *DEM*: primero, el material léxico del fichero *DEM* y, por consiguiente, del *DEMel* representa el caudal léxico desde los inicios de la lengua hasta comienzos del siglo XV y focaliza así, consecuente y exclusivamente, el español medieval; segundo, ese material está completamente lematizado, no lo están el *CORDE* y el *CE* (este último solo parcialmente); tercero, conforme a la idea directriz de Müller (1987: V) de «describir en toda su extensión el léxico tradicional fijado en la lengua escrita de la época estudiada», la base documental del fichero *DEM* (v. más abajo) puede cumplir el criterio de representatividad, una cuestión muy discutida en la lingüística de corpus (Torruella 2016: 98–105); cuarto, el *DEMel* indica la categoría gramatical de los lemas y señala las variantes gráficas; quinto, las palabras y expresiones documentadas normalmente están apuntadas en las fichas con el contexto para ilustrar el uso semántico, y a menudo las acepciones están acompañadas de una definición provisional;

sexto, se especifica, en forma de sigla, la fuente con su datación donde aparece la palabra. De esta manera, la consulta del futuro *DEMel* permitirá al usuario una presentación cronológica de las documentaciones y le ofrecerá, desde una perspectiva general, una sinopsis de la evolución del uso semántico, gráfico, morfológico, fraseológico y también la primera documentación de las palabras.

Por supuesto, el *DEMel* no podrá competir con un proyecto como el que representa el *NDHE*, ya que la base material del *DEMel*, el fichero *DEM*, forma parte de la lexicografía histórica tradicional y no fue concebido con el objetivo de crear un diccionario electrónico. No obstante, el *DEM* es un diccionario que siempre ha intentado satisfacer las más altas pretensiones lingüísticas y filológicas, lo que se refleja en la concepción y estructura de los artículos del *DEM* que, según Müller (1987: V), respetan

> […] tres aspectos centrales: la documentación de los vocablos, que ofrece ejemplos de su empleo según un ordenamiento cronológico-semántico; información lexicográfica adicional, que complementa los datos más importantes contenidos en los antiguos diccionarios españoles; y por último, comentarios acerca de la interpretación lexicológica e histórico-lingüística del material. De este modo el *DEM* integra las características de diccionario descriptivo de una etapa de la lengua y de diccionario etimológico-histórico.

El fundamento de esta obra es el archivo *DEM*, que abarca, como núcleo, el fichero *DEM* en que se hallan recogidas las fichas en papel que constituyen el repertorio lexicográfico, lematizado y ordenado alfabéticamente, y que consta de 376 gavetas. Además, el archivo contiene otras 104 gavetas con informaciones adicionales (etimológicas, bibliográficas, extralingüísticas y otras)[3]. En su conjunto es el producto de un inmenso trabajo lingüístico, filológico y, por supuesto, lexicográfico —prediccionarístico en el sentido de Quemada (1990: 64) porque el fichero *DEM* es el resultado del intento de confeccionar un diccionario—. Por consiguiente, la elaboración, estructuración y sistematización del fichero se operaron desde el principio en función de la finalidad de crear el tipo de diccionario descrito por Müller y, teniendo en cuenta estos hechos, nos parece justificable mantener la etiqueta *diccionario* en la designación del proyecto: *DEMel.* Adicionalmente, la denominación misma nos permite transparentar y recalcar la relación esencial con el *DEM*. A continuación, expondremos muy brevemente algunas características más del *DEM* y su archivo antes de detallar los aspectos técnico-informáticos y describir el estado actual del *DEMel.*

Como ya hemos explicado, el archivo *DEM* consta de 480 gavetas. Estas abarcan cerca de 865 000 fichas. Si tenemos en cuenta que en una ficha pueden

3 Para una descripción de las diversas partes del archivo *DEM,* véase Arnold (2016).

estar registradas dos o más atestaciones, hay que contar con más de un millón de documentaciones en total. El lemario comprende alrededor de 33 000 lemas. El punto de partida del archivo surgió de la idea de Bodo Müller, a finales de los años 60 y principios de los 70 del siglo XX, de redactar un pequeño diccionario estudiantil para facilitar la lectura de textos medievales. Desde 1971 hasta 1984, el *DEM* recibió el «apoyo material de la DFG [sc. Deutsche Forschungsgemeinschaft]» y después de la Academia de Ciencias de Heidelberg «que acordó acoger al *DEM* en su propio programa de investigación» (Müller 1987: VI). Así, en el Seminario de Filología Románica de la Universidad de Heidelberg se empezó a confeccionar un archivo léxico del español medieval basándose en textos desde el siglo X hasta comienzos del siglo XV que se fue ampliando de manera progresiva hasta aproximadamente 1995. Paralelamente, había empezado, en el año 1987, la publicación del *DEM* en formato de libro. Con el fin del proyecto, a finales del año 2007, la publicación que había llegado hasta el fascículo 26 (ALLÉN-ALMOHATAC) se interrumpió. Tampoco el apoyo del *Centro de Estudios Medievales y Renacentistas* de la Universidad de La Laguna (Tenerife), colaborando con el proyecto gracias a un contrato de cooperación acordado en abril de 2005, pudo evitar el fracaso del *DEM*. Así, de manera similar al *DHLE*, el *DEM* quedaba inacabado. No obstante, las reacciones de la comunidad científica respecto al *DEM* fueron muy positivas como lo muestran por ejemplo las reseñas de Clavería Nadal (1991: 1107), que consideraba el *DEM* como un diccionario que «abre nuevas y prometedoras perspectivas a la investigación histórica de la época medieval del español», o de Neumann-Holzschuh (1996: 581), que subrayaba la «riqueza de documentación y [una] metodología irreprochable», o un artículo de Metzeltin (1992: 441a) en el que el autor caracterizaba el *DEM* como «un diccionario de gran envergadura, precioso por la riqueza documental, el análisis semántico pormenorizado y la historia lingüística del material»[4]. También la Real Academia Española, en la persona del entonces Secretario Guillermo Rojo (2006: IX), se deshizo en elogios sobre Bodo Müller y «[l]as investigaciones del ilustre romanista», cuyo *DEM* representaba según Rojo, «una base fundamental para el conocimiento del castellano en su primera época», que permite «ampliar considerablemente las fuentes de conocimiento casi exclusivamente literarias que se tenían de la lengua en esa etapa originaria». Rojo (2006: IX) terminaba su carta a Bodo Müller con las siguientes palabras: «El tratamiento filológico y lexicográfico de esos datos ha presentado, desde el primer momento, un carácter

4 Vid. también Bracchi 1996: 419–420, Darbord 1990: 283, Pharies 1991: 79–80, Verd Conradi 1989: 361 o, para más información, Arnold *et al.* 2016: 30.

modélico»[5]. En un artículo de *El País* que apareció algunos meses después del congreso, Pilsel (2005), refiriéndose a Guillermo Rojo, retomó parcialmente las palabras de este al señalar: «El tratamiento filológico y lexicográfico del diccionario es de un carácter modélico tan exquisito que la Real Academia […] quiere rendir homenaje al octogenario profesor para reconocer la deuda que el hispanismo tiene con él contraída».

A pesar del abandono de la publicación del *DEM* en versión impresa, cabe destacar que la calidad científica del archivo *DEM* es de suma relevancia para la lexicografía histórica. Además de las características ya mencionadas anteriormente, hay que resaltar las particularidades del fichero *DEM* en cuanto a la selección de los documentos que constituyen su base material. En total se trata de más de 600 documentos, pero también de concordancias, vocabularios y otras fuentes que el *DEMel* registrará en la bibliografía electrónica junto con las respectivas siglas empleadas en el fichero/archivo *DEM*. Este conjunto de documentos comprende, aparte de los documentos hispanolatinos (siglos X–XII), textos literarios, hasta alrededor de 1400, el gran corpus de prosa jurídica (p. ej. fueros, ordenanzas y colecciones diplomáticas), historiográfica, técnica y científica (astronomía, mineralogía, agronomía, botánica, farmacología, etc.), y sobresalen entre estos documentos las traducciones al español medieval como las de la Biblia y las de textos árabes o de otras lenguas. De esta manera, el *DEMel*, tomando en consideración la riqueza léxica que ofrece el fichero *DEM* y las informaciones adicionales del archivo, podrá suministrar datos de gran valor para el conocimiento del español medieval. Además, el *DEM* y asimismo el material de su fichero destacan por otra cualidad. En el transcurso del tratamiento del material se puso de manifiesto que, de las palabras o acepciones que fueron publicadas en los dos primeros tomos del *DEM*, no menos de un 22 %, es decir, casi una cuarta parte, no había sido registrada hasta la fecha. En más de la mitad de los casos, se dieron significados inéditos y en similar proporción se ofrecieron testimonios más antiguos que retrasaban considerablemente la fecha de la primera documentación, muchas veces tres, cuatro o más siglos. El alcance del problema de las dataciones se refleja muy claramente en la cantidad registrada de predataciones (el 43 %, p. ej. sólo en el fascículo 25 de 2004, Fajardo Aguirre 2006: 57). Hay que percatarse de que, con el correr de los

5 La carta fue publicada en las actas del congreso internacional *Cosmos léxico*, organizado en la Universidad de Paderborn en octubre de 2004, como homenaje al maestro (véase Arnold/Langenbacher-Liebgott 2006).

siglos, una gran parte de las palabras desapareció, pero más del 40 % del léxico general de los siglos X–XIV se mantiene vivo en la lengua actual. Según Müller (2004: 71) «[u]na pervivencia aún mayor resulta si tenemos en cuenta todo el diasistema de la lengua, incluso dialectos, hablas y el multiforme español de América». Es obvio que una gran parte del léxico no ha sido investigada en toda su extensión en el marco de la lingüística, la literatura y las ciencias de la cultura. De ello resulta el gran potencial que ofrecerá el *DEMel* no solo para los lingüistas y filólogos, sino también para los historiadores e investigadores en los campos de la teología, la medicina, la mineralogía y la botánica, la historia del derecho, la sociología o para los estudios árabes y hebraicos.

5 Primera fase del trabajo y estado actual del *DEMel*

Lo que parece evidente, pues, es que los fondos de investigación paralizados deben ser retomados de una forma rápida y productiva para el beneficio público con ayuda de las tecnologías modernas, principal objetivo planteado por el *DEMel*. Por eso este apartado se dedicará a la descripción de los trabajos ya efectuados durante el primer año del proyecto, el estado actual de su desarrollo y una visión global de las etapas planificadas.

5.1 El material del *DEMel*

5.1.1 La configuración de las fichas y su preparación para el proceso de digitalización

Como se ha expuesto, el material físico del archivo *DEM* se encuentra en gavetas, de las cuales cada una alberga por término medio cerca de 1700 fichas en papel. En lo que se refiere a la forma, las fichas del archivo *DEM* corresponden en su mayoría al formato estándar DIN A6 (10,5 x 14,8 cm). Además, hay unos pocos casos (alrededor del 1 % de la cantidad total) que son más grandes, pero están dobladas de manera que no sobresalen de las gavetas. Aparte del tamaño, las fichas, recopiladas a lo largo de muchos años, son muy heterogéneas, porque representan el resultado de un trabajo realizado por varios redactores, colaboradores y estudiantes. Por lo tanto, el material no solo comprende extractos manuscritos y fotocopias pegadas, sino también informaciones impresas. Aunque las fichas no están normalizadas (en el sentido estricto del término), se puede distinguir a grandes rasgos entre dos tipos: primero, las fichas con los lemas —estas fichas estructuran todo el lemario por orden alfabético y pueden constar de un solo lema con su(s) variante(s) gráfica(s), o de más lemas que pertenecen a una

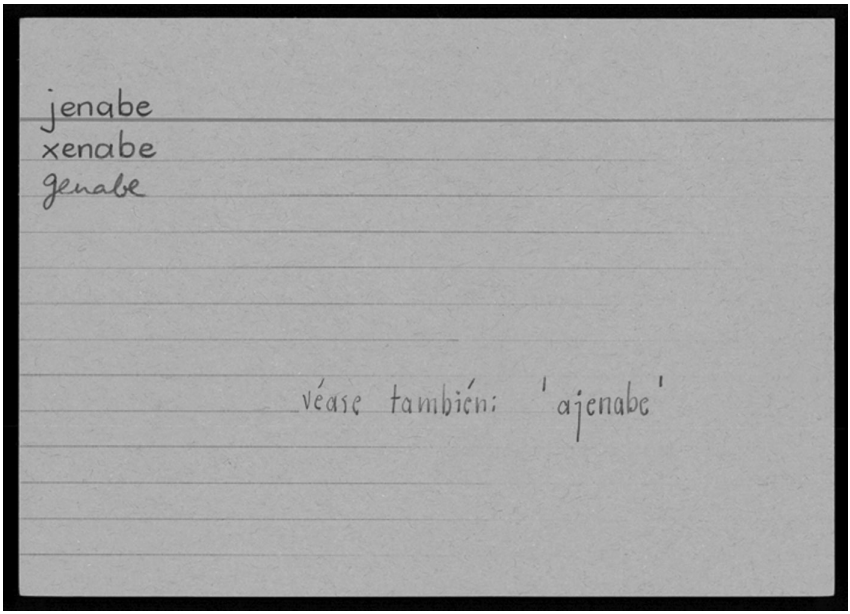


Gráfico 1a: Ficha de tipo *lema*

misma familia léxica—[6]; segundo, las fichas con las documentaciones que están relacionadas con un lema específico.

Acerca del contenido y de su disposición, las fichas que incluyen los lemas (primer tipo) tienen en general la estructura siguiente (v. gráfico 1a): a la izquierda, arriba del todo, aparece el lema[7] (aquí: *jenabe*), eventualmente seguido por otros lemas de la misma familia léxica. Las variantes gráficas se relacionan por debajo de cada lema (aquí: *xenabe* y *genabe*). En algunos casos, hay también remisiones a otros lemas o entradas (aquí: «véase también: *ajenabe*»).

En comparación con este tipo de fichas, las documentaciones (segundo tipo) contienen informaciones más detalladas (v. gráfico 1b): arriba, a la izquierda, aparece otra vez el lema en cuestión (aquí: *jenabe*), arriba, a la derecha, se puede

6 Respecto a la problemática de los conceptos *familia léxica*, *familia de palabras*, etc. véase Pena/Campos Souto (2009).

7 La lematización corresponde al español moderno. Si un lema aparece entre corchetes, es que «designa una forma lemática que no pudo ser documentada en español medieval» (Müller 1987: VII).

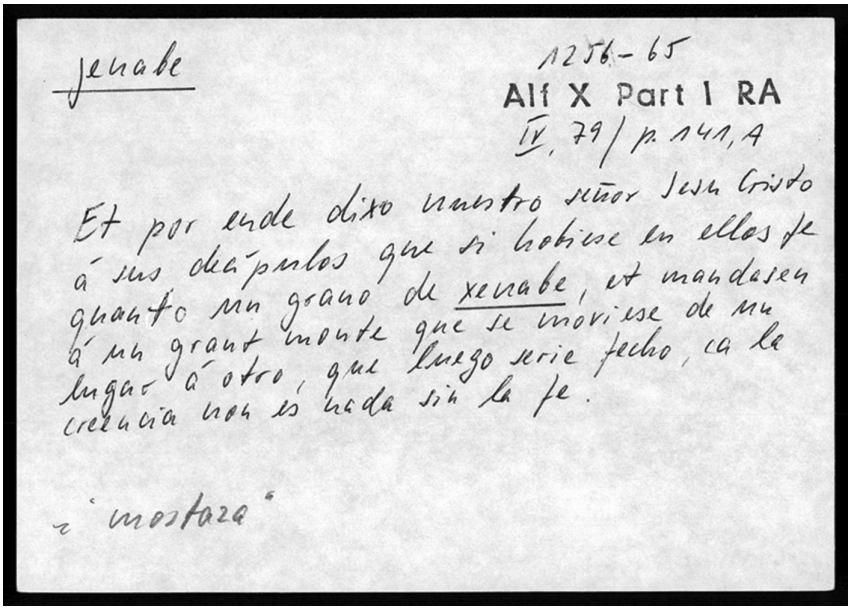


Gráfico 1b: Ficha de tipo *documentación*

ver la datación (aquí: *1256–65*), la sigla de la fuente (aquí: *Alf X Part I RA*) y, si las hay, informaciones adicionales, como en nuestro caso, el tomo, el título, la ley, la página u otras indicaciones. En la mitad de la ficha, hay normalmente un extracto del texto en el cual aparece el lema. En la mayoría de las ocasiones, esta forma documentada (aquí: *xenabe*) está señalada, es decir, subrayada o marcada con colores. En algunos casos, debajo del extracto, encontramos indicaciones sobre la acepción del lema (aquí: «*mostaza*»)[8].

Para obtener una calidad excelente de las fichas digitales que satisfaga nuestras exigencias y cumpla con los estándares de la Deutsche Forschungsgemeinschaft, encomendamos el escaneo a una empresa especializada en grandes proyectos de digitalización. Dados los recursos financieros disponibles, acordamos con la empresa que la preparación de todo el material fuera realizada por nosotros mismos para contribuir al máximo, por nuestra parte, a optimizar el rendimiento

8 Vid. Arnold *et al.* 2016: 35–37. Se trata de un significado provisional porque, aunque la lematización está completa, el análisis semántico solo se había realizado en forma definitiva en los artículos lexicográficos publicados.

del escáner y a facilitar el escaneo (v. abajo *a*), para evitar que no se escaneen los dorsos de las fichas (v. abajo *b*) o para cerciorarnos de que formatos más grandes de lo normal fueran escaneados separadamente (v. abajo *c*), puesto que:

a) Había fichas con cintas elásticas, y no eran raras las fichas grapadas o que —por diversas razones— tenían un clip. Queda claro que era necesario quitar todo lo que podía impedir el proceso de escaneo. De vez en cuando, los extractos de los textos no correspondían al formato indicado, de ahí que los papeles fueran doblados. En este caso, teníamos que cortar el exceso y crear una ficha adicional con las informaciones que sobrepasaban dicho formato;
b) A veces, las informaciones necesarias simplemente no cabían en la ficha de manera que también los dorsos fueron rotulados o pegados con trozos de texto. En estos casos, había que copiar las informaciones del dorso en formato DIN A6 para generar así una nueva ficha que, una vez escaneada, podremos identificar indudablemente como el dorso de la ficha original;
c) De vez en cuando, encontramos en el fichero papeles con informaciones adicionales acerca de una ficha, informaciones de tipo lingüístico o extralingüístico, pero en formato DIN A4 (doblados a A6). En estos casos, sacamos y archivamos estos formatos grandes que serán escaneados separadamente por la empresa.

Esta primera etapa del proyecto comenzó en noviembre de 2016 —con una fase piloto en la que hemos comprobado la calidad de aproximadamente 3600 fichas ya escaneadas y que terminó con una reunión de los equipos de trabajo de Rostock y Paderborn (filólogos e informáticos) y un representante de la empresa— y finalizó con éxito en mayo de 2017. El trabajo efectuado durante los primeros meses del proyecto consistía pues, primero, como hemos expuesto antes, en la revisión de las fichas, ficha por ficha, gaveta por gaveta, con el fin de preparar el material para el proceso de escaneo. Llevado a cabo por el equipo de Paderborn donde se encontraba instalado el archivo *DEM*, este trabajo incluyó también el cálculo de unas 865 000 fichas en total y el control de los respectivos dorsos. Paralelamente, el equipo de Rostock se dedicó entre noviembre de 2016 y mayo de 2017 a la revisión de la bibliografía del *DEM* con las siglas y dataciones para prepararla de manera que esté apta para el tratamiento informático, y además a cuestiones de coordinación y gestión relacionadas con la empresa, la elaboración de las especificaciones para la digitalización así como a la elaboración de un programa para la transferencia e integración de las copias digitales al sistema de Goobi (v. 5.1.2), la configuración de este sistema para el tratamiento de las copias digitales, la modelización de datos y la concepción de la aplicación para el registro de los mismos. Después del escaneo, el archivo *DEM* completo será trasladado a la Universidad de Rostock.

5.1.2 El tratamiento de las fichas digitales

El resultado del escaneo es un gran número de fichas digitales, es decir, una gran cantidad de copias digitales que corresponden al orden de las fichas en papel en las gavetas. Bien es verdad que las informaciones que nos interesan están representadas por las copias digitales, no obstante, hay que registrarlas electrónicamente, es decir, extraer las informaciones relevantes de las copias digitales para poder garantizar un tratamiento posterior en el ordenador. A causa de los diversos tipos de escritura (muchas letras y caracteres diferentes, textos escritos a mano o con máquina de escribir, sellos, etc.), no pareció recomendable descifrar los datos automáticamente.

Para que el futuro usuario pueda buscar en el fichero digital es necesario, en primer lugar, informatizar las anotaciones relevantes de las fichas. A tal fin, podemos beneficiarnos del programa *open-source* Goobi[9] que fue adaptado a las necesidades del *DEMel* por los informáticos de la Universidad de Rostock que forman parte del grupo de investigación. Este software permite una primera estructuración de las copias digitales para obtener segmentos que contengan un lema (y —si los hay en la ficha en cuestión— otros de la misma familia léxica) y sus documentaciones. En nuestra terminología interna, nosotros llamamos estos segmentos *secciones lemáticas* («Lemma-Abschnitte»).

Las *secciones lemáticas* difieren mucho en su extensión de dos (lema + forma documentada) a más de 1700 fichas digitales. Difiere también la cantidad de informaciones en las fichas. Actualmente, nos concentramos solo en el registro de las fichas de tipo *lema*. Como ya ha sido mencionado anteriormente, a veces aparece un solo lema en la ficha, pero más frecuentemente el tipo de fichas *lema* muestra también la(s) variante(s) gráfica(s) del lema, así como una remisión o varias a otros lemas, etc. (v. gráfico 1a). El gráfico 2 visualiza la obtención de los lemas y de las remisiones.

Después del registro de los datos, cada gaveta o fichero digital pasa a un proceso de revisión en el que todas las informaciones introducidas serán previamente controladas y, si es necesario, corregidas por los coordinadores del proyecto. En este paso, se discuten también los casos dudosos como p. ej. grafías extrañas, letras ilegibles… El resultado de los dos primeros pasos, es decir, la obtención de datos y revisión, será un listado electrónico —si bien provisional— de todos los lemas del fichero *DEM*, el lemario del *DEMel*.

9 Goobi <http://www.intranda.com/digiverso/goobi> [18/10/2017].

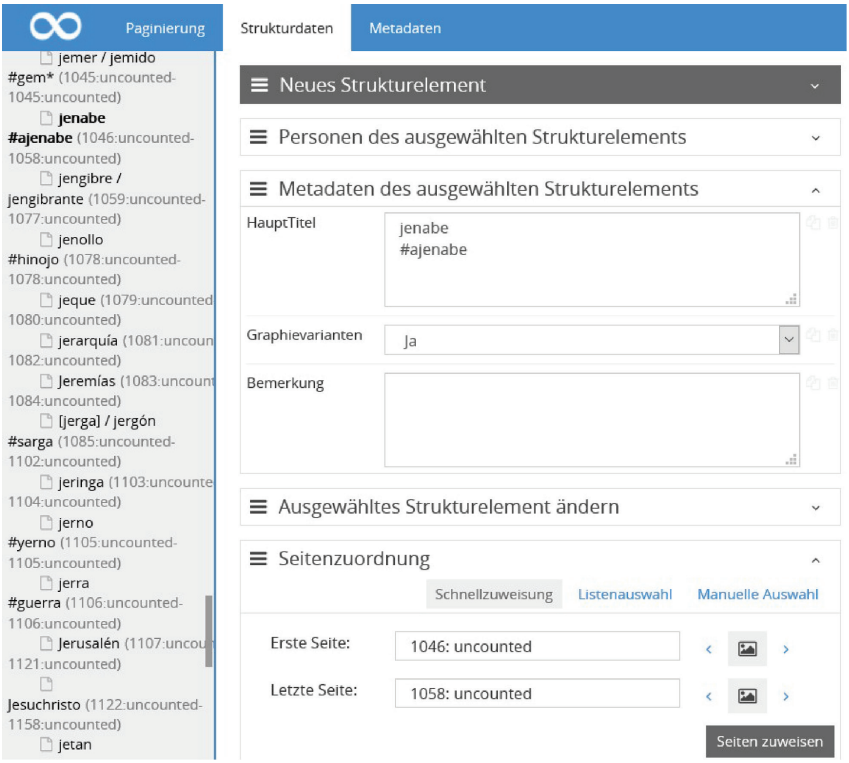


Gráfico 2: Las *secciones lemáticas* en Goobi

5.2 Estado actual del *DEMel* y su futuro

Actualmente, los lingüistas y filólogos del proyecto se encuentran todavía en la revisión de las *secciones lemáticas* y los informáticos están en plena elaboración de un editor, que servirá para el tratamiento posterior de las informaciones obtenidas. Por su parte, los informáticos se ocupan también de la creación del modelo de datos, un paso muy importante para el éxito del proyecto. Al mismo tiempo, todo el equipo de investigación está dedicándose a desarrollar una interfaz de entrada para el registro de las documentaciones (v. gráfico 3).

En general, en las fichas de tipo *documentación* se encuentran anotadas las formas de las palabras medievales en el contexto de su uso respectivo, la fuente del ejemplo y su datación. De vez en cuando, se encuentran además informaciones gramaticales, semánticas y etimológicas u otras (de manera no sistemática).

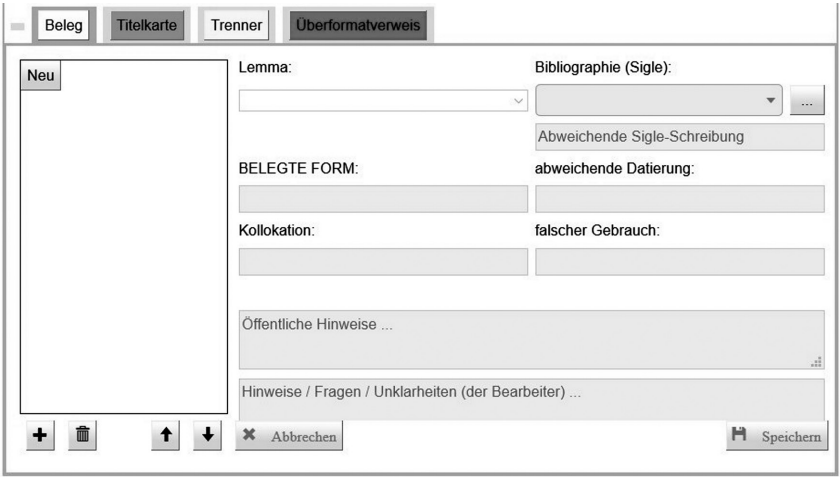


Gráfico 3: Interfaz de entrada (provisional) para el registro de las documentaciones

Una vez revisadas, se introducirán en la interfaz, para cada lema, datos como la(s) forma(s) documentada(s) y las colocaciones eventuales; además, se asignarán las formas documentadas a las siglas ya integradas antes en la programación de la interfaz. Así, las siglas preinstaladas en la interfaz están enlazadas automáticamente con la fuente y su datación. Como base de información servirá la bibliografía del *DEM*, revisada y ajustada a los requerimientos técnicos del *DEMel*.

A finales de 2017, llegarán los últimos resultados del escaneo que deberán pasar por los procesamientos mencionados. Simultáneamente, el equipo filológico empezará con la introducción de los datos en la interfaz de entrada propia y el equipo informático con la elaboración de la interfaz de búsqueda para el usuario[10].

En general, está previsto que se permita partir de una búsqueda según varios criterios como p. ej. una palabra en su forma completa o parcial (prefijos o sufijos). En la casilla «consulta» de la futura interfaz de búsqueda se escribirá el objeto de estudio, es decir, la palabra (completa o parcial). Para ello, podrán utilizarse las formas truncadas con los comodines usados universalmente (*,?). En el caso de que se quiera trabajar con algún documento específico, se podrá realizar una búsqueda en una obra concreta (según la elección de la sigla) o definir un

10 Para posibles diseños de la interfaz de búsqueda, vid. Arnold *et al.* (2016: 37–38) y Arnold/Langenbacher-Liebgott (2018).

período de tiempo en el que se quiere consultar la palabra. Con estas opciones, se le facilitará al usuario el hallazgo de datos según su interés de investigación.

En cuanto a la presentación de los resultados (es decir, el orden de las informaciones) en la aplicación web, donde se encontrarán todas las particularidades del *DEMel* (la descripción del proyecto, la bibliografía, el lemario, el manual de instrucciones, la interfaz de búsqueda, las fichas digitales, etc.), esta todavía está en estado de desarrollo. Lo que sí es seguro es que será posible para el usuario ver las fichas digitales correspondientes a su búsqueda para que pueda «hojear» las documentaciones y recibir así informaciones adicionales, p. ej. leer el extracto del texto (que no será informatizado en esta etapa del proyecto).

6 Conclusión

Como señala Müller (1987: V) en el primer fascículo del *DEM*, un «diccionario de un período histórico de la lengua […] no logrará abarcar ni reconstruir el vocabulario de la época», y, por supuesto, esto vale también para el material lexicográfico del *DEM*. Pero, naturalmente, es importante proporcionar siempre más informaciones, ya que es acertado lo que indica Kabatek (2016: 10) a propósito de que, «una lingüística histórica con una base de datos fiable más amplia es precisamente la que produce los análisis más complejos y completos de las evoluciones». En este sentido, el *DEMel* se entiende como un elemento complementario a las ya existentes obras y proyectos para obtener así una visión más completa del español medieval. Gracias al apoyo material de la Deutsche Forschungsgemeinschaft y a los valiosos consejos del *Trier Center for Digital Humanities* (Universidad de Tréveris), el *DEMel* deberá estar disponible en la red a finales de 2019.

Referencias bibliográficas

ADMYTE = Archivo Digital de Manuscritos y Textos Españoles. <http://www.admyte.com/admyteonline/home.htm> [último acceso: 20/10/2017].

Arnold, Rafael (2016): «La etimología en el *DEM*, con una breve descripción del "Fichero etimológico" en el Archivo del *DEM*», en Mariano Quirós García *et al.* (eds.), *Etimología e historia en el léxico del español. Estudios ofrecidos a José Antonio Pascual (Magister bonus et sapiens)*. Madrid: Iberoamericana/Fráncfort: Vervuert, 57–70.

Arnold, Rafael *et al.* (2016): «El *Diccionario del Español Medieval electrónico* (*DEMel*)», en María Victoria Domínguez-Rodríguez *et al.* (eds.), *Words across History: Advances in Historical Lexicography and Lexicology*.

Universidad de las Palmas de Gran Canaria: Servicio de Publicaciones y Difusión Científica, 30–39.

Arnold, Rafael/Jutta Langenbacher-Liebgott (eds.) (2006): *Cosmos léxico. Contribuciones a la lexicología y a la lexicografía hispánicas*. Fráncfort/Berlín/Berna: Lang.

Arnold, Rafael/Jutta Langenbacher-Liebgott (2018): «El caudal léxico del español medieval y el nuevo proyecto *DEM electrónico* (*DEMel*)», en Roberto Antonelli *et al.* (eds.), *Atti del XXVIII Congresso Internazionale di Linguistica e Filologia Romanza (Roma, 18-23 Iuglio 2016),* vol. 1. Estrasburgo: Éditions de Linguistique et de Philologie, 789–798.

Biblia Medieval <http://corpus.bibliamedieval.es/> [último acceso: 20/10/2017].

Biblioteca Digital de Textos del Español Antiguo. <http://www.hispanicseminary.org/textconc-es.htm>[último acceso: 20/10/2017].

Bracchi, Remo (1996): «Recensione – Bodo Müller, *Diccionario del Español Medieval*, fascículos 6–9 *acebuche-achar*», *Salesianum* 58, 419–420.

Campos Souto, Mar (2015): «El *NDHE* como muestra de la nueva lexicografía digital», *Estudios de Lexicografía* 3 (= *Monográfico sobre el Nuevo diccionario histórico de la RAE*, dirigido por José Antonio Pascual), 71–93.

Carrasco Manchado, Ana Isabel (2011): «Nuevas herramientas para la historia de la Edad Media hispánica: los corpus textuales informatizados», *En la España Medieval* 34, 343–372.

CDH = Real Academia Española: *Corpus del Diccionario histórico*. <http://web.frl.es/CNDHE/view/inicioExterno.view> [último acceso: 20/10/2017].

CE = *Corpus del español*. <http://www.corpusdelespanol.org> [último acceso: 20/10/2017].

CHARTA = *Corpus Hispánico y Americano en la Red: Textos Antiguos*. <http:corpuscharta.es> [último acceso: 20/10/2017].

Clavería Nadal, Gloria (1991): «Reseña – Bodo Müller, *Diccionario del español medieval*. Fascículos 1–5. Carl Winter-Universitätsverlag, Heidelberg, 1987–1990; 354 pp.», *NRFH* 39/2, 1102–1107.

CNDHE = Real Academia Española: *Corpus del Nuevo diccionario histórico del español*. <http://web.frl.es/CNDHE/view/inicioExterno.view> [último acceso: 20/10/2017].

CODEA+ 2015 = *Corpus de documentos españoles anteriores a 1800*. <http://corpuscodea.es/> [último acceso: 20/10/2017].

CORDE = Real Academia Española: *Corpus diacrónico del español*. <http://corpus.rae.es/cordenet.html> [último acceso: 20/10/2017].

CORPES XXI = Real Academia Española: *Corpus del español del siglo XXI*. <http://web.frl.es/CORPES/view/inicioExterno/view> [último acceso: 20/10/2017].

CREA = Real Academia Española: *Corpus de referencia del español actual*. <http://corpus.rae.es/creanet.htm> [último acceso: 20/10/2017].

Darbord, Michel (1990): «Compte rendu – Bodo Müller, *Diccionario del Español Medieval*, fascículos 1–4», *Revue de Linguistique Romane* 54, 283.

Davies, Mark (2009): «Creating useful historical corpora: A comparison of *CORDE*, the *Corpus del español*, and the *Corpus do português*», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Madrid: Iberoamericana/ Fráncfort: Vervuert, 137–166.

Davies, Mark (2017): «Compare to other corpora: for historical research (compared to *CORDE*)» <http://www.corpusdelespanol.org/compare.asp> [último acceso: 20/10/2017].

DEM = Müller, Bodo (1987–2005): *Diccionario del español medieval*, vol. 1, fascículos 1–10, vol. 2, fascículos 11–20, vol. 3, fascículos 21–26. Heidelberg: Winter.

DEMel = *Diccionario del Español Medieval electrónico*. <http://go.upb.de/DEMel/> y <http://www.romanistik.uni-rostock.de/forschung/sprachwissenschaft/DEMel/> [último acceso: 20/10/2017].

DHECan = Corrales, Cristóbal/Dolores Corbella: *Diccionario Histórico del Español de Canarias*. <http://web.frl.es/DHECan.html> [último acceso: 20/10/2017].

DHLE = Real Academia Española (1960–1996): *Diccionario histórico de la lengua española*, *a-bajoca*. Madrid: Real Academia Española. V. también <http://web.frl.es/DH.html> [último acceso: 20/10/2017].

Enrique-Arias, Andrés (2012): «Dos problemas en el uso de corpus diacrónicos del español: perspectiva y comparabilidad», *Scriptum Digital* 1, 85–106 <http://www.scriptumdigital.org/documents/04_ENRIQUE-ARIAS_Wok.pdf> [último acceso: 20/10/2017].

Fajardo Aguirre, Alejandro (2006): «La datación del léxico en la lexicografía histórica del español», en Rafael Arnold y Jutta Langenbacher-Liebgott (eds.), *Cosmos léxico. Contribuciones a la lexicología y a la lexicografía hispánicas*. Fráncfort/Berlín/Berna: Lang, 51–59.

Goobi <http://www.intranda.com/digiverso/goobi> [último acceso: 18/10/2017].

HSMS = *Hispanic Seminary of Medieval Studies*. <http://www.hispanicseminary.org/intlang.htm> [último acceso: 20/10/2017].

Kabatek, Johannes (2016): «Un nuevo capítulo en la lingüística histórica iberorrománica: el trabajo crítico con los corpus. Introducción a este volumen», en Johannes Kabatek (ed.) con la colaboración de Carlota de Benito Moreno, *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín/Boston: De Gruyter, 1–17.

Kasten, Lloyd A./Florian J. Cody (eds.) (2001[2]): *Tentative Dictionary of Medieval Spanish*. Nueva York: Hispanic Seminary of Medieval Studies.

Kasten, Lloyd A./John J. Nitti (eds.) (2002): *Diccionario de la prosa castellana del Rey Alfonso X*, 3 vols. Nueva York: Hispanic Seminary of Medieval Studies.

Lapesa, Rafael (1972): «Carta de Rafael Lapesa a Bodo Müller», *Archivo DEMel*. Rostock: Universidad de Rostock, 1–5.

Metzeltin, Miguel (1992): «Spanisch – Etymologie und Geschichte des Wortschatzes/Español – Etimología e historia del léxico», en Günter Holtus, Michael Metzeltin y Christian Schmitt (eds.), *Lexikon der Romanistischen Linguistik*, vol. VI/1. Tubinga: Niemeyer, 440–457.

Müller, Bodo (1987–2005): *Diccionario del español medieval*, vol. 1, fascículos 1–10, vol. 2, fascículos 11–20, vol. 3, fascículos 21–26. Heidelberg: Winter.

Müller, Bodo (2003): «Etymologische und wortgeschichtliche Erforschung und Beschreibung der romanischen Sprachen: Spanisch», en Gerhard Ernst *et al.* (eds.), *Romanische Sprachgeschichte*, vol. 1. Berlín/Nueva York: De Gruyter, 376–396.

Müller, Bodo (2004): «Aspectos del léxico medieval desde la perspectiva del *Diccionario del español medieval* (*DEM*)», en Jens Lüdtke y Christian Schmitt (eds.), *Historia del léxico español: enfoques y aplicaciones. Homenaje a Bodo Müller*. Madrid: Iberoamericana/Fráncfort: Vervuert, 61–72.

NDHE = Nuevo Diccionario Histórico del Español. <www.frl.es> [último acceso: 20/10/2017].

Neumann-Holzschuh, Ingrid (1996): «Compte rendu – Bodo Müller, *Diccionario del Español Medieval,* fascículo 9 *acreer – achar*, Heidelberg 1993; fascículo 10 *achar – además*. Heidelberg 1994», *Revue de Linguistique Romane* 40, 581–582.

Octavio de Toledo y Huerta, Álvaro Sebastián (2016): «Aprovechamiento del *CORDE* para el estudio sintáctico del primer español moderno (ca. 1675–1825)», en Johannes Kabatek (ed.) con la colaboración de Carlota de Benito Moreno, *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín/Boston: De Gruyter, 57–89.

Pascual, José Antonio (2015): «Introducción a una celebración lexicográfica: a propósito de la reciente publicación de un millar de palabras del *NDHE* en el

portal de la RAE», *Estudios de Lexicografía* 3 (= *Monográfico sobre el Nuevo diccionario histórico de la RAE*, dirigido por José Antonio Pascual), 7–13.

Pascual, José Antonio/Mar Campos Souto (2017): «Presentación». <http://web.frl.es/DH/ayuda/presentacion.html> [último acceso: 20/10/2017].

Pena, Jesús/Mar Campos Souto (2009): «Propuesta metodológica para el establecimiento de familias léxicas en una consideración histórica: el caso de *hacer*», *Cuadernos del Instituto Historia de la Lengua* 2, 21–51.

Pharies, David (1991): «Review – *Diccionario del español medieval* by Bodo Müller, fascículos 1–4», *Hispanic Review* 59/1, 79–80.

Pilsel, Drago (2005): «Un profesor de Heidelberg elabora un magno diccionario de español medieval. El profesor Bodo Müller lleva trabajando 30 años en la obra», *El País*, 2 de enero. <http://elpais.com/diario/2005/01/02/cultura/1104620403_850215.html> [último acceso: 20/10/2017].

Pruvost, Jean (2005): «Quelques concepts lexicographiques opératoires à promouvoir au seuil du XXIe siècle», *Ela. Études de linguistique appliquée* 137, 7–37.

Quemada, Bernard (1987): «Notes sur *lexicographie* et *dictionnairique*», *Cahiers de lexicologie* 51, 229–242.

Quemada, Bernard (1990): «La nouvelle lexicographie», en Maria Teresa Cabré *et al.* (eds.), *La Lingüística Aplicada* 9 (= *Noves perspectives/noves professions/noves orientacions*). Barcelona: Fundació Caixa de Pensions/Universitat de Barcelona, 55–78.

Real Academia Española (2017a): «*CDH*: Corpus nuclear del *NDHE*». <http://www.rae.es/recursos/banco-de-datos/cdh> [último acceso: 20/10/2017].

Real Academia Española (2017b): «Nuevo diccionario histórico: Corpus del *NDHE*». <http://www.rae.es/recursos/diccionarios/nuevo-diccionario-historico> [último acceso: 20/10/2017].

Reinhardt, Jan (2014): «Iberoromance I: Historical and etymological lexicography», en Rufus H. Gouws *et al.* (eds.), *Dictionaries. An International Encyclopedia of Lexicography*, vol. 4 (= *Supplementary Volume with Focus on Electronic and Computational Lexicography*). Berlín/Boston: De Gruyter, 636–645.

Rodríguez Barcia, Susana (2016): *Introducción a la lexicografía*. Madrid: Síntesis.

Rojo, Guillermo (2006): «[Carta a Bodo Müller]», en Rafael Arnold y Jutta Langenbacher-Liebgott (eds.), *Cosmos léxico. Contribuciones a la lexicología y a la lexicografía hispánicas*. Fráncfort/Berlín/Berna: Lang, IX.

Rojo, Guillermo (2010): «Sobre codificación y explotación de corpus textuales: otra comparación del *Corpus del Español* con el *CORDE* y el *CREA*», *Lingüística* 24, 11–50.

Rojo, Guillermo (2014): «Hispanic Corpus Linguistics», en Manel Lacorte (ed.), *The Routledge Handbook of Hispanic Applied Linguistics*. Nueva York: Routledge, 371–387.

Rojo, Guillermo (2016a): «*Citius, maius, melius*: del *CREA* al *CORPES XXI*», en Johannes Kabatek (ed.) con la colaboración de Carlota de Benito Moreno, *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín/Boston: De Gruyter, 197–212.

Rojo, Guillermo (2016b): «Corpus textuales del español», en Javier Gutiérrez-Rexach (ed.), *Enciclopedia de lingüística hispánica*, vol. 2. Londres/Nueva York: Routledge, 285–296.

Ruiz Mantilla, Jesús (2017): «Una donación de Inditex resucita los trabajos del diccionario histórico», *El País*, 4 de febrero. <https://elpais.com/cultura/2017/02/03/actualidad/1486135672_668793.html> [último acceso: 20/10/2017].

Salas Quesada, Pilar/Abelardo Torres Morcillo (2015): «Aproximación a los fundamentos del *NDHE* a través de las herramientas informáticas usadas en su elaboración y presentación», *Estudios de Lexicografía* 3 (= *Monográfico sobre el Nuevo diccionario histórico de la RAE*, dirigido por José Antonio Pascual), 15–69.

Sánchez Sánchez, Mercedes/Carlos Domínguez Cintas (2007): «El banco de datos de la Real Academia Española: *CREA* y *CORDE*», *Per Abbat* 2, 137–146.

Torruella, Joan (2016): «Tres propuestas en el ámbito de la lingüística de corpus», en Johannes Kabatek (ed.) con la colaboración de Carlota de Benito Moreno, *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín/Boston: De Gruyter, 90–112.

Verd Conradi, Gabriel María (1989): «Reseña – Bodo Müller, *Diccionario del español medieval*, fascículos 3–4», *Archivo Teológico Granadino* 52, 361.

Pilar García Mouton

Geolingüística y Humanidades digitales: el *Atlas Lingüístico de la Península Ibérica* (*ALPI*)

Resumen: Este trabajo presenta los avances del proyecto que supone la edición definitiva, en soporte digital, del *Atlas Lingüístico de la Península Ibérica* (*ALPI*) de Tomás Navarro Tomás, proyecto que integra una obra histórica de la Geolingüística románica en las Humanidades digitales. A partir de los datos accesibles en <alpi.csic.es> relativos a los nombres del jabalí, se ejemplifican algunas de las posibilidades de la herramienta *ALPI-CSIC* para estudiar el léxico.

Palabras clave: Geolingüística, Humanidades digitales, *Atlas Lingüístico de la Península Ibérica*, *ALPI*, Tomás Navarro Tomás, nombres del jabalí

Abstract: This paper presents the advances of the project that involves the definitive edition, in digital support, of the Tomás Navarro Tomás's *Linguistic Atlas of the Iberian Peninsula* (*ALPI*), a project that integrates a historical work of the Romance Geolinguistics in the Digital Humanities. Using the data accessible at <alpi.csic.es> regarding the names of the wild boar, some of the possibilities of that the *ALPI-CSIC* tool offers to study the lexicon are exemplified.

Keywords: Geolinguistics, Digital Humanities, *Linguistic Atlas of the Iberian Peninsula, ALPI*, Tomás Navarro Tomás, names of the wild boar

1 Introducción

La invitación de los coordinadores de este libro nos permite mostrar algunas posibilidades de la edición definitiva del *Atlas Lingüístico de la Península Ibérica* (*ALPI*) para trabajar con sus datos léxicos y cartografiarlos[1]. Como es sabido, por su metodología las obras geolingüísticas son especialmente útiles para estudiar el léxico (García Mouton 1990, 2016), ya que sus mapas proporcionan información rigurosa sobre la existencia o la ausencia, la extensión y la vitalidad de una voz (Alvar 1982, Corrales/Corbella 2002–2004). Y, a partir de su distribución en el

1 Evidentemente el *ALPI* ofrece también datos fonéticos, morfosintácticos y etnográficos.

espacio, se pueden obtener conclusiones determinantes acerca de la historia de una palabra y su etimología.

El interés fundamental del *ALPI* radica en su condición de atlas supranacional con una red de encuesta que cubre con 527 puntos España[2], Portugal y la zona catalanohablante del Rosellón. A partir de preguntas generales, como corresponde a un atlas de gran dominio, recoge la forma de hablar de personas que vivían y trabajaban en el campo, con la instrucción propia de su entorno, en una época que consideró como tesoros lingüísticos unas hablas rurales prácticamente desconocidas. Conviene recordar que el *ALPI* fue un proyecto de Ramón Menéndez Pidal y su equipo del Centro de Estudios Históricos de la Junta para Ampliación de Estudios (Pérez Pascual 2017), dirigido en los años 30 por Tomás Navarro Tomás e interrumpido por la guerra civil, y que, si bien las encuestas se retomaron a fines de los años 40 y se terminaron, la publicación se detuvo después de un primer volumen con 75 mapas de fonética ([Navarro Tomás] 1962), de modo que casi todos los materiales quedaron inéditos[3].

Desde los años 50, comenzaron a hacerse en España otros atlas lingüísticos[4] (García Mouton 2009), dedicados a determinadas regiones, a dominios reducidos y a variedades de las distintas lenguas españolas, pero no se proyectó ningún otro atlas que ofreciese una visión tan amplia ni llenase el vacío que dejó el *ALPI*. De ahí su valor como marco para estudiar las variedades románicas peninsulares y como punto de comparación anterior a los grandes cambios demográficos y culturales de la posguerra.

2 La edición definitiva

Este valor histórico se aprecia en los trabajos basados en los mapas de su primer volumen y en los materiales inéditos (Navarro Tomás 1975, Fernández-Ordóñez 2011, Pato 2011). Desde el año 2007, un proyecto intramural del CSIC[5], que coordino, utiliza las nuevas tecnologías combinadas con los Sistemas de Información Geográfica para publicarlo definitivamente en un soporte mucho más accesible que el de los atlas editados en papel. El objetivo del equipo, del que

2 Exceptuando casi todo el País Vasco y las islas Canarias.

3 Hasta que David Heap (2002) localizó, fotocopió los cuestionarios del *ALPI* y comenzó a colgarlos en la página web alpi.ca.

4 Manuel Alvar, el impulsor de estos estudios, dirigió la mayor parte de los atlas regionales.

5 *Elaboración y edición de los materiales del Atlas Lingüístico de la Península Ibérica (ALPI),* referencia 200410E604, 2009–2014, IP Pilar García Mouton.

forman parte Inés Fernández-Ordóñez (Universidad Autónoma de Madrid-RAE), David Heap (University of Western Ontario), Maria Pilar Perea (Universitat de Barcelona), João Saramago (Centro de Lingüística, Universidade de Lisboa) y Xulio Sousa (Instituto da Lingua Galega, Universidade de Santiago de Compostela), es ofrecer en línea todos los contenidos del *ALPI* con la posibilidad de hacer búsquedas en ellos y de cartografiar los resultados, algo con lo que Navarro Tomás no podía soñar cuando planeó la obra (García Mouton 2017).

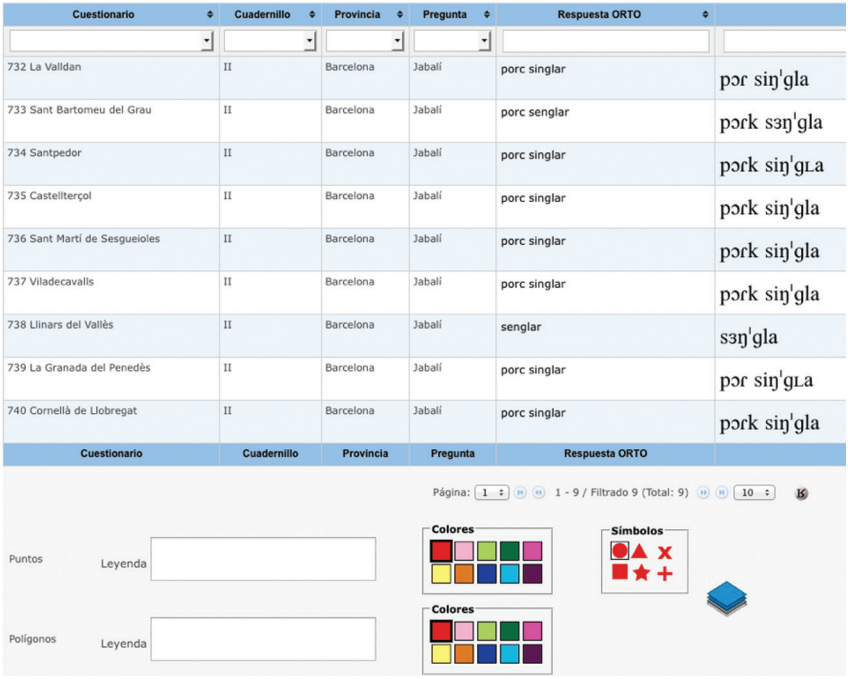
Con ese fin se diseñó una herramienta informática a la medida[6] para incorporar los datos de los 1050 cuadernos de encuesta que componen los cuestionarios, una herramienta que relaciona, con la técnica de las «migas de pan», el nombre y el número de pregunta y las respuestas correspondientes con todo lo que apuntaron los encuestadores en sus cuestionarios: refranes, anotaciones marginales, dibujos y cualquier otra información.

Como hemos explicado en otras ocasiones (García Mouton 2010), tomamos la decisión de volcar las transcripciones del equipo entrenado por Navarro Tomás en el uso del Alfabeto Fonético de la *Revista de Filología Española* (ARFE), creado por él mismo y muy detallado, al Alfabeto Fonético Internacional (IPA), asumiendo las renuncias que suponía, conscientes de que con ello ampliábamos las posibilidades de acceso a su contenido por parte de los especialistas y de otros tipos de usuarios. Se pudo tomar esta decisión con responsabilidad científica porque las nuevas tecnologías permiten ofrecer a quien necesite ver la transcripción original la misma imagen que utilizamos para introducir una respuesta en la base de datos. Para ello fue necesario digitalizar los cuestionarios. A falta de contabilizar las últimas, se digitalizaron 16 541 imágenes en dos formatos, PDF y JPG, que ocupan 5,823 GigaBytes. Esta información se conserva en PDF de gran calidad, transformada a JPG en la herramienta. Cada una de las imágenes recoge dos páginas de los textos originales. Las imágenes se organizaron internamente y se colocó cada pregunta para poder mostrar en la web toda la información de la cuestión que se seleccione.

Desde el principio se apostó por el software libre (GPL), con programas como Notepad++, Quantum GIS e Inkscape en el desarrollo de la aplicación, el diseño cartográfico y la simbología[7]. Para almacenar los datos se empleó la base de datos PostgreSQL con los módulos Postgis, que le aportan funcionalidad geoespacial y, como lenguaje de programación para las herramientas de trabajo y de servicio al usuario final, PHP5 + HTML + JavaScript + CSS.

6 En los comienzos del proceso de definición informática intervino Juan Carlos Martínez y, en todo su desarrollo, Ángel Díaz del Castillo.

7 Contamos con un servidor de páginas web Apache y un servidor de aplicaciones Tomcat.



Cuestionario	Cuadernillo	Provincia	Pregunta	Respuesta ORTO	
732 La Valldan	II	Barcelona	Jabalí	porc singlar	pɔr siŋˈgla
733 Sant Bartomeu del Grau	II	Barcelona	Jabalí	porc senglar	pɔrk sɜŋˈgla
734 Santpedor	II	Barcelona	Jabalí	porc singlar	pɔrk siŋˈgʟa
735 Castellterçol	II	Barcelona	Jabalí	porc singlar	pɔrk siŋˈgla
736 Sant Martí de Sesgueioles	II	Barcelona	Jabalí	porc singlar	pɔrk siŋˈgla
737 Viladecavalls	II	Barcelona	Jabalí	porc singlar	pɔrk siŋˈgla
738 Llinars del Vallès	II	Barcelona	Jabalí	senglar	sɜŋˈgla
739 La Granada del Penedès	II	Barcelona	Jabalí	porc singlar	pɔr siŋˈgʟa
740 Cornellà de Llobregat	II	Barcelona	Jabalí	porc singlar	pɔrk siŋˈgla
Cuestionario	Cuadernillo	Provincia	Pregunta	Respuesta ORTO	

Imagen 1: Resultados de una búsqueda con las dos opciones de cartografiado

A lo largo de una serie de reuniones periódicas, el equipo definió los campos de consulta, el teclado virtual y una serie de ayudas con las equivalencias entre los símbolos ARFE y los símbolos IPA, además de teclas rápidas que agilizan el proceso de retranscripción fonética en la aplicación que sirve para introducir vía web los materiales en la base de datos. Evidentemente hubo que incorporar la fuente DoulosSILR para visualizar los textos fonéticos en la aplicación y en el servidor de mapas, y se recurrió a una aplicación que gestiona la información para agilizar el proceso de revisión por parte de los responsables. También se han diseñado herramientas internas para cortar los dibujos que los encuestadores hicieron en los cuestionarios, relacionarlos con las respuestas y ordenarlos. Un software adaptado permite navegar por las páginas de los cuadernos de encuesta y mostrar el cuestionario original al usuario, así como hacer búsquedas en los contenidos y generar mapas de forma dinámica a partir del resultado[8]. Los

8 Los formatos SVG y PNG facilitan la consulta en dispositivos móviles.

Imagen 2: Inicio de la página web *ALPI* <alpi.csic.es>

mapas se pueden generar en dos formatos distintos: por puntos, con distintos símbolos y colores combinables para reflejar una elaboración previa, o por polígonos de Thiessen en varios colores.

Con una parte sustancial de los materiales elaborada, en mayo de 2016 abrimos en el CSIC la página web <alpi.csic.es> del *ALPI* en catalán, español, gallego, inglés y portugués. Allí, junto a otros contenidos relativos a la historia del atlas, su equipo, la metodología, los resultados, la galería fotográfica, las publicaciones, etc., el apartado *Consulta mapas* reúne todas las informaciones relacionadas con estas diez cuestiones en la red de encuesta completa: 458 *Guisantes*, 493 *Mariquita*, 496 *Lagartija*, 502 *Aguzanieves*, 520 *Jabalí*, 600a *Cadera*, 704 *Cuna*, 751 *Llevar a cuestas*, 753a *Dar volteretas* y 825 *Levadura*.

En esta primera entrega, desde cualquier ordenador se pueden hacer consultas relativas a uno o más lugares de la red de encuesta, a una o más provincias, por zonas o en la red completa. Los resultados se obtienen en transcripción fonética (IPA) o en la ortografía del área lingüística, con el aliciente de que se puede acceder a la imagen de la respuesta tal como el encuestador la transcribió (en ARFE).

Editar el atlas en soporte digital y en el Alfabeto Fonético Internacional exigió desde el primer momento un riguroso trabajo de elaboración y retranscripción,

a lo largo del que se incorporó a la base de datos cualquier información útil para el usuario, que el soporte relaciona con los contenidos propiamente lingüísticos. Esto permite búsquedas inimaginables en un atlas lingüístico convencional. Por otra parte, hay que insistir en que la consulta no acaba con la obtención de un listado; según sus intereses, el usuario puede ir refinando sus búsquedas unas sobre otras y cartografiar el resultado final.

Gracias a un proyecto de la Fundación BBVA, financiado en la primera edición de las Ayudas a Proyectos de Investigación en Humanidades digitales, del que disfrutamos los años 2015 y 2016, está prácticamente terminada la elaboración de los materiales correspondientes a 103 preguntas que pronto se incorporarán al apartado *Consulta mapas* de la web. Estas preguntas[9], seleccionadas por su interés lingüístico, dan forma a un pequeño *ALPI* interactivo que tendrá las mismas prestaciones de la obra definitiva.

3 Un ejemplo: la pregunta 520ª, *jabalí*

La pregunta 520a, que recoge los nombres del *jabalí* en toda la red de encuesta, una de las diez preguntas actualmente accesibles en la web, sirve para ejemplificar algunas posibilidades de la herramienta *ALPI*. De esta pregunta interesan varios aspectos, entre otros, la extensión real del arabismo, la adaptación de su *-í* final a la morfología romance y un procedimiento de nominación recurrente que combina distintas bases léxicas con adjetivos (García Mouton 1997).

La lengua científica tiene que ser unívoca, no admite ambigüedad, como le ocurre hasta cierto punto a la lengua culta, que suele consagrar un nombre por

9 Serán 44 preguntas de Fonética *(abeja, ahogarse, azada, caballo, asa, avispa, castillo, brazo, caja, cazador, cereza, clavo, coz, los domingos, encía, enero, escuchar, flor, fuente, fuerza, grano, hierro, hierba, hijo, hoja, jornal, los gatos, leche, leña, liebre, llave, llorar, mucho, muerte, molino, ojo, pecho, sed, rabia, sudor, tomarlo, trébedes, uncir, yerno);* 15 de Morfosintaxis *(gatito, gatazo, no saben freír un huevo, sus corderos están en nuestro prado, traje demasiada cebada, cantad una jota, os vais a caer, ¿se te calmó el dolor?, al padre le vieron llorando, a la madre no le dieron limosna, nos verá cuando vengamos, ¿a qué hora has llegado?, si tuviera dinero, lo compraría, estaba medio dormida, lo queréis para vosotros)* y 42 de Léxico *(guisantes, mariquita, lagartija, aguzanieves, jabalí, cadera, cuna, llevar a cuestas, dar volteretas, levadura, yugo, arco iris, vía láctea, amapola, sanguijuela, urraca, gallina clueca, semental de la cerda, vivienda del cerdo, cría de la vaca, comida del cerdo, nombre del cerdo, vejiga, ombligo, sobaco, meñique, difteria, hijo menor, formas infantiles del nombre de la abuela, modo de llamar al cerdo, voz del asno, grieta en la pared, modo de llamar al charlatán, columpio, gallina ciega, jugar a esconderse, caer de bruces, el arado, el escardillo, atabladera, colmena, cepillo (de carpintero)).*

concepto, el que incorpora a los diccionarios. El lenguaje popular, en cambio, mantiene una variedad léxica grande, por ejemplo a la hora de nombrar a los animales que no pertenecen al ámbito doméstico, mientras que los que tienen utilidad clara reciben un nombre general: *caballo, vaca, toro, oveja, caballo, gallina,* etc. Sus denominaciones varían dependiendo de que el animal sea más o menos conocido, doméstico y rentable a la comunidad que lo nombra. Aunque el jabalí (SUS SCROFA) pertenece al grupo de los animales salvajes, o silvestres, resulta conocido por su abundancia, su parecido con el cerdo, el interés que despierta como pieza de caza y los destrozos que causa a los campesinos.

Jabalí es uno de los pocos zoónimos árabes casi generales en la Península que han pasado a la lengua culta[10]. De hecho, el *Diccionario de la lengua española* (*DLE*) reenvía los demás nombres reconocidos del animal —*puerco, puerco jabalí, puerco montés, puerco salvaje* y *cochino montés*— a *jabalí,* como forma principal. En principio, jabalí (Corriente 2003[2]) era un adjetivo continuador del árabe hispánico *ǧabalí* 'montés', 'montaraz', que en árabe calificaba al sustantivo *hinzîr* 'cerdo'. Del mismo étimo derivan el portugués *javalí* y el gallego *xabaril,* mientras que el catalán *senglar* comparte un étimo latino con el francés *sanglier* y con el italiano *cinghiale* (< (PORCUS) SINGULARIS).

Los datos del *ALPI* evidencian un semantismo transparente similar en casi todos los nombres del animal: una base 'cerdo' (*porco*, *puerco, porc, tocino, bacó, bácoro, cochino, cocho, cerdo, marrano, marrão, gorrino, guarro,* etc., según las zonas) a la que se añade un adjetivo que lo opone al doméstico: 'salvaje' (*cerval, salvatge, suatge*), 'de monte' (*montés, montês, muntanés, muntanyés, jabalí*), 'del bosque' (*feréstic, fréstic, feréstec*), 'fiero' (*fero, fiero, fer*), 'bravo' (*bravo*), 'solo' (*singlar, senglar*), 'pinchudo' (*espinho*), etc. Este recurso de oponer lo de monte, silvestre o asalvajado a lo doméstico se utiliza en otras lenguas, como en euskera, *basaserri*; en alemán, *wildschwein*; en inglés, *wild boar* o en rumano, *porc sălbatic*.

De esta constatación y de la distribución de las formas en el mapa, se deduce el camino que debió seguir *jabalí* para introducirse en las hablas románicas: en un primer momento entró como adjetivo prestado con el sentido de 'montés', 'salvaje' y, después, perdida la carga adjetival, se desvinculó del sentido etimológico y se lexicalizó hasta prescindir del sustantivo base[11]. A ello debieron contribuir

10 La actividad agrícola de árabes y moriscos justifica la abundancia de arabismos en el léxico de la agricultura; en cambio, los zoónimos árabes son escasos, salvo en el caso de las aves de caza de altanería.

11 En textos literarios clásicos conviven *puerco salvaje* o *puerco jabalí*; *jabalí* o *puerco montés* y *puerco de monte* o *jabalí*; en Chile se oye *chancho jabalí;* en Andalucía, *cochino jabalí*.

los conocidos procesos eufemísticos en la historia de las denominaciones del cerdo, como apuntan Corominas y Pascual (*DECH*): parece que *marrano*, del árabe *máhram* 'cosa prohibida', habría sustituido a *puerco* y, a su vez, habría sido sustituido por *cochino*. A medida que estas voces se fueran percibiendo como groseras, la redenominación volvería a ponerse en marcha y, de ese modo, se generalizaron *guarro* y *gorrino* en el s. XVIII y se creó un eufemismo culto, *cerdo,* que se superpuso a todos los demás nombres[12].

En el *ALPI* los casos en los que *jabalí* y sus variantes aparecen en lexías compuestas se concentran en la mitad sur peninsular, presumiblemente porque allí el arabismo mantuvo más tiempo su valor adjetivo. Los más septentrionales se localizan en Badajoz, *cochino jabalín* (373); Ciudad Real, *guarro jabalín* (479)[13]; Córdoba, *guarro jabalín* (501), *marrano jabalín* (504); Albacete, *gorrino jabalís* (482), *gorrino jabalí* (485), *marrano jabalí* (488); Castellón, *cerdo jabalí* (758), *porc jabalí* (762); Valencia, *porc xabalí* (765, 772, 774, 775), *porc jabalí* (780). El resto, más al sur: en Alicante, *porc jabalí* (781, 783, 784, 787), *cochino jabalí* (790); Cádiz, *cochino jabalín* (533, 535) y Granada, *marrano jabalí* (546, 549, 550). No se encuentran en la mitad norte, donde en cambio abundan construcciones románicas equivalentes: *porco bravo, cocho bravo* y *porco montés* dominan en Galicia y *porc fer*, *porc senglar, porc singlar,* en Cataluña, donde *senglar, singlar*[14] sufrió el mismo proceso que *jabalí* y acabó por despojarse de la base 'cerdo' para nombrar, ya no como adjetivo, al cerdo salvaje, lo mismo que ocurrió con *sanglier* en francés y *cinghiale* en italiano.

3.1 La verdadera extensión de *jabalín*

Desde la Edad Media se documenta *jabalín* como variante de *jabalí.* Parece evidente que la lengua popular trató de integrar en el sistema romance la terminación *-í* del adjetivo árabe: añadiéndole una *-n*, *jabalín,* como en *celemín* (Lapesa

12 Entre cazadores es normal llamar *cochino* o *guarro* al jabalí. Hace unos años, para insultarse, los niños andaluces se canturreaban: «Cochino, marrano, puerco, sevillano…». Las rimas infantiles muestran esa sinonimia: «Pato i ganso i ansarón/tres cosas suenan i una son;/cochino i puerco i lechón/otras tres en una son» (Margit Frenk 1987: 1007).

13 El *Atlas Lingüístico y etnográfico de Castilla-La Mancha* (*ALeCMan*), gracias a su red de encuesta más densa, documenta en el mapa 75 este tipo de ejemplos hasta el nordeste de Cuenca y el sur de Toledo: *guarro jabalí* en Ciudad Real, y en Cuenca y Albacete, *gorrino jabalí*, *gorrino jabalín* y *gorrino jabalís*.

14 Se puede observar en lugares cercanos entre sí: *porc senglar* (725, 726, 728, 731, 733), *senglar* (707, 727, 729, 738); *porc singlar* (706, 713, 720, 722), *singlar* (701, 702, 703, 704, 705, 718).

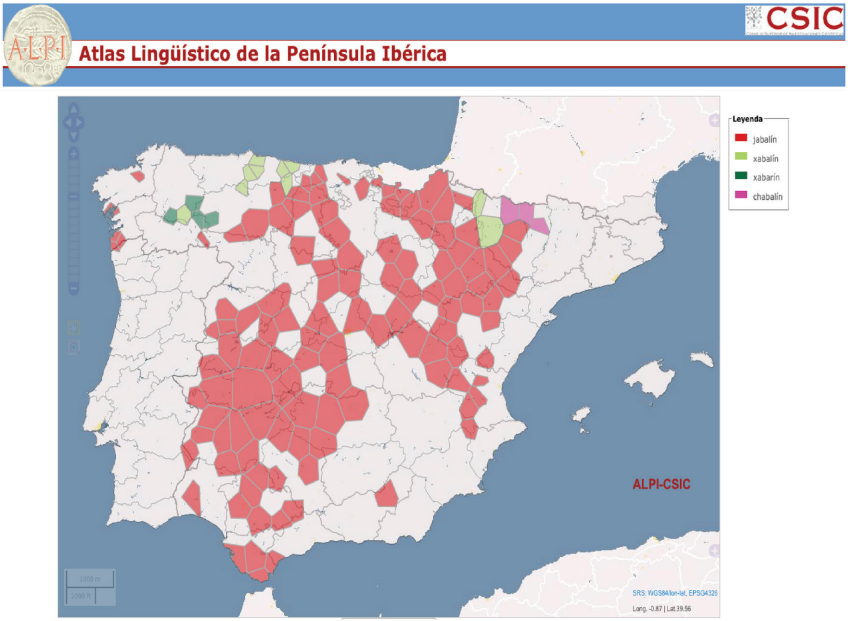


Imagen 3: Mapa (con polígonos) de la extensión de *jabalín* [(rojo), *xabalín* (verde claro), *xabarín* (verde intenso), *chabalín* (rosa)]

1981: 143); una *-l, jabaril,* como en *albañil;* e, incluso, una *-s, jabalís*[15]. Este tipo de adaptación se encuentra también en gallego, *xabaril, xebarín, xabalín, jabalín*; en portugués, *javaril*; en asturleonés, *xabaril, xabalín, jabalín, jabaril*; o en aragonés, *xabalí, xabalín, chabalín*. Sobre *jabalín* se formó el femenino *jabalina*, que ya recoge el primer diccionario académico (1734). Actualmente el *DLE* localiza *jabalín* en Andalucía y Salamanca como palabra poco usada.

Y, sin embargo, el *ALPI* deja constancia de que *jabalín* tenía en la primera mitad del siglo XX bastante uso en el medio rural y que su ámbito no se limitaba a Andalucía y Salamanca. Probablemente se asentó con el apoyo del sufijo diminutivo *-ín, -ino*, vinculado a las hablas asturleonesas y a sus zonas de expansión, aunque *jabalín* también tiene presencia oriental.

15 En la zona que actualmente conserva *jabalís,* el femenino tiende a ser *jabalisa.* El *ALeCMan* documenta *jabalís* (m.) – *jabalisa* (f.) (CU 605, AB 306, 307, 309), *jabalí* – *jabalisa* (GU 400, AB 103, 211, 405, 600), *jabalisa* (CR 408), *gorrino jabalís* – *gorrina jabalisa* (AB 310), *gorrino jabalí* – *gorrina jabalisa* (AB 208, 404).

Imagen 4: Mapa (con puntos) de la extensión de *jabalín* (rojo), *xabalín* (verde claro), *xabarín* (verde intenso), *chabalín* (rosa)

Jabalín no pasó a la lengua escrita, que optó por *jabalí*[16] y su plural *jabalíes*; en cambio, *jabalina* sí lo hizo, como el femenino más difundido, probablemente apoyado por sus casi sinónimos *gorrina* y *cochina*.

Los mapas evidencian que *jabalí* y *jabalín* cubren casi todo el dominio de expansión del castellano. Y su adaptación a la fonética de las diferentes áreas lingüísticas ratifica la antigüedad del arabismo.

3.2 *Jabato* y *javardo*: geografía lingüística y etimología

Para concluir, el mapa se completa por el sudoeste con otras dos formas asentadas en aparente continuidad: en el sur de Portugal llaman *javardo* al jabalí y, en las tierras linderas de Huelva y Sevilla, *jabato.* Aunque habitualmente *jabato*

16 El *ALeCMan* documenta el femenino *jabalía* (GU 309, AB 312) y *gorrina jabalía* (CU 609).

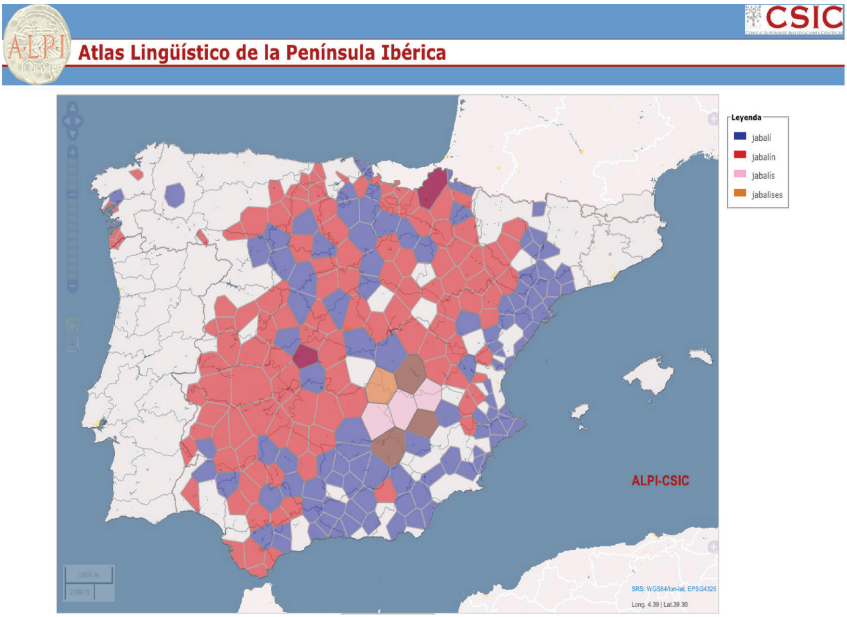


Imagen 5: Mapa de la extensión de *jabalí* (azul), *jabalín* (rojo), *jabalís* (rosa), *jabalises* (marrón)[17]

se refiera a la cría de jabalí, en bastantes puntos de la Andalucía occidental, en Huelva (517, 518, 520, 521, 522) y en Sevilla (523, 526, 531), nombra al jabalí adulto[18]. La cercanía entre *jabato*, con un sufijo *-ato* productivo en las lenguas románicas occidentales para dar nombre a crías de animales (Pharies 2002: 122–123), y el portugués *javardo* merecería un estudio, como ya apuntaron Corominas y Pascual s.v. *jabardo,* voz que en castellano está semánticamente ocupada como 'enjambre pequeño que se separa de una colmena'.

17 En alguna ocasión, *jabalín* convive como segunda respuesta con *jabalí*. Es lo que ocurre en los puntos con un rojo más intenso, resultante de la superposición de rojo y azul.

18 En los tres casos sevillanos, como segunda respuesta tras *jabalí, jabarín*, y en uno de los de Huelva (517), como segunda respuesta tras *jabalí*.

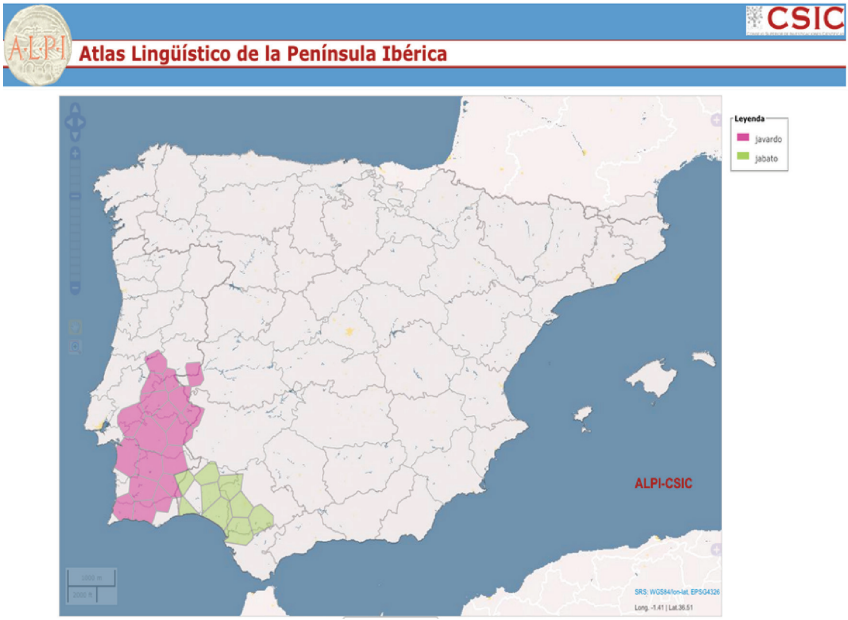


Imagen 6: Mapa de la extensión de *jabato* (verde) y *javardo* (rosa)

Referencias bibliográficas

ALPI = [Navarro Tomás, Tomás (dir.)/Aurelio M. Espinosa hijo/Luís F. Lindley Cintra/Francesc de Borja Moll/Armando Nobre de Gusmão/Aníbal Otero/Lorenzo Rodríguez Castellano/Manuel Sanchis Guarner] (1962): *Atlas Lingüístico de la Península Ibérica*, I, *Fonética*. Madrid: CSIC.

web ALPI = García Mouton, Pilar (coord.)/Inés Fernández-Ordóñez/David Heap/Maria Pilar Perea/João Saramago/Xulio Sousa (2016–): *ALPI-CSIC* <alpi.csic.es>, edición digital de Navarro Tomás, Tomás (dir.): *Atlas Lingüístico de la Península Ibérica*. Madrid: CSIC [último acceso: 15/09/2017].

ALeCMan = García Mouton, Pilar/Francisco Moreno Fernández (dirs.) (2003–): *Atlas Lingüístico (y etnográfico) de Castilla-La Mancha.* <www.linguas.net/alecman> [último acceso: 15/09/2017].

Alvar, Manuel (1982): «Atlas lingüísticos y diccionarios», *Lingüística Española Actual* IV, 253–323. Recogido en Manuel Alvar (1991): *Estudios de geografía lingüística*. Madrid: Paraninfo, 49–115.

Corrales Zumbado, Cristóbal/Dolores Corbella Díaz (2002–2004): «El *ALEICan* en los diccionarios», en Rosa M.ª Castañer y José M.ª Enguita (eds.), *Archivo de Filología Aragonesa* LIX-LX, *In memoriam Manuel Alvar*, II, 1203–1222.

Corriente, Federico (2003[2]): *Diccionario de arabismos y voces afines en iberorromance*. Madrid: Gredos.

DECH = Corominas, Joan/José Antonio Pascual (1980–1991): *Diccionario Crítico Etimológico Castellano e Hispánico*, 6 vols. Madrid: Gredos.

DLE = Real Academia Española/Asociación de Academias de la Lengua Española (2014[23]): *Diccionario de la lengua española*. Barcelona: Espasa Libros.

Fernández-Ordóñez, Inés (2011): *La lengua de Castilla y la formación del español*. Madrid: Real Academia Española.

Frenk, Margit (1987): *Corpus de la antigua lírica popular hispánica (siglos XV a XVIII)*. Madrid: Castalia.

García Mouton, Pilar (1990): «El estudio del léxico en los atlas lingüísticos», en Francisco Moreno Fernández (recop.), *Estudios sobre variación lingüística*. Alcalá de Henares: Universidad de Alcalá, 27–75.

García Mouton, Pilar (1997): «Zoónimos no latinos en español», en Sylvie Mellet (ed.), *Les zoonymes. Actes du colloque international tenu à Nice les 23, 24 et 25 janvier 1997*. Niza: Université de Nice – Sophia Antipolis, 233–248.

García Mouton, Pilar (2009): «La investigación geolingüística española en la actualidad», en Dolores Corbella y Josefa Dorta (eds.), *La investigación dialectológica en la actualidad*. Tenerife: Agencia Canaria de Investigación, Innovación y Sociedad de la Información, 333–346.

García Mouton, Pilar (2010): «El procesamiento informático de los materiales del *Atlas Lingüístico de la Península Ibérica* de Tomás Navarro Tomás», en Gotzon Aurrekoetxea y José Luis Ormaetxea (eds.), *Tools for Linguistic Variation*. Bilbao: Universidad del País Vasco, 167–174.

García Mouton, Pilar (2016): «Corominas tenía razón: *jamila* no *jámila*», en Mariano Quirós García, José Ramón Carriazo Ruiz, Emma Falque Rey y Marta Sánchez Orense (eds.), *Etimología e historia en el léxico español*, Estudios ofrecidos a José Antonio Pascual. Madrid/Fráncfort: Iberoamericana/Vervuert, 293–302.

García Mouton, Pilar (2017): «El *Atlas Lingüístico de la Península Ibérica (ALPI)* en línea. Geolingüística a la carta», *Estudis Romànics* 39, 335–343.

Heap, David (2002): «Segunda noticia del *ALPI* (a los cuarenta años de la publicación de su primer tomo)», *Revista de Filología Española* LXXXII, 5–19.

Lapesa, Rafael (1981[9]): *Historia de la lengua española*. Madrid: Gredos.

Navarro Tomás, Tomás (1975): *Capítulos de geografía lingüística de la Península Ibérica*. Bogotá: Instituto Caro y Cuervo.

Pato, Enrique (2011): «Sobre geografía léxica española: Distribución y áreas léxicas de la *mustela*», *Dialectologia* 6, 45–53.

Pérez Pascual, José Ignacio (2017): *Los primeros pasos de un largo caminar. Los comienzos del Atlas Lingüístico de la Península Ibérica*. San Millán de la Cogolla: Cilengua.

Pharies, Ralph (2002): *Diccionario etimológico de los sufijos españoles y de otros elementos finales*. Madrid: Gredos.

Rolf Eberenz

Hacia un diccionario de la alimentación y la culinaria medievales y renacentistas

Resumen: Se presentan aquí las estructuras generales de nuestro proyecto de un diccionario castellano de la alimentación y la cocina antiguas, actualmente en curso de elaboración, así como los tipos de información que se pretende proporcionar. Se hace especial hincapié en las clases de lexemas estudiados, más numerosas y amplias que las que suelen recogerse en los repertorios lexicográficos sobre este tema, y en los problemas que plantea la selección de las voces, sobre todo en relación con los sufijados, las lexías complejas y los compuestos. Se describe igualmente la microestructura de las entradas, en las que ocupan un espacio importante los datos enciclopédicos relativos a los platos en cuestión, sus ingredientes, su elaboración y su estatus socioalimentario.

Palabras clave: Lexicografía diacrónica, Historia de la lengua, Historia social

Abstract: Here we present the general structures of our project for a Castilian dictionary of historical food and cooking, currently under development, as well as the types of information that are intended to provide. Special emphasis is placed on the classes of lexemes studied, which are more numerous and more broad than those usually found in the lexicographical repertoires concerning this subject. We will pay attention to the problems raised by the selection of words, especially in relation to suffixed ones, phrases and compounds. The micro-structure of the entries is also described, in which the encyclopedic data on the dishes in question, their ingredients, their elaboration and their social status will occupy an important space.

Keywords: Diachronic Lexicography, History of Language, Social History

1 Coordenadas generales

¿Por qué un diccionario de la alimentación antigua? Antes de responder a esta pregunta conviene recordar que los hábitos alimenticios de cada sociedad constituyen un sistema más o menos estructurado, un código cultural que interesa tanto a los historiadores como a los antropólogos y, por supuesto, a los lingüistas. Acercarse al léxico de este código es un reto apasionante que permite adentrarse en un mundo relacionado con nuestras vivencias de cada día, aunque en el caso de la alimentación antigua también bastante diferente de la realidad moderna. Ello tiene que ver con el hecho de que la variedad de productos disponibles en

la época era mucho menor que hoy en día. Sobre todo, faltaba todavía una serie de vegetales corrientes en la Europa actual, como la patata, el tomate, el maíz, el pimiento americano, el cacao y, sobre todo al principio, el azúcar en cuanto edulcorante barato y en cierto modo trivial. Si el azúcar colonial empezó a llegar a la Península durante el siglo XVI, los frutos de las demás plantas no se difundieron en la alimentación de las sociedades del Viejo Mundo hasta el XVII o, incluso, más tarde.

Por «alimentación antigua» entendemos los hábitos nutricionales documentados en lengua castellana del siglo XIII a principios del XVII. Hemos fijado el límite cronológico final del proyecto en 1611, año en que se publica el *Tesoro de la lengua castellana o española* de Sebastián de Covarrubias, autor muy atento al léxico de la comida (*cfr*. Eberenz 2014b), y se edita el *Arte de cocina, pastelería, vizcochería y conservería*, de Francisco Martínez Montiño, particularmente interesante para este tema.

Elaborar un diccionario castellano de la alimentación medieval y renacentista puede sorprender por otro motivo. En la Edad Media, los territorios castellanohablantes de la península ibérica no se caracterizaban precisamente por una importante tradición culinaria como la que existía, por ejemplo, en Cataluña, en Italia o en Francia. Las dos Castillas se presentan más bien como un espacio que recibe numerosas influencias externas: hispanoárabes, catalano-valencianas, francesas e italianas, impulsos que contribuyen, más tarde, a configurar el código alimentario del Siglo de Oro español. Y los hábitos culinarios de los siglos XV y XVI se difunden después también en los virreinatos americanos, donde se combinan con las tradiciones amerindias. Muchos de los conceptos peninsulares pasan, pues, al Nuevo Mundo, en el que algunos se transforman y cambian de contenido. Piénsese en nociones tan emblemáticas como *alfeñique, manjar blanco, nogada* o *tortilla*, que en América designan hoy en día realidades diferentes de las que se conocen en España.

No vamos a detenernos demasiado en las generalidades ni comentaremos las principales fuentes del proyecto puesto que ya lo hemos hecho en otras publicaciones (Eberenz, 2014a y 2015). Algunos de los textos más importantes que se utilizan para esta investigación figuran al final del presente trabajo. Cronológicamente, van del *Arte cisoria* de Enrique de Villena (1423) al ya mencionado *Arte de cocina* de Martínez Montiño (1611). Se han explorado igualmente los datos del *Corpus Diacrónico del Español* (*CORDE*) y del *Corpus del Nuevo Diccionario Histórico del Español* (*CNDHE*).

Es sabido que los libros de cocina en lengua castellana, fuentes principales para este tipo de estudio, son poco numerosos y tardíos. La primera obra que de alguna manera se refiere a la transformación de los alimentos para la mesa

es, precisamente, el *Arte cisoria* de Enrique de Villena; la segunda, el anónimo *Manual de mugeres*, algo anterior a 1525. Le sigue el *Libro de cozina* o *Libro de guisados* de Ruperto de Nola, de 1525, versión castellana del *Llibre del coc* catalán, cuya primera edición data de 1520. La obra recoge buena parte de la tradición gastronómica catalana y valenciana del siglo XV, enriquecida por un cierto número de platos hispanoárabes e italianos. El texto catalán suscita uno de los problemas más interesantes para el estudio del léxico alimentario, el del significado de los nombres de platos, muchos de los cuales son opacos, foráneos, alterados y de difícil interpretación. Esta dificultad se planteó varios al traductor castellano y a otros autores que más tarde recogieron y reformularon recetas de Nola, especialmente a Juan Vallés, compilador de una obra titulada *Regalo de la vida humana*, de hacia 1560, y a Diego Granado, autor del *Libro del arte de cozina*, de 1599. Las últimas obras del corpus son el *Libro del arte de cozina* de Domingo Hernández de Maceras (1607) y el ya citado tratado de Martínez Montiño (1611). Como se ha dicho, estos textos contienen varios platos de otras tradiciones culinarias, algunos de los cuales probablemente no solían prepararse en España. Por otro lado, tales tratados reflejan una cocina muy particular, la de la nobleza y, hasta cierto punto, de la burguesía acomodada de las ciudades, pero se refieren solo muy parcialmente a los hábitos alimentarios de las clases populares.

Sin embargo, hay otras fuentes que pueden aprovecharse para obtener una visión más completa de lo que era la alimentación de la sociedad castellanohablante. La clase de obras más interesante es la de los tratados médicos, que contienen abundantes datos no solo sobre platos dietéticos sino también sobre la alimentación corriente. Incluyen igualmente información valiosa algunas obras relativas a la medicina veterinaria, a la agricultura y a la botánica. Por fin, se han tenido en cuenta bastantes obras literarias que abundan en referencias a la comida, como el *Libro de buen amor,* el *Corbacho*, *La Lozana Andaluza,* el *Viaje de Turquía* y el *Quijote*.

Varios, si no la mayoría, de estos textos ya se habían utilizado para las concordancias del *CORDE* y del *CNDHE*. Hemos vaciado de nuevo muchos de ellos para obtener datos más completos para el conocimiento de los distintos alimentos y platos[1].

1 Sobre este y otros puntos del proyecto se puede encontrar más información en nuestra página electrónica (<https://people.unil.ch/rolfeberenz/> [último acceso: 10/10/2017]), en la que se actualizan periódicamente p. ej. la nomenclatura, el corpus, la bibliografía y unas muestras de entradas.

Los diccionarios de alimentación y gastronomía suelen centrarse en las denominaciones de platos, y algunos se limitan incluso a esta parcela del léxico. En la época estudiada existían muchas preparaciones culinarias poco conocidas hoy en día, como el *almodrote*, la *capirotada*, el *gigote* o la *olla podrida*. Por otra parte, algunos términos actualmente corrientes muestran un notable diferencial semántico. Así, el *pisto* no era un plato de verduras sino un caldo concentrado de carne triturada y prensada que se usaba para el régimen de los enfermos. Una *salsa* podía ser lo mismo que hoy en día, pero la palabra también se refería a un conjunto de especias y hierbas aromáticas, lo que explica, por ejemplo, que en portugués *salsa* designe el perejil. Una *sopa* no era una preparación líquida o cremosa sino una rebanada de pan tostada o frita que se embebía en caldo o en otro líquido y servía a menudo de soporte a otro alimento, p. ej. a cierta cantidad de carne o verdura. Y *tisana* no representaba una infusión de hierbas sino una bebida medicinal de granos de cebada triturados y cocidos.

2 Clases de palabras

Se tienen en cuenta todas las voces que de alguna manera están relacionadas con los alimentos y con su transformación, no solo los nombres de platos. Así pues, se prevé incluir en el diccionario los siguientes conjuntos:

- *Sustantivos* referentes a:
  - Preparaciones culinarias o platos
  - Productos semielaborados como el pan, el aceite, el vino, el vinagre, etc.
  - Las diferentes clases de vegetales que se aprovechan para la confección de los platos: hortalizas (especialmente las legumbres), frutas, hierbas aromáticas, etc.
  - Alimentos de origen animal, como carnes y pescados, pero también huevos, leche, miel, etc.
  - Sustancias minerales como el agua y la sal
  - Las diferentes comidas de un día corriente
- *Adjetivos* con los que se denominan las propiedades de las materias comestibles, especialmente sus sabores, colores, consistencia, estado de conservación, etc.
- *Verbos*, sobre todo los relativos a las diferentes manipulaciones de los alimentos.

El vocabulario de la alimentación antigua en las diferentes lenguas europeas se ha recogido, a veces más bien de pasada, en numerosas monografías de orientación sociohistórica, y en varios trabajos sobre un número limitado de términos,

así como en los glosarios que complementan las ediciones críticas de los libros de cocina. En cambio, hay pocos repertorios lexicográficos completos. Una notable excepción es el excelente diccionario de Enrico Carnevale Schianca, *La cucina medievale. Lessico, storia, preparazioni* (2011), que se basa en los múltiples recetarios de la península itálica, redactados tanto en latín como en las distintas variedades romances. Este diccionario nos está sirviendo, en cierto modo, de modelo, aunque nuestro repertorio cubrirá no solo la Edad Media sino también parte del Siglo de Oro. Por otro lado, debido a la importancia que posee la literatura médica para la alimentación, se utiliza también el *Diccionario español de textos médicos antiguos* (*DETEMA*) dirigido por Mª Teresa Herrera (1996).

3 Macroestructura

Veamos, en primer lugar, la nomenclatura de términos que se tratan en el diccionario. Se incluyen, en principio, todos los lexemas simples y las lexías complejas suficientemente atestiguados en la documentación. En cambio, se descartan conceptos utilizados únicamente en la literatura médica o farmacéutica, pues sobre ellos ya existen muchos estudios y repertorios. He aquí, como muestra de la macroestructura, la relación de lemas iniciados por la letra M (la lista es provisional):

> Macarrón, Machacar, Machar, Machucar, Macis, Madrecilla, Madrezuela, Magro, Majadero, Majar, Malcocinado, Malva, Malvasía, Maná, Manecilla, Manga, Manguito, Manir, Manjar, Manjar blanco, Manjar de ángeles, Manjar imperial, Manjar principal, Manjar real, Manteca, Mantecada, Mantecar, Mantecoso, Mantenimiento, Mantequilla, Manzana, Marisco, Masa, Masa blanca, Masa encerada, Masa fina, Masa negra, Mastuerzo, Matalahúva, Materia, Material, Mazacote, Mazamorra, Mazapán, Mecedor, Mecer, Mecha, Mechadera, Mechar, Mejorana, Melado, Melcocha, Melindre, Melliza, Melocotón, Melón, Membrillate, Membrillo, Memoria, Mendrugo, Meneador, Menear, Menestra, Menudillo, Menudo, Meollo, Merendar, Merienda, Merluza, Mermelada, Mero, Mezclar, Miel, Miga, Migajón, Mijo, Milgrana, Ministrar, Mirrauste, Mojama, Mojí, Mojo, Molde, Moler, Molino, Molleja, Mollete, Mondar, Mondo, Mondongo, Montés, Montesino, Mora, Morcilla, Morcón, Morena, Moreta, Mortero, Morteruelo, Moscatel, Mostachón, Mostaza, Mosto, Moyuelo, Mulso, Murta, Musco.

Al seleccionar las voces que constituyen la nomenclatura surgen dos problemas específicos, el de los derivados y el de las lexías complejas. Entre los derivados hay, ante todo, bastantes diminutivos y algunos aumentativos, así como una serie de sufijados en *–ada* (*almendrada, limonada, nogada,* etc.). En cuanto a los diminutivos, se comentan solo los que están suficientemente lexicalizados, esto es, los que muestran un significado claramente diferenciado del que posee la palabra base. Veamos algunos ejemplos. A partir del siglo XVI, *empanadilla*

ya no denomina simplemente una *empanada* pequeña sino una pasta de pequeñas dimensiones, dulce y frita, de modo que coincide hasta cierto punto con lo que en la época se llamaba *fruta de sartén.* Lo mismo ocurre con *pastelillo*, que desde el siglo XVI se refiere a un pastel pequeño, relleno de sustancias dulces, como mazapán o diacitrón, y de especias. Asimismo, se observa un proceso de diferenciación semántica en la palabra *tortilla,* entre su actual significado peninsular y el usual en las normas americanas del español. Para el valor con que hoy se usa en España, la expresión más antigua es «huevos en *tortilla*», atestiguada en Villena, mientras que más tarde los autores hablan de «*tortilla* de huevos», y esta expresión se mantiene durante varios siglos en contraste con «*tortilla* de pan» o «*tortilla* de masa».

Entre las lexías complejas hay, por ejemplo, un gran número de nombres de *pasteles*, para los cuales conviene tener en cuenta que en la época *pastel* denominaba siempre una preparación salada, rellena de carne o verdura. Así, se encuentran numerosas expresiones del tipo «*pastel* de codornices», «*pastel* de hígado de carnero», «*pastel* de lomo de ternera», etc. Puesto que estos giros son transparentes y se pueden interpretar sobre la base del significado de cada uno de sus elementos léxicos, y dado que la variedad de pasteles era muy importante, preferimos no tratar estos conceptos en entradas independientes sino caracterizarlos brevemente en los artículos dedicados a sus principales ingredientes, es decir, en los ejemplos mencionados, a *codorniz, carnero* y *ternera*. En principio, nos limitamos a describir en artículos separados las lexías complejas opacas, cuyo significado no se puede desentrañar a partir de la semántica de los elementos que las componen. Por ejemplo, se dedica una entrada a la lexía *fruta de sartén* porque no resulta transparente, aunque es interesante. La expresión tiene que ver con el hecho de que los postres consistían generalmente en fruta fresca. Sin embargo, en el siglo XV se pusieron de moda ciertas pastas fritas, con lo cual nació la distinción entre *fruta verde* (o fresca) y *fruta de sartén.* Otro caso bien conocido de la gastronomía áurea es el de la *olla podrida*, término documentado desde la primera mitad del siglo XVI y relativo a un plato de varias clases de carne y hortalizas. La explicación de por qué se llama *podrida* ha hecho correr mucha tinta, y algunas hipótesis son bastante fantasiosas. Como en su origen se trata de un plato rústico, lo más plausible es que estemos ante una creación léxica de carácter lúdico, expresivo, nacida en el medio popular y referente a la gran variedad de ingredientes. También se podría mencionar la *cazuela mojí* (a veces se habla de *berenjenas mojíes*), nombre de un plato de origen hispanoárabe bien documentado en el siglo XVI y cuyo segundo elemento refleja el árabe andalusí *mušḥí*, que significa 'relleno'. Evidentemente, todas estas expresiones y otras más o menos emblemáticas de la cocina áurea se van a incluir en el diccionario.

Además, se crean entradas independientes para lexías complejas como *salsa negra* y *salsa verde*, que denominan dos célebres preparaciones de la gastronomía clásica. Su adjetivación apunta a la importancia de los colores en la culinaria tardomedieval y renacentista. La *salsa negra* se llama así porque se hacía con harina tostada y vino tinto —que le daban su color oscuro—, además de con miel o azúcar y con especias. Tenía un sabor pronunciado y acompañaba especialmente a la carne de caza. En cuanto a la *salsa verde*, llamada también *salsa de perejil,* figura ya en ciertos tratados de medicina medievales, incluso en latín (hacia 1300, el médico catalán Arnau de Vilanova habla de *salsamentum viride*). Contiene sobre todo perejil y otras hierbas aromáticas, pan tostado, triturado y desleído con vino blanco, a veces pimienta y un poco de miel. En este caso, se construirá una sola entrada «*salsa verde, salsa de perejil*», con una remisión a ella en el lugar que por el orden alfabético corresponde a *salsa de perejil.* Una situación similar se da en las diferentes denominaciones del clavo de especia: se reúnen en una entrada única «*clavo, girofre, clavo de girofre, clavo de especia*», con una remisión bajo *girofre*.

Hay igualmente unos llamativos términos compuestos, como *aguamiel, ajopollo, hierbabuena, malcocinado, salpimienta, salpreso*, etc. Es interesante, por ejemplo, el término *malcocinado,* nombre de un guisado de despojos, especialmente de tripas. Se trata de un plato de ínfima calidad que se vendía en puestos callejeros y se documenta desde el siglo XIV. Con el tiempo, la palabra pasó a denominar también los puestos en que se vendía. ¿Por qué el manjar se llamaba *malcocinado*? Sin duda porque las carnes en cuestión se cocinaban de forma improvisada en la calle, y porque se trataba de un plato preparado con ingredientes malos o poco apreciados.

4 Microestructura: organización de las entradas

Cada artículo se inicia con el lema, normalmente en la forma que tiene en español moderno y que registra el diccionario de la Real Academia. Cuando en la lengua antigua se grafía sistemáticamente de otro modo, lo señalamos entre paréntesis, p. ej. *hierba* (ant. *yerva*), *garbanzo* (ant. *garvanço*). Siguen las variantes formales, incluidas las dialectales, y la información gramatical relativa al término. En los casos en que la palabra tiene diversas acepciones, estas se presentan en diferentes apartados numerados y ordenados según su antigüedad. Cada acepción se caracteriza con una breve definición del valor semántico que la voz posee en los contextos alimenticios de la época. La definición resume en pocas palabras los datos enciclopédicos que se exponen después. Se evita recurrir a las definiciones del *Diccionario de autoridades* de la Real Academia, como es costumbre en muchos

estudios sobre el léxico medieval, puesto que a menudo las palabras no significan lo mismo en la Edad Media que en el Siglo de Oro, al que se refiere dicho diccionario. Los problemas de definición se plantean sobre todo a propósito de los nombres de vegetales y de sus partes. Cuando hablamos de plantas conocidas, optamos por definiciones muy someras, sin mencionar todos los rasgos botánicos como se hace en el diccionario académico. En cambio, mencionamos sistemáticamente el nombre latino de la planta según la clasificación de Linné, lo que permite aclarar de qué especie se trata.

Sigue el primer testimonio conocido de la voz, o varios de los primeros testimonios cuando poseen un interés especial. Estas primeras documentaciones se extraen a menudo del *CNDHE* o de repertorios lexicográficos existentes. Tienen, a nuestro modo de ver, un valor un tanto relativo ya que con frecuencia los años en que se copiaron o publicaron los textos en cuestión difieren de su época de redacción. Sin embargo, permiten situar la palabra estudiada, sea en el acervo patrimonial del castellano medieval, sea en un contingente léxico más reciente, cuyas innovaciones pueden responder a necesidades de denominación.

A continuación, se presenta la parte central del artículo. Partimos de la idea de que quienes consultan un diccionario de alimentación y gastronomía se interesan en primer lugar por los nombres de ciertos platos y por lo que estos representan en la realidad histórica, a saber, qué ingredientes contienen, cómo se preparan, en qué momentos y contextos sociales se comen, qué propiedades dietéticas poseen, etc. La información enciclopédica es, pues, primordial en un repertorio de este tipo. Por otro lado, importa que esta información se pueda leer con cierta facilidad. Buena parte de la documentación sobre el tema se puede encontrar en las concordancias consultables en línea, y muchos recetarios históricos pueden leerse o, incluso, descargarse en Internet. Por lo tanto, no merece la pena proporcionar un gran número de citas. Nos ha parecido más útil sintetizar e interpretar los datos que se pueden extraer de la documentación. Para estas explicaciones se ha elegido una forma de discurso entre expositiva y narrativa. Dicho de otro modo: se procura exponer en qué consiste el manjar o el ingrediente o la manipulación culinaria en cuestión, cómo se elabora el plato, etc. Ello implica a menudo la necesidad de parafrasear los testimonios más importantes, especialmente las recetas, pero insertando frecuentes menciones de las fuentes en que se atestiguan los elementos referidos. Dentro de esta parte de los artículos reservamos, pues, las citas para los casos donde son realmente imprescindibles y poseen un gran valor ilustrativo. También se realizan a menudo remisiones a otros términos tratados en el diccionario mediante el símbolo (⇨), por un lado para no repetir la información ya expuesta en esas entradas, por otro para poner en evidencia las estructuras del léxico alimenticio.

Asimismo, se intenta determinar las coordenadas antropológicas, étnico-religiosas y sociales en que se sitúan los conceptos, especialmente cuando se trata de alimentos o platos. Respecto de la antropología, se podría citar el conocido conjunto de operaciones culinarias designadas con los verbos *asar*, *cocer* y *freír*, estudiados por Claude Lévi-Strauss (2004 [1965]). Otro ejemplo es el de las nociones *puches/gachas, torta* y *pan.* Las *gachas* o *puches* constituyen la forma más primitiva y rústica de consumir la harina de cereales, cocida con agua y sal. Las *tortas* representan un grado de elaboración más alto, mientras que el *pan*, de miga esponjosa gracias al uso de levadura, simboliza en cierto modo la alimentación «civilizada». Este sistema de gradación —digamos de rusticidad a civilización— se manifiesta, por ejemplo, en el refrán «A falta de *pan,* buenas son *tortas*».

Para las coordenadas étnico-religiosas se podría mencionar el caso de la *berenjena*, hortaliza característica de los platos hispanomusulmanes (o judíos), ámbito cultural al que pertenecen al principio también las preparaciones llamadas *cazuelas*. En cuanto a las implicaciones sociales, vamos a referirnos brevemente a la historia del concepto *ensalada.* Las ensaladas se pusieron de moda entre los nobles castellanos durante el siglo XV, y la nueva costumbre se debió, con toda probabilidad, a influencia italiana. De hecho, la palabra no puede ser otra cosa que un préstamo del italiano *insalata*, ya que en español falta un verbo **ensalar*. Partiendo de la conocida distinción de Lévi-Strauss (1978 [1964]: 340–347; 2004 [1965]: 9–19) entre lo *crudo* y lo *cocido*, se puede recordar que en las sociedades europeas de la época los miembros de las clases altas no consumían hortalizas sin cocer, contrariamente a los campesinos que, en ocasiones, se contentaban con cebollas o ajos crudos, complementados por unos mendrugos de pan y algún pedazo de queso o tasajo. La introducción de las ensaladas constituyó, pues, en cierto modo una ruptura del código alimenticio tradicional, una innovación significativa en los hábitos nutricionales de las clases altas de la península ibérica.

Llegamos a la última parte de la microestructura de los artículos, dedicada a la etimología. En conformidad con la concepción general de nuestro diccionario, consideramos esta sección un tanto secundaria, sobre todo cuando se trata de palabras de origen conocido, sea románico, sea germánico o árabe. Todos los artículos sobre lexemas simples se terminan con un apartado titulado *Antecedentes*, con una breve noticia etimológica. En cambio, parece útil ampliar esta información en los casos donde el término en cuestión tiene una trayectoria internacional o interlingüística compleja, como ocurre, por ejemplo, con *mostaza.* El nombre castellano tradicional de la planta *Brassica nigra* era *xenabe* (< lat. *sĭnāpi*). Con los granos de este vegetal y, a menudo, con *mosto* de uva se

elaboraba la *mostaza*, de modo que el segundo ingrediente explica la denominación de esta salsa. Con el tiempo, *mostaza* empezó a aplicarse también a la planta y a sus granos, suplantando a *xenabe*. En otras lenguas se usan términos análogos, p. ej. en catalán *mostalla/mostassa*, en francés *moutarde* y en las variedades italianas *mostarda/mustarda*, lo que suscita la cuestión de la zona de origen del producto —probablemente Francia— y de sus vías de difusión.

5 Ejemplo de entrada

potaje m. **1.** 'Preparación culinaria compleja'. 1as docs.:
◊ […] pero sabrán / qué quiere dezir *potage*; / regulage con formage / ayan si comieren pan, / quẽl passage nin ostage / nunca gelo soltarán. (1379–a1425 Alfonso de Villasandino: *Poesías* [*Cancionero de Baena*], ed. B. Dutton y J. González Cuenca; *CNDHE*)
◊ Et no han en muchos logares nj arbeillas nj fauas nj otro *potage* si no del broet dela carne (c1400 *Viaje de Juan de Mandevilla*. Escorial M.III.7, ed. J. L. Rodríguez Bravo y Mª del M. Martínez Rodríguez; *CNDHE*)

El término aparece, pues, por primera vez en una serie de galicismos sufijados con *-aje*[2], todos ellos de connotación afectada, en un poema de la lírica cancioneril (1ª cita); y, por los mismos años, se encuentra en una traducción con impronta aragonesa del relato de viaje de Mandeville (2ª cita). Un poco más tarde, *potaje* es empleado varias veces y con toda normalidad por Ruy González de Clavijo cuando describe las comidas de los tártaros de Tamorlán, en secuencias como «truxieron mucha carne e mucho arroz e *potajes* de muchas maneras» (a1412 *Embajada*, 178). Y unos años después, Alfonso de La Torre (1430–1440: 306) reprueba la variedad de nombres de *potajes* existentes en su época y la afición de sus contemporáneos a los «*potajes* en que aya colores para agradar la vista e olores de mulçimiento a los otros sentydos». La crítica de la excesiva diversificación de los platos en la mesa de los pudientes es frecuente en la literatura de los ss. XV y XVI, y en las enumeraciones de manjares considerados superfluos casi nunca faltan los *potajes*[3]. En esa época, la palabra se refiere, pues, a una clase de

2 El sentido de este ejemplo de *potaje* no resulta claro; pero teniendo en cuenta la historia de la palabra, la interpretación como 'bazofia' referida por López Quero (2011: 493) parece poco plausible.

3 Las mismas críticas a los *potajes* las formulan Hernando de Talavera (1477–1496 *Tratado*, 23, 45), Juan de Flores (1470–1492 *Triunfo de amor,* ed. J. Fernández Jiménez; *CNDHE*) y, más tarde, Antonio de Torquemada (1553: 331). Y Juan B. de la Concepción, en sus austeras instrucciones para los cocineros de la Orden Trinitaria (1607 *Oficios*, 477), advierte que no va a referirse a las «diferencias de *potajes* ni guisados».

guisados de gran complejidad y refinamiento. Tiene probablemente un matiz de extranjerismo de moda, aunque su significado resulta siempre algo borroso. Tal valor parece haberse consolidado con la difusión del tratado de Nola, en cuya versión original catalana de 1520 *potatge* aparece con llamativa frecuencia como denominación general de los guisados, a menudo seguido de un complemento que precisa la naturaleza de la preparación (Valles Rojo 2007: 37). La traducción castellana de 1525 ofrece en su lugar invariablemente *potaje*, en expresiones como «*potage* de manos de carnero», «*potage* de carnero adobado», etc. Este uso perdura hasta cierto punto en el tratado de Vallés (a1563), quien habla p. ej. de «carnero adovado que se dize *potage* pardo» (656). Por otro lado, en ambas obras se encuentran también salsas y preparaciones de verduras llamadas *potajes*. El autor del *Viaje de Turquía* (1550–1560: 797) equipara *potaje* al it. *minestra* y al turco *sorbas*, manjares que se pueden comer con cuchara. Este significado del término prevalece todavía en el libro de cocina de Granado (1599), aunque el autor aplica la voz a numerosos guisados de pescado y marisco, casi todos tomados del italiano Scappi (1570) y por tanto influenciados por el it. *potaggio*.

2. 'Bebida, brebaje'. 1as docs.:

◊ e porque Sócrates burlava de los athenienses e de sus dioses, condepnáronle a muerte que beviesse un *potaje* de venino. (1446 Pero Díaz de Toledo: *Traducción del Libro llamado Fedrón, de Platón*, ed. N. G. Round; *CNDHE*)

◊ [La Razón contra la Luxuria:] […] / Das a las gentes vltrajes, / de muerte non las reseruas; / tú fallas las tristes yeruas, / tú los crueles *potajes*. (c1453–a1456 Juan de Mena: «Debate razón» *[Cancionero de Gómez Manrique]*, ed. F. Vidal González; *CNDHE*)

Esta segunda acepción, bien documentada en la poesía de cancionero y en algunas traducciones de textos médicos del s. XV (*CNDHE*; *DETEMA*) así como en el *Universal vocabulario* de Palencia (1490: s.v. *jus*), se debe sin duda a un reanálisis, por autores familiarizados con el latín, de *potaje* como miembro de la familia léxica del lat. *pōtŭs* 'bebida' y *potare* 'beber', según afirma explícitamente Covarrubias (1611: s.v. *potage*).

3. 'Preparación líquida o semilíquida de agua, caldo, leche, etc., así como, preferentemente, de cereales y verduras'. 1a doc.:

◊ E si [sc. el herido] estouiere tan afrontado que no pudiesse comer, se le deue dar *potaje* claro, fecho assi como amidon, o semola, con caldo de pollo o carnero, o semejantes cosas de substancia liuiana e de buena digestion. (1494 Ketham, 166–167)

Es posible que el tercer valor de *potaje* represente en cierta manera una síntesis de las acepciones 1. y 2. Aunque durante el s. XVI e incluso a comienzos del XVII se conserve el sentido primero, p. ej. en las recetas de Maceras (1607) y de Montiño

(1611), el término se aplica con creciente frecuencia a manjares más o menos líquidos. Así, Bartolomé de Torres Naharro hace decir a uno de sus personajes que los señores «sorben *potajes*» (*CNDHE*), y en los tratados de culinaria tal tendencia resulta también acusada. En este contexto se puede recordar la pregunta sobre la diferencia entre *potage, caldo* y *cozina* que, en el *Diálogo de la lengua* de Juan de Valdés (1535–1536), Marcio dirige al personaje del mismo Valdés; pregunta a la que este contesta: «*potage* llaman a lo que acá [sc. en Italia] llamáis *minestra*» (206; *cfr.* Gemmingen 1987: 492–493). La respuesta confirma la forma líquida de los potajes, pero no el tipo de ingredientes, ya que la *minestra* italiana podía incluir carne u hortalizas (Carnevale 2011: 405–406). Sin embargo, poco a poco empiezan a predominar los potajes de verduras y cereales. Ya Alonso de Herrera (1513: 19v) menciona un plato de este tipo hecho con panizo y leche de almendras o de cabra. Nola (1525: 32v–33r) explica la preparación de dos *potajes modernos*, que contienen solo verduras y cuya caracterización como *modernos* podría referirse, precisamente, al cambio semántico que estamos comentando. Más tarde abundan los potajes de legumbres y verduras previstos para la Cuaresma, pues los menciona Maceras (72–73) y los describe Montiño (60r-v, 147r-v, 154r-v, 161r-v, 308v-309r, 317v), en recetas que incluyen zanahorias, habas, calabaza, arvejas, trigo, castañas, etc. La forma semilíquida y las hortalizas que constituyen sus ingredientes fundamentales se convierten entonces en las características esenciales de estas preparaciones. Al mismo tiempo, surge el derivado *potajería* 'conjunto de verduras y legumbres empleadas en los potajes', atestiguado por primera vez en un documento sobre la administración de la casa imperial de Carlos I, de la segunda mitad del s. XVI (Varela 2009: 2.1810), y utilizado de forma corriente por Montiño. Tanto este último valor de *potaje* como la introducción de *potajería* se deben sin duda a influencia francesa, lengua en la que se observa la misma evolución semántica.

ANTECEDENTES: El esp. *potaje* tiene una historia algo enigmática, como ocurre también con sus equivalentes léxicos en las variedades italianas de la Edad Media (Carnevale 2011: 534), puesto que este tipo léxico muestra durante mucho tiempo una sorprendente ambigüedad semántica, igualmente reflejada en la tradición lexicográfica. Por lo visto, los castellanohablantes intentaron vincular *potaje* a algún étimo latino conocido, de donde la reinterpretación de la palabra como derivado de POTŬS 'bebida'. En principio, todas estas voces se remontan al fr. *potage*, derivado de *pot* 'olla' (< lat. POTTUS). También la palabra francesa poseía originariamente un significado general, de 'guisado que incluía carne y verduras' (*FEW*, 9.268). Por otra parte, es posible que el castellano no adoptara la voz directamente del francés sino del catalán, donde *potatge* era más corriente y antiguo. Pero probablemente la palabra catalana procede a su vez del francés

o del occitano (*DECLlC*, 6.752–6.753), dado que en catalán *pot* no designa una vasija tan fundamental para preparar guisados como el fr. *pot* o el cast. *olla*.

Otro ejemplo de acepciones que se suceden en el tiempo es la historia de la palabra *manteca*. En su origen, el término se refiere a la 'materia grasa de la leche' y a la 'sustancia grasa y pastosa que se obtiene batiendo la nata', es decir, a lo que actualmente en España se llama *mantequilla.* Los autores hablan a veces de *manteca de cabras, ovejas* o *vacas*; y también es frecuente la expresión *manteca de ganado.* Este producto no se solía usar para cocinar, como se hacía, por ejemplo, en Francia (la «cuisine au beurre») sino para confeccionar ciertos dulces como las *empanadillas* o el *manjar blanco*. En segundo lugar, *manteca* significa 'grasa de los animales, especialmente del cerdo', como lo hacen también las palabras *enjundia, gordura, grasa, grosura, unto*, etc. Este segundo valor de *manteca* está atestiguado desde fines del siglo XV y, al principio, convive con el primero. Hay, pues, cierta ambigüedad, que en muchos casos se resuelve mediante un complemento: el nombre del producto lácteo se precisa con los elementos léxicos ya mencionados, mientras que la grasa de cerdo se llama *manteca de puerco.* Pero poco a poco se va generalizando la segunda acepción. Según el *Tesoro* de Covarrubias (1611), *manteca*, sin complemento, es la «gordura de animal, especialmente del lechón», mientras que el producto lácteo se denomina *manteca de ganado*. Esta última expresión todavía existe en ciertas variedades americanas del español. ¿Qué significaba entonces *mantequilla*? La voz aparece normalmente en plural y denomina unas pastillas o pelotillas del mismo material pero mezclado con azúcar y que se consumía como dulce muy apreciado. Se encuentra sobre todo en entornos textuales de ambientación pastoril.

6 Punto final: las relaciones semánticas

Para terminar, volvamos a la idea de que los hábitos alimenticios constituyen un código cultural que tiene implicaciones lexicológicas. Sobre todo, se descubren en ellos redes de relaciones semánticas entre palabras de significado afín. Un caso particularmente llamativo es el de los múltiples nombres sinónimos de vegetales usados en la elaboración de platos, problema onomasiológico que posee con frecuencia una dimensión variacional. He aquí algunos ejemplos:

- Vegetales:
  – *albahaca*, *basilicón*
  – *alcachofa*, *alcana*, *(al)canaria*, *alcaucil*
  – *anís*, *matalahúva*
  – *cebada*, *hordio*

  - *cilantro*, *culantro*, *celiandro*, *coriandro*
  - *clavo*, *girofre*, *clavo de girofre*, *clavo de especia*
  - *granada*, *mi(e)lgrana*
  - *guisante*, *bisalto*, *arveja*
  - *hierbabuena*, *hierba de huerto*, *hierba santa*, *hortolano*, *menta*
  - *mejorana*, *mayorana*, *(al)moraduj*
  - *melocotón*, *durazno*, *prisco*, *pérsico*
  - *nuez moscada*, *nuez de especia*
  - *pistacho*, *alfóncigo*
  - *sésamo*, *ajonjolí*, *alegría*
  - etc.
- Animales:
  - *abadejo*, *truchuela*, *curadillo*
  - etc.
- Productos:
  - *aceite*, *olio*
  - *almizcle, musco*
  - *gachas*, *puches*, *poleadas*
  - *jamón*, *pernil*, *lunada*
  - *jugo*, *zumo*
  - *pasta*, *masa*
  - *tuétano*, *caña de vaca*
  - etc.
- Utensilios de cocina:
  - *alambique*, *alquitara*
  - *almirez*, *mortero*
  - *cobertera*, *tapadera*
  - *asador*, *espeto*
  - *mano (de mortero)*, *majadero*, *pisador*, *meneador*
  - etc.

Hay otros conceptos, más generales, que muestran una curiosa pluralidad de denominaciones, especialmente entre los verbos:

- *amasar*, *apastar*, *sobar*, *heñir*
- *cuajar*, *helar*
- *desatar*, *deshacer*, *desleír*, *destemplar*
- *enjugar*, *secar*
- *gastarse*, *consumirse* [un líquido en un proceso de transformación]
- *machar*, *machacar*, *machucar*, *majar*
- *mecer*, *menear*, *revolver*, *tornear*, *traer (a una mano)*

– *tajar*, *cortar*
– *tender*, *extender*, *adelgazar*, *aplanar* [una masa]
– etc.

Una situación algo particular se da en las denominaciones de vegetales, pues en varios de los casos antes enumerados estamos ante términos semánticamente equivalentes, que se refieren a una misma planta, como ocurre, por ejemplo, con *clavo, girofre, clavo de girofre, clavo de especia*. Ahora bien, en otras series se manifiesta una evidente variación diatópica o diamésica: *cebada* es, en principio, la palabra castellana, *hordio* la voz aragonesa, aunque no falta en algunos textos castellanos. El mismo contraste diatópico se observa en *granada* y *mi(e)lgrana*, pues *mi(e)lgrana* es, nuevamente, el nombre aragonés, pero también aparece en algunas obras castellanas. Por otro lado, por ejemplo *basilicón* y *coriandro* pertenecen solo al vocabulario científico, mientras que los nombres populares de las respectivas plantas son *albahaca* y *cilantro/culantro*. En este contexto resultan particularmente interesantes los pares sinonímicos en los que un miembro es románico y otro de origen árabe, como acontece con *anís* frente a *matalahúva*, *mejorana/mayorana* junto a *(al)moraduj* y *alegría* o *sésamo* en comparación con *ajonjolí.* Sin embargo, puede haber situaciones más complejas, como la de *melocotón, durazno, prisco* y *pérsico.* El nombre castellano más antiguo y tradicional de este árbol y de su fruto —del género *Prunus persica*— era *durazno*, voz usada en la literatura culinaria de la época y, hasta hoy, en buena parte de América. *Prisco* se documenta en textos antiguos con impronta aragonesa y se conoce hasta nuestros tiempos como regionalismo de Aragón, aunque varios autores de la época (p. ej. Laguna 1555: 104) consideran que *prisco* y *durazno* no designan exactamente la misma variedad de dicha fruta. *Prisco* procede del lat. *PERSICUS*, pero la palabra latina tiene en castellano igualmente un continuador culto, *pérsico,* que se documenta p. ej. en la traducción castellana del *Lilio de medicina* de Gordonio (1495: 1.806, 2.1004, etc.). En cuanto a *melocotón*, se refiere originariamente a la fruta nacida del injerto de durazno en membrillo, hecho mencionado por primera vez en la *Obra de agricultura* de Alonso de Herrera (1513: 74r). En los siglos XVI y XVII, *melocotón* pasa a ser también el nombre de la *Prunus persica*, por lo que los términos *durazno* y *melocotón* se emplean a menudo como sinónimos, aunque más tarde *melocotón* se generaliza en el español peninsular.

Los ejemplos expuestos plantean sobre todo cuestiones onomasiológicas relacionadas con ciertas plantas y con sus respectivos frutos, al tiempo que permiten adentrarse en el terreno de la terminología técnica utilizada en los tratados de distintas disciplinas científicas. Por otro lado, nos encontramos a veces con

problemas de semántica de la lengua corriente: ¿Cómo funcionan en la diacronía y en la sincronía de la época, por ejemplo, ciertos pares sinonímicos como *tajar* y *cortar*, *enjugar* y *secar* o *zumo* y *jugo*? Sirva de botón de muestra la relación entre *zumo* y *jugo*. En los numerosos contextos en que se documentan las dos palabras se pueden observar dos acepciones comunes relativas a la alimentación: (1) 'líquido contenido en un objeto sólido, especialmente en vegetales como frutas y hierbas, y que se saca exprimiéndolo' y (2) 'líquido que se desprende de la carne cuando esta se asa o cuece'. Tanto *zumo* como *jugo* cuentan con testimonios que arrancan del siglo XIII, pero *zumo* es claramente más frecuente que su competidor, y ello en todo el período estudiado. En los libros de cocina son recurrentes los *zumos* de fruta, sobre todo los de naranja, limón, uva verde (*agraz*) y granada, con los que se suele regar la carne asada, particularmente la de ave. Pero también en relación con el líquido que se desprende de la carne al ser asada, predomina claramente *zumo.* Solo el traductor castellano de Nola así como Vallés y Granado, los tres marcados por la influencia catalana, usan alguna vez *jugo.*

En muchas ocasiones nos encontramos con una noción general que se ramifica en varios conceptos más específicos. Así, lo que hoy en día llamamos *comida* o *alimento* recibía en la época denominaciones bastante diferentes. Las más usuales eran *cosas de comer, el comer, vianda* y *manjar,* mientras que *comida* con este valor no se empleó hasta el s. XVI. Los autores de obras científicas preferían *mantenimiento, gobierno, cibo, nutrimento,* etc., en tanto que en los contextos de expediciones militares y marítimas se hablaba de *bastimento.* Asimismo, lo que actualmente expresamos con giros como *hacer, guisar* o *preparar [la comida]*, se designaba con otros verbos. Los más corrientes en los primeros tiempos del castellano eran *adobar* y *guisar*, aunque *adobar* se especializó pronto con otros significados. También se empleaban, de forma más esporádica, *aparejar* y *cocinar.* En cambio, faltaban *preparar* y *confeccionar*, ya que su uso se limitaba a los productos farmacéuticos y dietéticos.

Además, habrá que tener en cuenta algunos campos temáticos relacionados de forma más indirecta con el universo de la alimentación y la culinaria. Pensamos, por ejemplo, en los utensilios de cocina, en las principales clases de preparaciones que se describen en los recetarios (*ollas, cazuelas, escudillas; sopas, tortas, empanadas, pasteles*, *salsas, conservas,* etc.), en las comidas de un día corriente (*desayuno, almuerzo, yantar/comida, merienda, cena*, sin olvidar los conceptos de *colación* y *refección*), los sabores y los colores, los procedimientos de transformación de los alimentos mediante calor (*asar, cocer, hervir, freír; avahar, escaldar, escalfar, estofar, perdigar, sancochar, tostar,* etc.; *cfr.* Eberenz 2016: 87–92), las técnicas de partir un alimento en trozos o partículas más o menos pequeñas (*cortar/tajar, quebrantar, machacar, majar, moler, picar*, etc.).

¿Cómo se presentarán estas relaciones semánticas? Cuando se trata de conjuntos de pocos términos, nos limitamos a insertar bajo la entrada de cada palabra una remisión a los artículos sobre palabras afines. Sin embargo, sobre algunas áreas temáticas pensamos redactar textos sintéticos que destaquen las características de las relaciones entre los distintos términos.

Por fin, no hay que olvidar las grandes series de conceptos que constituyen verdaderas taxonomías biológicas, como las hortalizas, las frutas, las hierbas aromáticas, las especias, las distintas carnes y pescados, etc. Cada una de estas clases se caracteriza por el hecho de ser muy extensa, aunque la literatura culinaria y los tratados de medicina, agricultura, etc. que aquí se toman en consideración seleccionan un número relativamente limitado de nociones. De estos conjuntos bien conocidos no se van a presentar síntesis ni habrá remisiones a ellos en los artículos sobre los respectivos conceptos. Sin embargo, no se descuidarán los hiperónimos de las diferentes categorías, a saber, *hortaliza, verdura, fruta, hierba,* etc., pues algunos ofrecen una configuración semántica interesante que no siempre es idéntica a su significado actual.

Referencias bibliográficas

Alonso de Herrera, Gabriel (1513): *Obra de agricultura acopilada de diversos autores*. Alcalá de Henares: Arnao Guillén de Brocar. <books.google.es> [último acceso: 10/10/2017].

Carnevale Schianca, Enrico (2011): *La cucina medievale. Lessico, storia, preparazioni*. Florencia: Leo S. Olschki Editore.

CNDHE = Instituto de Investigación Rafael Lapesa de la Real Academia Española: *Corpus del Nuevo Diccionario Histórico del Español*. <http://web.frl.es/CNDHE> [último acceso: 10/10/2017].

CORDE = Real Academia Española: Banco de datos (*CORDE*) [en línea]. *Corpus diacrónico del español*. <http://www.rae.es> [último acceso: 10/10/2017].

Covarrubias, Sebastián de (1611 [1987]): *Tesoro de la lengua castellana o española*. Ed. Martín de Riquer, Barcelona: Editorial Alta Fulla.

DECLlC = Coromines, Joan (1980–1991): *Diccionari etimològic i complementari de la llengua catalana*. Barcelona: Curial/«La Caixa».

DETEMA = Herrera, Mª Teresa (1996): *Diccionario español de textos médicos antiguos*. Madrid: Arco/Libros.

Eberenz, Rolf (2014a): «El léxico español de la alimentación y la culinaria en su historia: fuentes y líneas de investigación», en Vicente Álvarez Vives, Elena Díez del Corral Areta y Natacha Reynaud Oudot (coords.), *Dándole*

cuerda al reloj: ampliando perspectivas en lingüística histórica del español. Valencia: Tirant Humanidades, 23–46.

Eberenz, Rolf (2014b): «Alimentos, platos y bebidas en el *Tesoro de la lengua castellana o española* (1611) de Sebastián de Covarrubias», en Rolf Eberenz (ed.), *Discursos hispánicos sobre alimentación y culinaria. Aproximaciones literarias y lingüísticas*. Madrid: Visor Libros, 385–407.

Eberenz, Rolf (2015): «El vocabulario castellano de la culinaria medieval y renacentista – las fuentes tratadísticas y sus rasgos léxicos», *Revue de linguistique romane* 79, 495–523.

Eberenz, Rolf (2016): «De lo crudo a lo cocinado: sobre el léxico fundamental de la culinaria en la historia del español (siglos XIII a XVII)», *Revista de Filología Española* 96.1, 81–112.

Embajada = González de Clavijo, Ruy (a1412 [1999]): *Embajada a Tamorlán*. Ed. Francisco López Estrada, Madrid: Castalia.

FEW = Wartburg, Walther von (1922–2002): *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes*. Leipzig/Bonn/Basilea: Teubner/Klopp/Zbinden.

Gemmingen, Barbara von (1987): «Das verborgene Kochbuch des Herrn Oudin oder: Untersuchungen zum kulinarischen Wortschatz im zweisprachigen Wörterbuch», *Zeitschrift für Romanische Philologie* 103, 491–499.

Gordonio, Bernardo de (1495 [1993]): *Lilio de medicina*. Ed. Brian Dutton y Mª Nieves Sánchez, 2 vols., Madrid: Arco/Libros.

Granado, Diego (1599): *Libro del arte de cozina, en el qual se contiene el modo de guisar de comer en qualquier tiempo, assi de carne, como de pescado, para sanos y enfermos, y conualecientes, assi de pasteles, tortas y salsas, como de conseruas a la vsança Española, Italiana, y Tudesca, de nuestros tiempos*. Madrid: Luis Sánchez.

Ketham, Johannes de (1494 [1990]): *Compendio de la humana salud*. Ed. Mª Teresa Herrera, Madrid: Arco/Libros.

Laguna, Andrés de (1555): *Pedacio Dioscorides Anazarbeo, Acerca de la materia medicinal y de los venenos mortíferos, traduzido de lengua griega en la vulgar castellana*. Amberes: Juan Latio. <books.google.es> [último acceso: 10/10/2017].

Lévi-Strauss, Claude (1964 [1978]): *Le cru et le cuit*. París: Plon.

Lévi-Strauss, Claude (1965 [2004]): «Le triangle culinaire», *Food & History* 2, 9–19.

López Quero, Salvador (2011): «El léxico gastronómico medieval del *Cancionero de Baena*», *Zeitschrift für Romanische Philologie* 127, 476–502.

Maceras = Hernández de Maceras, Domingo (1607 [1999]): *Libro del arte de cozina*. Ed. Santiago Gómez Laguna, Salamanca: Universidad de Salamanca.

Manual de mugeres en el qual se contienen muchas y diversas reçeutas muy buenas (a1525 [1995]): ed. Alicia Martínez Crespo, Salamanca: Universidad de Salamanca.

Montiño = Martínez Mo[n]tiño, Francisco (1611): *Arte de cozina, pasteleria, vizcocheria, y conserueria*. Madrid: Luis Sánchez. <http://books.google.es> [último acceso: 10/10/2017].

Nola (1520) = Robert, Mestre (1520 [1977]): *Libre del coch. Tractat de cuina medieval*. Ed. Veronika Leimgruber, Barcelona: Universitat de Barcelona/ Curial Edicions Catalanes.

Nola (1525) = Nola, Ruperto [o Ruberto] de (1525): *Libro de cozina compuesto por maestre Ruberto de Nola cozinero que fue del señor rey don Hernando de Napoles […]*. Toledo: Ramón de Petras. <www.bne.es> [último acceso: 10/10/2017].

Oficios = San Juan Bautista de la Concepción (c1607 [1998–2002]): *De los oficios más comunes de la religión de Descalzos de la SS. Trinidad*, en *Obras completa*s, vol. 3: *Espíritu de la Reforma Trinitaria*. Ed. Juan Pujana y Arsenio Llamazares, Madrid: Editorial Católica, 417–791.

Palencia = Fernández de Palencia, Alfonso (1490 [1967]): *Universal vocabulario en latín y en romance*. Reproducción facsímil. de la ed. de 1490. Nota preliminar de S. Gili Gaya, 2 vols., Madrid: Comisión Permanente de la Asociación de Academias de la Lengua Española. <www.bne.es> [último acceso: 10/10/2017].

Real Academia Española (1726–1739 [1979]): *Diccionario de autoridades*. Edición facsímil, Madrid: Gredos.

Scappi, Bartolomeo (1570): *Opera*. Venecia: Michele Tramezzino. <http://books.google.es> [último acceso: 10/10/2017].

Talavera *Tratado* = Castro, Teresa de (1477–1496 [2001]): «El *Tratado sobre el vestir, calzar y comer* del arzobispo Hernando de Talavera», *Espacio, tiempo y forma. Serie III, Historia medieval* 14, 11–92.

Torquemada, Antonio de (1553 [1995]): «Coloquio que trata de la desorden que en este tiempo se tiene en el mundo, y principalmente en la christiandad, en el comer y beber […]», en *Obras completas*, vol. 1, Madrid: Turner, 325–340.

Torre, Alfonso de la (1430–1440 [1991]): *Visión deleytable*. Ed. Jorge García López, Salamanca: Universidad de Salamanca.

Valdés, Juan de (1535–1536 [2003[7]]): *Diálogo de la lengua*. Ed. Cristina Barbolani, Madrid: Cátedra.

Vallés, Juan (a1563 [2008]): *Regalo de la vida humana*. Ed. Fernando Serrano Larráyoz, 2 vols., Pamplona/Viena: Gobierno de Navarra/Oesterreichische Nationalbibliothek.

Valles Rojo, Julio (2007): C*ocina y alimentación en los siglos XVI y XVII*. Valladolid: Junta de Castilla y León.

Varela Merino, Elena (2009): *Los galicismos en el español de los siglos XVI y XVII*. 2 vols., Madrid: CSIC.

Viaje de Turquía. Diálogo entre Pedro de Hurdimalas y Juan de Voto a Dios y Mátalas Callando […] (1560 [2000]): Ed. Marie-Sol Ortola, Madrid: Castalia.

Villena, Enrique de (1423 [1984]): *Arte cisoria*. Ed. Russell V. Brown, Barcelona: Editorial Humanitas.

José Calvo Tello, Ulrike Henny-Krahmer y Christof Schöch

Textbox: análisis del léxico mediante corpus literarios

Resumen: En este artículo presentamos *textbox*, una colección de corpus literarios históricos en varias lenguas romances (tres de ellos en español), codificada en XML-TEI, publicada bajo licencia Creative Commons y accesible en su totalidad en GitHub. En primer lugar, explicamos las razones que hicieron necesaria la creación de los corpus dentro del grupo de investigación de Wurzburgo. En segundo lugar, describimos sus características: los objetivos que cumple, su marcación, los metadatos y su publicación. En tercer lugar, analizamos lingüísticamente varias de las colecciones de novelas aunando visualización, análisis estadístico, metodologías estilométricas como Zeta y recursos léxicos como *CORDE* o diccionarios de la RAE. Con esto queremos mostrar las posibilidades de análisis de léxico que corpus accesibles en su totalidad ofrecen a la comunidad científica interesada en la investigación del léxico histórico español y de textos literarios históricos en lenguas romances en general.

Palabras clave: Corpus literarios, TEI, Estilometría, Lexicografía

Abstract: In this article we will present *textbox*, a collection of historical literary corpora in several Romance languages (three of them in Spanish), encoded in XML-TEI, published under a Creative Commons license and fully accessible on GitHub. Firstly, we will explain the motivation for the creation of the corpus within our research group. Secondly, we will describe the characteristics of the corpus: the goal, markup, metadata and its publication. Thirdly, we will provide a linguistic analysis of several of the collections of novels combining visualization, statistical analysis, stylometric methodologies such as Zeta and lexical resources such as *CORDE* and dictionaries of the RAE. The purpose of this article is to show the possibilities of lexical analysis offered by fully accessible corpora to the scientific community interested in the investigation of the historical Spanish lexicon and historical literary texts in Romance languages.

Keywords: Literary corpora, TEI, Stylometry, Lexicography

1 Introducción

1.1 Antecedentes

TEI (Text Encoding Initiative) es una iniciativa internacional comenzada por diferentes proyectos de humanidades cuyo objetivo era la definición de

un formato con el que marcar diferentes tipos de textos (Burnard/Sperberg-McQueen 2012). Desde sus inicios, hace 30 años, ha sido adoptado por miles de proyectos alrededor del mundo. Uno de los tipos de texto para el que más se utiliza es el literario, especialmente para ediciones filológicas críticas. En el caso del español podrían mencionarse proyectos como *Soledades* (Rojas Castro 2016) o las *7 Partidas Digital* (Fradejas Rueda 2017, actualmente en desarrollo). Sin embargo otros proyectos cuyo objetivo es realizar ediciones sencillas de numerosos textos también han utilizado este formato para literaturas de diferentes lenguas y épocas: *TextGrid* (TextGrid Consortium 2016) para textos literarios en alemán, *Theâtre Classique* (Fièvre 2007–2017) para el teatro francés de los siglos XVII y XVIII, o *Spectateurs* (Ertler 2013) para este género textual en diferentes lenguas europeas.

En el caso del español, otro gran proyecto también decidió utilizar TEI como su formato de codificación. Nos referimos a la Biblioteca Virtual Miguel de Cervantes, uno de los proyectos pioneros en Humanidades Digitales en español, que codificó miles de textos en XML-TEI. Lamentablemente este proyecto nunca ha puesto a disposición los archivos en su formato original, sino solamente en su versión HTML. La preferencia por la publicación en HTML o similar (como libros electrónicos) ha sido la tendencia mayoritaria en los proyectos de textos literarios en español. El humanista digital interesado en análisis textuales en esta lengua se ha tenido que familiarizar con la conversión del HTML al formato que realmente desee, ya fuese desde proyectos académicos como el Cervantes Virtual, así como aquellos realizados por instituciones o comunidades no académicas como ePubLibre o Gutenberg Project. Tanto es así que Agenjo (2015) señalaba la necesidad, hace apenas algunos años, de que la comunidad hispánica se decidiese a implementar este formato en sus proyectos.

Aun así, ciertos proyectos decidieron publicar sus textos en XML-TEI, algunos de ellos con el claro objetivo de que fuesen utilizados en análisis cuantitativos. Entre ellos caben destacar el ya mencionado proyecto *Spectateurs* o *Moralische Wochenschriften* en su título original (Ertler 2013), o los proyectos sobre poesía *ADSO* (Navarro-Colorado *et al.* 2015) y *DISCO* (Ruiz *et al.* 2017). Cada uno de estos proyectos ha publicado varios miles de textos en formato XML-TEI, añadiendo diferentes tipos de anotación y metadatos. Sin embargo las necesidades de nuestra investigación requerían de corpus con otros objetivos, por lo que fue necesario constituir nuestros propios corpus.

1.2 Necesidades

El grupo de investigación en Humanidades Digitales *Estilística computacional del género literario*, o CLiGS (por sus siglas en alemán, *Computergestützte Literarische Gattungsstilistik*), iniciado en 2014 y ubicado en la cátedra de Filología Computacional en la Universidad de Wurzburgo (Alemania), está financiado por el Ministerio de Educación e Investigación Alemán (BMBF). En este proyecto, romanistas e informáticos analizan el género literario desde metodologías cuantitativas como estilometría, machine learning, topic modeling o análisis de sentimientos.

Para la creación de los corpus que sostuviesen esta investigación se tuvieron en cuenta no solo las necesidades puntuales del proyecto, sino también buenas prácticas en la creación de corpus y colecciones (Henny-Krahmer/Neuber 2017; Jannidis *et al.* 2017), como son la publicación del formato original, la citación de las fuentes de las que procede, la identificación de manera estándar de autores y obras, la publicación del esquema de validación del corpus o la edición de los textos bajo licencias Creative Commons.

En las siguientes secciones de este trabajo queremos presentar las principales características de esta colección de corpus, centrándonos en aquellos que contienen textos en español. Posteriormente, mostraremos varios ejemplos de análisis léxico y textual que pueden ser llevados a cabo mediante metodologías cuantitativas cuando se dispone de los datos de manera libre: en primer lugar, compararemos las frecuencias léxicas del vocabulario de las novelas españolas y latinoamericanas utilizando una medida de comparación surgida en la estilometría; en segundo lugar, utilizaremos las clases semánticas señaladas en el *Nuevo diccionario histórico del español* (*NDHE*) para responder a las hipótesis de si ciertas áreas léxicas son más frecuentes en ciertos subgéneros de la novela; y en tercer lugar, compararemos los lemas de las novelas con el lemario del *Diccionario* de la Real Academia para analizar por un lado la cobertura del léxico de las novelas por el diccionario y, por otro lado, el vocabulario específico, no estándar, en el sentido de no-inclusión en el lemario, de estos textos históricos literarios.

2 Objetivos y características de *textbox*

2.1 Objetivos y selección

Actualmente, *textbox* consiste en nueve colecciones de textos literarios de varios géneros en español, francés, italiano y portugués. En total se encuentran 550

obras con aproximadamente 15 millones de *tokens*, de las cuales 139 son novelas, 311 novelas cortas y cuentos, y 100 piezas de teatro. *Textbox* está disponible en https://github.com/cligs/textbox[1].

Inicialmente, *textbox* se ha concebido como plataforma de publicación y repositorio para los corpus literarios que constituyen la base de trabajo en el grupo de investigación CLiGS. El objetivo es publicar colecciones más extensas al final del proyecto, sobre todo las colecciones que se establecen en el ámbito de los estudios de doctorado y post-doctorado. Al mismo tiempo, hemos decidido empezar con la publicación inmediata de colecciones pequeñas y parciales, por varios motivos: esta estrategia nos permite desarrollar buenas prácticas no sólo en la preparación, sino también en la publicación de textos literarios en formato digital que pueden servir de modelo tanto para el trabajo del grupo CLiGS como para otros investigadores que pueden utilizarlo y aportar *feedback*, y, además, las colecciones pequeñas pueden utilizarse ya en experimentos y servir como ejemplos en la enseñanza.

Las colecciones individuales no son muestreos representativos de la totalidad de novelas, cuentos o piezas de teatro que se publicaron en la respectiva lengua y época, sino que se crearon considerando diferentes principios de selección y escenarios de utilización. El marco general de *textbox* son todas las lenguas y literaturas romances e invitamos a otros investigadores a colaborar en las colecciones existentes o a crear nuevos corpus.

Veamos ahora los criterios y fines que guiaron la creación de las colecciones actuales en *textbox* con textos en español, en su versión 3.0:

- Las dos colecciones de novelas en español, *Corpus of Spanish Novels from 1880 to 1940* y *Collection of 19th century Spanish American Novels (1880–1916)*, forman parte de dos proyectos de doctorado dentro de CLiGS que se dedican al análisis de los subgéneros de la novela en los siglos XIX y XX en España e Hispanoamérica, respectivamente. Las selecciones publicadas hasta la fecha se compilaron para facilitar análisis de autoría y, en consecuencia, están equilibradas respecto al número de textos por autor. En la versión actual de *textbox*, la colección de novelas españolas incluye 39 novelas de 13 autores con un total de 2,5 millones de *tokens* y la colección de novelas hispanoamericanas, 24 novelas de 8 autores diferentes con un total de 1,1 millones de *tokens*.

1 El DOI que redirige a la última versión es <https://doi.org/10.5281/zenodo.597430> [último acceso: 15/10/2017].

- El *Corpus of Spanish Short Stories from 1880–1940* consta de 12 colecciones de cuentos escritos por 6 autores diferentes. En total, este corpus incluye 193 cuentos individuales y 498 000 *tokens*. Este corpus mantiene una estrecha relación con el corpus de novelas españolas.

En este artículo nos limitamos a describir las colecciones españolas. Más información acerca de las colecciones de textos franceses, portugueses e italianos se encuentra en la página principal de *textbox* en GitHub.

Las fuentes principales de las obras incluidas son portales y páginas web que ofrecen textos literarios en los formatos HTML, EPUB o texto plano. Además, se utilizaron algunas fuentes que publican en formato PDF, imagen; parte de los materiales fueron escaneados por el mismo proyecto. Fuentes generales para todos los idiomas fueron Wikisource y el Proyecto Gutenberg. Una fuente muy importante para los textos en español fue la mencionada Biblioteca Virtual Miguel de Cervantes, pero también ePubLibre.

Si bien estas colecciones se han creado según algunos criterios primarios de selección, todas se pueden transferir o adaptar para diversas situaciones de uso, siendo el análisis del léxico español una de ellas, como demostraremos más adelante en la sección 3.

2.2 Textos y marcado

Todos los textos recopilados a partir de las diversas fuentes se prepararon según el esquema de datos común de *textbox* basado en XML-TEI P5[2]. El uso de un esquema común para todas las colecciones en *textbox* es ventajoso porque asegura una codificación de datos homogénea a través de las diversas colecciones. Esto permite, entre otras cosas, desarrollar scripts de conversión y programas de análisis comunes.

Utilizamos scripts en Python con expresiones regulares y XPath para transformar los textos fuentes en formato texto plano, HTML y EPUB a TEI. En el caso de fuentes en formato PDF o archivos de imagen, la conversión del texto a TEI está precedida por un procesamiento OCR.

Decidimos utilizar XML como formato base porque permite combinar los metadatos (en el TEI header) con los mismos textos (en el TEI body), lo que asegura la integridad de ambos. En el caso de los metadatos, XML permite registrarlos de una manera más detallada y más estructurada que, por ejemplo, un

2 El esquema de datos de *CLiGS textbox* está disponible en <http://github.com/cligs/reference/> [último acceso: 15/10/2017].

simple formato tabular. Con respecto a los textos, XML permite captar y preservar informaciones estructurales como títulos, capítulos y párrafos.

Ofrecemos dos formatos derivados del formato de base en TEI. Primero, una versión anotada con informaciones lingüísticas creada con la herramienta FreeLing (Padró/Stanislovsky 2012), que incluye las formas lematizadas de las palabras, informaciones acerca de su categoría gramatical y anotaciones semánticas básicas derivadas de WordNet (Miller 1995; Fellbaum 1998). El formato anotado es adecuado como punto de partida para muchos análisis que dependen de las informaciones lingüísticas, por ejemplo, para análisis léxicos que recurran a lemas. Segundo, incluimos una versión de los textos sin marcado estructural y sin metadatos, esto es, solo en texto plano completo, ya que la mayoría de las herramientas de análisis textuales solo aceptan texto plano como formato de entrada.

2.3 Metadatos

Los metadatos que registramos se dividen en administrativos y descriptivos. Los primeros incluyen el nombre del investigador responsable de la creación del archivo TEI, su fecha de creación, informaciones legales e informaciones sobre las condiciones de su uso, así como la documentación de modificaciones hechas posteriormente a su creación. Otro metadato administrativo importante es el identificador del texto que está compuesto por dos letras, indicando la colección de textos (por ejemplo «ne» para «novelas españolas») y un número consecutivo de cuatro dígitos (por ejemplo «ne0001», «tc0120»). Los campos más importantes de los metadatos descriptivos son los siguientes:

- *autoría*: se indican el nombre del autor, su identificador VIAF, continente y país de nacimiento.
- *obra*: la obra está descrita por su título, subtítulo e identificador (VIAF o biblioteca nacional).
- *ediciones*: se registra la primera edición de la obra y la edición digital o impresa del texto que se utilizó como base para el archivo TEI.
- *género*: el género de la obra literaria se describe a diferentes niveles. El campo «supergénero» sirve para distinguir entre los géneros principales (narrativo, drama, lírico); el campo «género», para distinciones dentro de un género principal, por ejemplo novela frente a cuento en la narrativa; el campo «subgénero», para subtipos dentro de un género, por ejemplo tragedia, comedia, tragicomedia en drama, o bien novela histórica o novela sentimental como subtipos de la novela.

- *contenido del texto*: para algunas de las colecciones se registran, además, metadatos que describen aspectos relacionados con el contenido de los textos, los cuales pueden ser útiles en la interpretación de los resultados del análisis. Estos campos son, entre otros, un resumen del contenido del texto, la perspectiva narrativa (primera, segunda o tercera persona), el género del protagonista (femenino, masculino o mixto) y el tipo de lugar en el que se desarrolla la acción (urbano o rural).

2.4 Publicación

La publicación de las colecciones en *textbox* se basa en dos infraestructuras que ofrecen, por un lado, la flexibilidad deseable en la administración de los archivos y, por otro, la seguridad necesaria para su almacenamiento a largo plazo. Las colecciones se publican en un repositorio público GitHub, una plataforma de desarrollo colaborativo con un control automático de versiones. Así, todos los cambios en los metadatos, en el marcado y en las transcripciones de los textos, son documentados y el trabajo tanto de crear como de desarrollar las colecciones se puede realizar en equipo. GitHub permite además recopilar sugerencias y discusiones acerca de las colecciones en los llamados «issues» y ofrece un mecanismo para realizar copias de las colecciones y reutilizarlas («clone», «fork» o descarga)[3].

Al ser GitHub una entidad comercial que no asegura la disponibilidad de las colecciones a largo plazo, definimos versiones estables de las colecciones (las llamadas «releases» en GitHub) periódicamente y las archivamos en Zenodo. Este es un servicio para almacenamiento de datos a largo plazo creado para investigadores en Europa, administrado por OpenAire y apoyado por CERN (Nielsen 2013).

Los textos están en dominio público bajo una licencia Creative Commons Attribution (CC-BY) que permite su uso libre. Escogimos la licencia de atribución solamente para animar a los usuarios de las colecciones a citar estas colecciones, de manera que se reconozca el esfuerzo hecho en la creación de textos íntegros, los marcados TEI y el conjunto de metadatos detallados.

3 <https://guides.github.com/introduction/git-handbook/> [último acceso: 15/10/2017].

Con esta descripción esperamos haber cubierto las principales características de *textbox*. En las siguientes secciones se ofrecen algunos ejemplos de análisis del léxico que se pueden realizar a partir de este proyecto.

3 Ejemplos de análisis digital

3.1 Pyzeta

Zeta es una medida de distintividad (en inglés también referido como *keyness*) propuesta por John Burrows (2007). Este tipo de medidas se utiliza para identificar rasgos característicos (por ejemplo, palabras) de un grupo de textos en comparación con otro (Scott 1997). Zeta es ampliamente utilizada por la estilometría y los Estudios Literarios Digitales, por su sencillez matemática, su preferencia por palabras de significado altamente interpretable y su implementación en la herramienta estilométrica en R *stylo* (Eder *et al.* 2016). Algunos ejemplos de aplicación de Zeta a problemas de historia de la literatura pueden verse en Craig/Kinney (2009), Hoover (2010) o Schöch (2017).

Recientemente nuestro proyecto ha desarrollado una nueva implementación de Zeta en Python llamada *pyzeta*, accesible mediante GitHub[4]. A diferencia de *stylo*, esta nueva implementación tiene como objetivo poder experimentar, evaluar y analizar de qué manera afecta tanto la variación de aspectos concretos del cálculo de Zeta como de parámetros (ej: la extensión de los segmentos textuales) (Schöch/Zehe/Calvo Tello 2018).

Para este análisis comparamos 24 novelas de España con 24 novelas de Hispanoamérica (Argentina, Cuba y México). Utilizamos los textos lematizados (con Freeling) que se encuentran en el repositorio, sobre los que se calculó el valor zeta de cada lema (parámetros: la medida zeta en su versión original, fragmentos de texto de 10 000 *tokens*). El resultado puede ser visualizado de diferentes maneras. En primer lugar, podemos observar los veinte lemas más distintivos tanto para el español de España (los valores en azul) como para el español de América (valores en verde). El eje horizontal representa el valor zeta de cada lema, siendo posibles los valores entre -1 y 1:

4 <https://github.com/cligs/pyzeta> [último acceso: 15/10/2017].

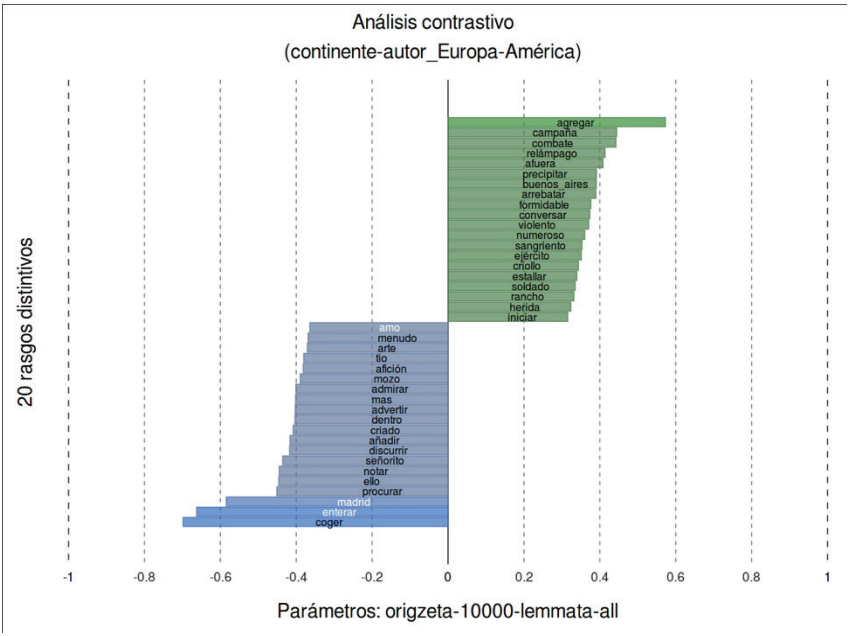


Imagen 1: Valores zeta para las veinte palabras más distintivas para España e Hispanoamérica

Coger, enterar y *Madrid* son los rasgos que mejor caracterizan las novelas de España, frente a *agregar, campaña* y *combate*, que serían los tres rasgos que mejor caracterizan las novelas hispanoamericanas.

Sin embargo mediante esta visualización no sabemos en qué manera los lemas aparecen en el otro grupo de novelas. *Coger* es un españolismo, ¿pero significa esto que no aparece en las novelas hispanoamericanas? Para responder a esta pregunta, podemos observar los valores mediante un *scatter plot* en el que los ejes representen la proporción de segmentos de textos en los que aparece la palabra en cada uno de los grupos. Podemos observar que la palabra más cercana a la esquina superior izquierda es, efectivamente, *coger*, con una posición cercana a 0.8 en eje vertical y a 0.1 en el eje horizontal:

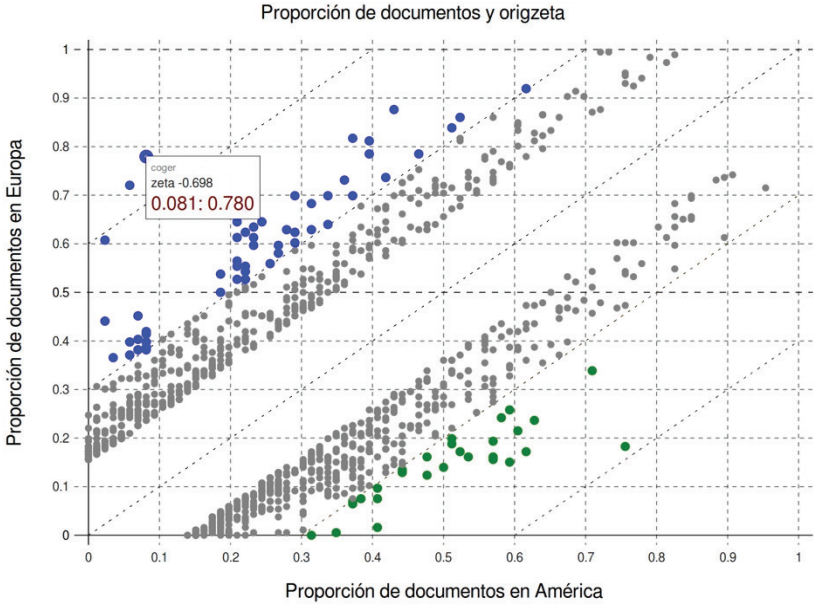


Imagen 2: Distribución de proporción en segmentos de texto de España e Hispanoamérica de los 500 lemas con valores zeta más extremos

Esto significa que *coger* aparece en el 80 % de segmentos de las novelas de España, mientras que solo aparece en el 10 % de segmentos de novelas hispanoamericanas. En el lado contrario del gráfico, el lema más cercano a la esquina inferior derecha es *agregar*, presente en más del 75 % de segmentos de novelas hispanoamericanas y tan solo en el 20 % de segmentos de novelas de España.

Observando los lemas de ambos grupos, podemos preguntarnos si estos resultados realmente aportan alguna información sobre las variedades lingüísticas geográficas, si esta información en realidad solo es representativa del lenguaje literario de prosa, o incluso si exclusivamente estamos observando fenómenos de las colecciones específicas de *textbox*. Las novelas de España parecen contener un vocabulario más burgués (*Madrid, señorito, mozo, afición, tío, arte*) frente al más vehemente de Hispanoamérica (*campaña, combate, arrebatar, violento, numeroso, sangriento, ejército, soldado*). ¿Son estas diferencias realmente lingüísticas?

Para evaluar esta pregunta, hemos comparado los valores zeta de los 40 lemas[5] mostrados en la primera figura (20 lemas con los valores zeta más altos y las 20 con valores más bajos) con el porcentaje que el corpus diacrónico *CORDE* aporta para España. El *CORDE* señala en su documentación que el 74 % del corpus proviene de España, por lo que ese sería el valor esperable del porcentaje de apariciones que una palabra debería tener si es utilizada por igual tanto en España como en Hispanoamérica. De cualquier manera, es posible que este porcentaje sufra ciertas modificaciones en subsecciones del corpus, por lo que debemos tomarlo como información útil pero inexacta. Este valor se muestra en las dos siguientes gráficas como líneas gruesas horizontales. En la siguiente gráfica observamos como boxplots el porcentaje de aparición en España dentro del *Corpus Diacrónico del Español –CORDE–* (1880–1940, textos narrativos) de cada uno de los 40 lemas característicos según zeta para las novelas hispanoamericanas y españolas:

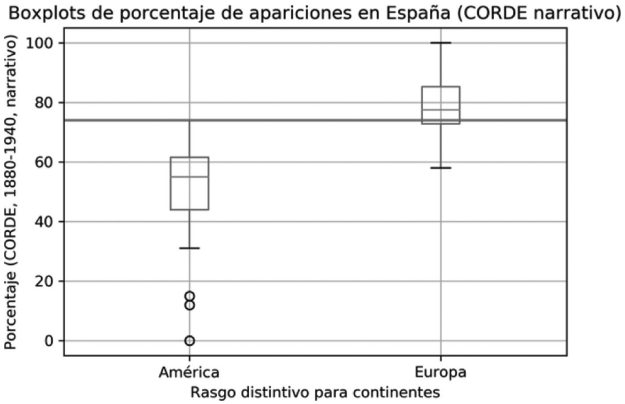


Imagen 3: Porcentaje de aparición en España de lemas distintivos (*CORDE* narrativo)

5 Si hubiésemos dispuesto de los textos de *CORDE* de manera abierta, mediante API o mediante listas de frecuencias léxicas por cada texto, podríamos haber evaluado una mayor cantidad de lemas. Al tener que hacer las búsquedas de manera manual, reducimos la cantidad a algo manejable.

Aunque algunas palabras que distinguen el español europeo están por debajo del 74 % señalado como base, en general se observa que las palabras identificadas por Zeta como españolismos lo son también en el *CORDE* narrativo. Esto indica que los rasgos léxicos arriba mencionados no son solo más frecuentes en nuestra colección o en los subgéneros de novela aquí representados, sino en general en la narrativa española e hispanoamericana de este período.

En cuanto a si estos rasgos son, en realidad, fruto de distinciones lingüísticas y no solo están presentes en la lengua literaria, hemos realizado la misma búsqueda dentro del *CORDE* general:

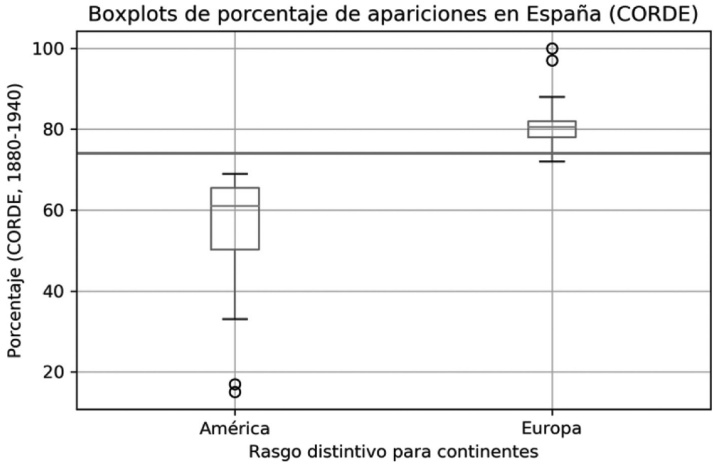


Imagen 4: Porcentaje de aparición en España de lemas distintivos (*CORDE*)

En este caso, los dos grupos de rasgos no se sobreponen: los lemas identificados como rasgos para diferenciar novelas españolas e hispanoamericanas efectivamente tienen rangos de frecuencias diferentes en el *CORDE* general, no solo en textos literarios de prosa.

3.2 Análisis de áreas léxicas del *NDHE*

En los últimos años diferentes metodologías provenientes del área de *Information Retrieval* se han centrado en el análisis distribucional del léxico en grandes corpus. Dependiendo de los objetivos que persiguiesen, el algoritmo puede representar cada palabra en un espacio de *n* dimensiones en el que son posibles

cálculos matemáticos básicos entre palabras, como la medición de su distancia, manipulación numérica, etc. (word embeddings como word2vec). Otro grupo de algoritmos denominados *topic modeling* (Blei 2012) han centrado su objetivo en buscar *clusters* de palabras que tiendan a coaparecer entre ellas en diferentes segmentos de texto. Sin embargo, el hecho de que un grupo de palabras coaparezca en diferentes textos no tiene que representar afinidad semántica, sino algún tipo de característica compartida (semántica, pragmática, textual, errores de OCR…).

Para representar temas que contengan menos ruido necesitaríamos que especialistas determinasen qué temas existen y que clasificasen cada palabra (en realidad cada acepción de la palabra) en relación a esos temas o grupos semánticos. Precisamente, esta es una de las tareas que ha abordado el *Nuevo diccionario histórico del español* (*NDHE* desde ahora), un proyecto dirigido por José Antonio Pascual y coordinado por Mar Campos Souto dentro del Instituto Rafael Lapesa de la RAE (Pascual/García 2008; Pascual/Souto 2014). Este proyecto ha decidido comenzar su macroestructura no por orden alfabético sino por una serie de familias o grupos semánticos, de las que actualmente están más desarrolladas las enfermedades, armas, instrumentos de medida e instrumentos musicales. Agradecemos a los responsables del *NDHE* que nos hayan facilitado un listado completo de los lemas y su pertenencia al grupo semántico, datos que han hecho posible esta sección del análisis.

Estos listados de palabras con características semánticas comunes nos permiten poder analizar de manera cuantitativa si ciertos temas son más frecuentes en ciertos subgéneros que en otros. Para esta sección hemos trabajado con la colección de novelas hispanoamericanas, sobre la que hemos querido evaluar tres hipótesis:

1. ¿Es el léxico de enfermedades en las novelas naturalistas estadísticamente más frecuente que en el resto de novelas?
2. ¿Es el léxico de armas en las novelas gauchas estadísticamente más frecuente que en el resto de novelas?
3. ¿Es el léxico de armas en las novelas históricas estadísticamente más frecuente que en el resto de novelas?

Para responder a cada una de estas preguntas partimos de nuevo desde la versión lematizada de nuestra colección y extrajimos las frecuencias léxicas relativas de cada lema por cada novela. Posteriormente, filtramos los lemas manteniendo solo aquellos que pertenecen a un grupo semántico específico en el *NDHE*. Por último, calculamos la media de la frecuencia relativa de los lemas del grupo semántico estudiado en cada novela. Estos valores fueron

divididos en dos grupos: los que pertenecen al subgénero analizado y los que no. Cada uno de los grupos contiene un valor de frecuencia del área léxica específica por cada novela.

Veámoslo con el ejemplo de léxico de enfermedades en las novelas naturalistas. De las 481 lemas que el *NDHE* contiene como pertenecientes a enfermedades, solo 28 aparecen en la colección de novelas hispanoamericanas de *textbox*. De esta manera, de los miles de lemas que contiene cada novela, decidimos centrar nuestro análisis exclusivamente en estos 28 términos[6]. Calculamos entonces la frecuencia relativa media del conjunto de términos en cada novela. De esta manera conseguimos una serie de valores independientes entre sí. A partir de ellos, podemos comparar los dos subgéneros y visualizarlo como boxplot:

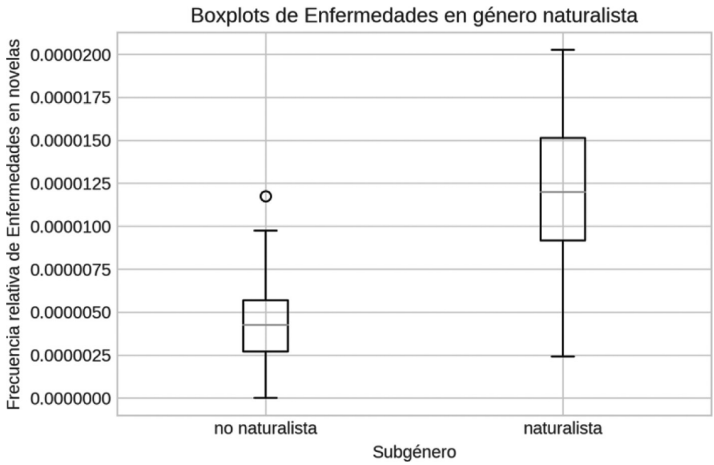


Imagen 5: Frecuencia de léxico de enfermedades en novelas naturalistas y no naturalistas

6 En concreto son: *apestar, catarral, catarroso, crup, diarrea, diarreico, difteria, epilepsia, epiléptico, esclerosado, hidrofobia, influencia, lepra, leproso, peste, pestilencia, pulmonar, pulmonía, pulmón, reumático, sarampión, sarna, sarnoso, sifilítico, sífilis, tetánico, tifo* y *tétano.*

Como se puede observar, la frecuencia del léxico de enfermedades en novelas naturalistas es notablemente superior a la del resto de novelas, aunque también hay cierta área donde ambos grupos de novelas comparten valores. Por eso es necesario realizar un test estadístico como t-test[7]. Para este caso específico, el test arroja un valor *p* de 0.0007, muy por debajo del estándar de 0.05, lo que señala que efectivamente el léxico de las enfermedades es significativamente más frecuente en las novelas naturalistas de nuestro corpus que en el resto de novelas.

Hemos formulado otras dos hipótesis en relación a las novelas gauchas e históricas y la frecuencia de léxico sobre las armas. Los resultados en estos casos son los siguientes:

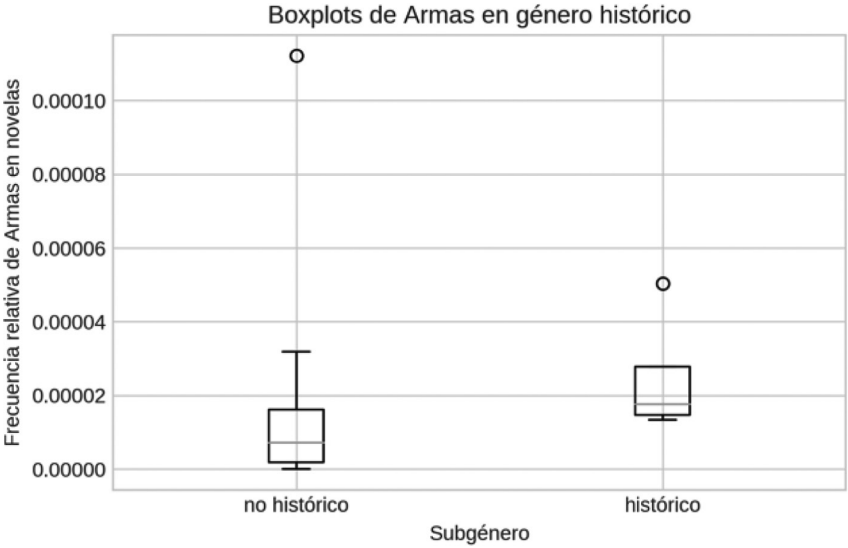


Imagen 6: Frecuencia de léxico de armas en género gaucho e histórico frente al resto

7 Hay que señalar que los datos de ambos grupos disponen de una distribución normal, por lo que se cumplen los requisitos para este test.

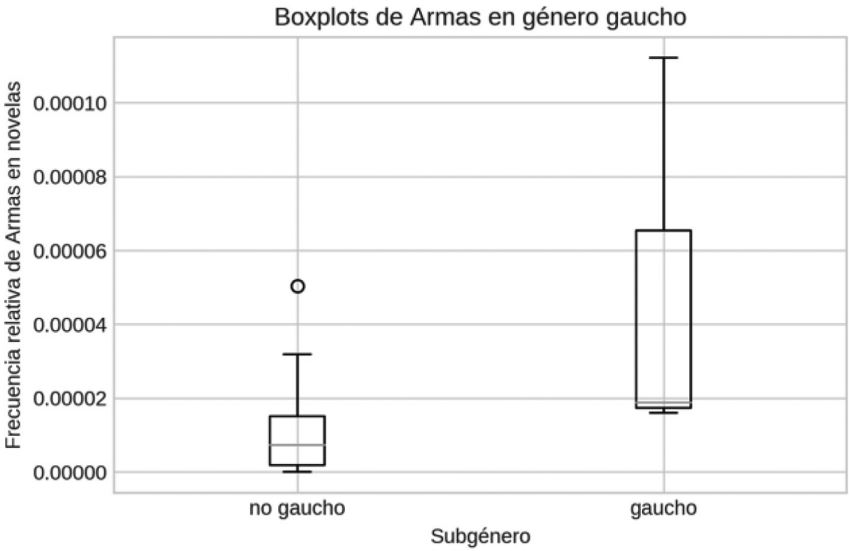


Imagen 7: Frecuencia de léxico de armas en género gaucho e histórico frente al resto

Como se observa, el léxico de las armas es más frecuente (tanto en la novela gauchesca como en la novela histórica) que en el resto de novelas, sin embargo las diferencias de ambos grupos son menos claras que en el caso anterior. Además, los valores no están distribuidos de manera normal, por lo que no se cumplen los criterios para realizar el t-test. Por esta razón hemos realizado el test no paramétrico Mann-Whitney-Wilcoxon. Para ambos, el valor *p* es menor de 0.05 (0.02 y 0.03 respectivamente), por lo que en ambos casos podemos confirmar que el vocabulario de las armas es significativamente más frecuente en los subgéneros de novela gaucho e histórico que en el resto.

3.3 Comparación del léxico con el *DLE*

El *Diccionario de la lengua española* de la Real Academia (*DLE*) se encuentra hoy en su vigesimotercera edición, publicada en el año 2014. La edición actual registra más de 93 000 lemas. El *DLE* se puede considerar el más importante diccionario de referencia y consulta del español. Ya desde 1884 se trabaja en la inclusión del léxico procedente de países hispanoamericanos y en la presente edición se señala que se ha procedido a una «revisión del tratamiento de las

marcas geográficas americanas y la de los extranjerismos»[8]. Por lo dicho, suponemos que el *DLE* se presta para analizar el léxico de los textos incluidos en *textbox* procedentes de diferentes países hispanohablantes. En concreto, queremos investigar las siguientes cuestiones:

- *Textbox* reúne obras literarias. Una comparación con el *DLE* puede poner de relieve las características léxicas específicas de este tipo de textos. ¿Qué tipo de palabras aparecen en los textos literarios pero no en el *DLE*? ¿Hay entre ellos palabras formadas por procesos creativos? Los resultados se pueden analizar diferenciados por autores individuales para investigar su léxico específico.
- Los textos en *textbox* son históricos. De las palabras que se encuentran en los textos de *textbox* y también en el *DLE*, ¿cuáles están marcadas como desusadas presumiblemente por ser históricos?
- Como *textbox* incluye obras de Hispanoamérica, ¿qué diferencias hay entre los textos de España y los textos de América cuando se compara su léxico con el *DLE*? ¿Qué porcentaje del léxico está marcado como proveniente de España o América, dialectal o coloquial?

El corpus de trabajo para este análisis se ha compilado desde las mismas colecciones en *textbox* que los análisis en *3.1* y *3.2*. Aquí se ha escogido un subconjunto de textos provenientes de Argentina y una igual cantidad de textos de España, 36 novelas en total. Cada autor (Alarcón, Bazán, Miró, Pereda, Valera, de España; Bunge, Cambaceres, Gutiérrez, Holmberg, Payró y Sicardi, de Argentina) está representado por tres obras. El corpus está equilibrado en cuanto al número de textos por autor y no tiene desequilibrios significativos en cuanto a los subgéneros de los textos.

Nuestra estrategia para extraer las informaciones lexicográficas del *DLE* es la siguiente: a partir del eBook del diccionario, convertimos el HTML a un formato XML simple mediante scripts XSLT. Los campos que podíamos extraer son el lema, la categoría gramatical de la palabra, si tiene una acepción con marca geográfica (España, Argentina o América), con marca de nivel de uso (coloquial, dialectal, desusado)[9]. Los textos del corpus de trabajo se utilizaron en su formato

8 <http://www.rae.es/diccionario-de-la-lengua-espanola/presentacion> [último acceso: 15/10/2017].

9 Esta estrategia de transformación es improvisada. Idealmente, si la RAE lo ofreciese, se podría utilizar una *interfaz de programación de aplicaciones* (API) para incorporar el diccionario a análisis de texto computacionales. Desafortunadamente, la función de búsqueda ofrecida por la RAE en la página web del diccionario solo se puede utilizar manualmente. Existen algunas interfaces alternativas del diccionario

TEI anotado con FreeLing para poder acceder a los lemas y las categorías gramaticales de las palabras. Por supuesto, este proceso de anotación reduce el número de palabras porque no todas son identificadas por el *tagger*. Este análisis se basa por consiguiente en el léxico más convencional y reconocible estructuralmente. Por ser un análisis del léxico, nos concentramos en las categorías gramaticales sustantivo, verbo, adjetivo y adverbio, dejando de lado las otras categorías.

El diccionario cuenta con 54 048 sustantivos (61 %), 12 140 verbos (14 %), 24 586 (28 %) adjetivos y 2148 (3 %) adverbios que suman 88 161 entradas[10]. En el corpus de trabajo se encuentran 17 185 sustantivos (55 %)[11], 6018 verbos (19 %), 5997 adjetivos (19 %) y 1920 adverbios (6 %), 31 120 lemas en total. Podemos ver que el número de lemas en las novelas aquí analizadas alcanza más o menos un tercio del número de lemas en el *DLE*. En cuanto a las proporciones de categorías de palabras, es notable que el porcentaje de adverbios es el doble en las novelas que en el diccionario mientras que el porcentaje de adjetivos es mucho más bajo. La mitad de los adverbios no encontrados en el *DLE* son formas complejas como «de hoy en adelante», que en el *DLE* no tienen estatus de lema. Un tercio termina en *-mente* y es por tanto probable que sean lemas producidos por un proceso productivo de formación de palabras y que por consiguiente no se encuentren recogidos en el diccionario[12]. Solo un análisis más detenido de los tipos de palabras en el *DLE* frente al corpus de trabajo podrá esclarecer los restantes casos de diferencias en la distribución de categorías de palabras entre ambos.

3.3.1 Palabras del corpus que (no) se encuentran en DLE

Analicemos primero cuáles de las palabras en los textos literarios se encuentran en el *DLE*. El 74 % de los lemas en el corpus tiene una entrada en el diccionario. Estos lemas constituyen el 97 % de los *tokens* de tipo sustantivo, verbo, adjetivo y adverbio, lo que es una buena cobertura. Diferenciado por categoría de palabra, la cobertura es:

(véanse <https://dirae.es/> y <http://recursosdidacticos.es/goodrae/> [último acceso: 15/10/2017]), pero tampoco ofrecen API.

10 La suma de entradas solo incluye las categorías de palabras sustantivo, verbo, adjetivo y adverbio, así que el número de entradas en el diccionario completo es mayor: 98.184. Los porcentajes se suman a más de 100 % porque una entrada puede tener varias acepciones con categorías de palabras diferentes.

11 No se consideran nombres propios.

12 En general, el *DLE* sí incluye adverbios en *-mente*, por ejemplo «definitivamente», «detenidamente». Presumimos que son los que ofrecen un estatus más lexicalizado y de uso más frecuente.

Tabla 1: Cuotas de lemas y *tokens* del corpus comparado con el *DLE*

Categoría gramatical	**Lemas en *DLE***	***Tokens* en *DLE***
sustantivo	77 %	98 %
verbo	69 %	98 %
adjetivo	79 %	96 %
adverbio	48 %	92 %
total	74 %	97 %

La diferenciación por categoría de palabra muestra que, en cuanto a lemas, la mejor cobertura es de los adjetivos y la peor la de los adverbios. Si se toma en cuenta el número de *tokens* en vez de lemas, la mejor cobertura corresponde a los sustantivos y verbos, y la peor a los adverbios otra vez. Los números varían más en la cobertura de lemas (entre 48 y 79 %) que en los *tokens* (entre 92 y 98 %). Esto se puede interpretar de manera que el *DLE* cubre el vocabulario más común y frecuente, pero no todo tipo de vocabulario especial y menos frecuente, como es de esperar.

Podemos diferenciar esos resultados todavía más, considerando las diferencias entre los textos provenientes de España y Argentina:

Tabla 2: Cuotas de lemas y *tokens* del corpus comparado con el *DLE* para textos españoles y argentinos

País	**Categoría gramatical**	**Lemas en *DLE***	***Tokens* en *DLE***
España	sustantivos	81 %	97 %
España	verbos	75 %	97 %
España	adjetivos	83 %	96 %
España	adverbios	52 %	92 %
ESPAÑA	TOTAL	78 %	96 %
Argentina	sustantivos	85 %	98 %
Argentina	verbos	81 %	99 %
Argentina	adjetivos	86 %	97 %
Argentina	adverbios	55 %	93 %
ARGENTINA	TOTAL	82 %	98 %

En total, y sorprendentemente si se considera que el *DLE* es históricamente un diccionario del español peninsular, la cobertura es ligeramente mejor para las novelas de Argentina que para las de España. La distribución entre las categorías de palabras no difiere tanto del panorama general, sin diferenciación entre países.

Más allá de los números, pasamos a analizar rasgos cualitativos de las palabras no cubiertas por el *DLE*. ¿Qué tipo de palabras son? ¿Podemos ver un uso innovador de voces en los textos literarios? ¿Hay diferencias entre los autores en cuanto al uso de palabras novedosas? En total, 8144 (26 %) de los lemas y 29 418 (3 %) de los *tokens* en el corpus no se encuentran en el *DLE*. Para analizar esas palabras, hay que revisarlas manualmente. Obviamente, no podemos hacerlo para todos los casos por lo que analizamos las 100 más frecuentes. Los resultados están sintetizados en la siguiente tabla:

Tabla 3: Tipos de lemas del corpus que no se encuentran en el *DLE*, 100 MFW

Tipo de palabra	**Ejemplos**	**Lemas**	***Tokens***
palabra con ortografía histórica	á, fué, habia, luégo, oir	22	6871
forma compleja	a cada paso, hacer cargo, pues bien, tal vez, tomar parte	49	3851
número	1, 2, 3, 4	5	2270
lematización	acurrucar, bajito, milicos, prostituta, pulperías	12	538
expresión oral	ju ju, pa, usté, verdá	4	350
tratamiento	donna, misia, ño, pae, señá	5	267
forma histórica	díjole, púsose	2	71
derivación	apresuradamente	1	36
total		100	14 254

De las palabras que no se encuentran en el *DLE*, los 100 lemas más frecuentes representan más o menos la mitad de los *tokens* no cubiertos por el diccionario (≈ 14 000 *tokens*). Para interpretar la lista asignamos un «tipo de palabra» a cada entrada que intenta explicar el porqué no se encuentra en el *DLE*.

La mayoría de los *tokens* son palabras con ortografía histórica como «fué» o «habia». Estas formas indican cómo interactúan el POS-tagger y el diccionario: FreeLing reconoce las formas con ortografía histórica y les asigna un lema correspondiente pero este lema no aparece en el *DLE*. Más allá de mostrar las dependencias entre los varios recursos lingüísticos utilizados, los resultados de la tabla 4 revelan las características del corpus. Entre las novelas en *textbox* hay varias que se basan en ediciones con una ortografía histórica. Para registrar las palabras con una ortografía que difiere de la norma actual ya hemos desarrollado un control ortográfico («spellcheck», véase Henny/Schöch 2016), pero todavía se necesita un esfuerzo para normalizar todas las formas ortográficas.

Alternativamente, sería necesario adaptar el POS-tagger de manera que asigne lemas normalizados a formas históricas.

En segundo lugar están las formas complejas, en su mayor parte adverbios, pero también formas verbales, como «tal vez» o «tomar parte». Estas sí aparecen en el *DLE*, pero no como lema sino como acepciones diferentes y adicionales al lema principal. Otra vez, el POS-tagger interpreta las formas complejas como lemas individuales, por lo que la comparación entre el *DLE* y el corpus falla en estos casos. Para poder comparar también las formas complejas sería necesario adaptar el POS-tagger o bien ser capaz de tomar en cuenta las acepciones diferentes del *DLE*. Lo último solo sería posible con una interfaz de programación adecuada o una versión digital del diccionario marcada estructural y semánticamente y, en consecuencia, reutilizable para este tipo de análisis léxico. En tercer lugar quedan los números, también como casos excepcionales. Y siguen las formas en las que la estrategia de lematización de FreeLing difiere otra vez de los lemas establecidas en el *DLE*, por ejemplo verbos reflexivos («acurrucar» frente a «acurrucarse»), diminutivos («bajito» frente a «bajo»), plurales («milicos», «pulperías» frente a «milico», «pulpería») o formas femeninas («prostituta» frente a «prostituto»).

Casos más interesantes son las expresiones orales que aparecen en las novelas («ju ju», «pa», «usté», «verdá») y los tratamientos como «misia» y «ño» porque son palabras realmente fuera del alcance del diccionario e indicadores del estilo utilizado en los textos literarios. Finalmente, también hay construcciones verbales históricas («díjole», «púsose») y formaciones de palabras regulares como la construcción de adverbios en *-mente* («apresuradamente»). En conjunto, el análisis de las palabras más frecuentes no encontradas en el *DLE* revela más las características de los datos (las formas de las palabras en el corpus) y de las herramientas (las estrategias de lematización en FreeLing frente al *DLE*) que particularidades estilísticas de los textos literarios en *textbox*. Por ende, con el estado actual de las herramientas y los datos, un análisis comparativo del léxico como el que realizamos aquí puede en primer lugar servir para mejorar estos. En segundo lugar, podemos tratar de superar las influencias técnicas para llegar a resultados significativos desde las perspectivas lingüística y literaria.

Escogimos dos autores específicos para analizar su vocabulario «no académico»: el argentino Eduardo Ladislao Holmberg (1852–1937) y la española Emilia Pardo Bazán (1851–1921). Ambos fueron representantes del naturalismo (Levine/Novoa 2012, Baquero Goyanes 1986). Holmberg escribió las primeras novelas de ciencia ficción en Argentina (Gasparini 2008) así como novelas policíacas. Analizamos el vocabulario de los dos autores que no figura

en el *DLE* de manera cualitativa, centrándonos en las palabras que destacan desde el punto de vista estilístico, sin repetir los fenómenos generales que ya se discutieron más arriba (véase tabla 3). Los resultados están sintetizados en las tablas 4 y 5:

Tabla 4: Palabras que no se encuentran en el *DLE*: Holmberg

Autor	Tipo de palabra	Ejemplos	Lemas	*Tokens*
Holmberg	extranjerismo	aurum, my lord, robe-de-chambre, rotissérie, spleen	11	12
Holmberg	derivación	desprecavidamente, enchalecar, fisionómico, mediumnidad, pesquisante	15	17
Holmberg	composición	ante-marcial, antero-posterior, semiconvicción, semi espontáneo	4	5
Holmberg	término especializado	faquirismo, filodendro, fosfóreo, giratriz, soporífico	8	8
Holmberg	neologismos y erratas	etmóidues, paralelipípedo	2	2

Tabla 5: Palabras que no se encuentran en el *DLE*: Pardo Bazán

Autor	Tipo de palabra	Ejemplos	Lemas	*Tokens*
Pardo Bazán	diminutivo	arroyuelo, carita, golpecito, nubecilla, pequeñín	44	60
Pardo Bazán	aumentativo	bigardón, caseretón, ricachón	5	5
Pardo Bazán	extranjerismo	champagne, esprit, highlife, kummel, nobis	31	37
Pardo Bazán	derivación	alguacilatos, aportuguesada, barbacanesca, enarcamiento, noticierismo	22	28
Pardo Bazán	composición	autoconfesión, neosegundo, semisueño, ultrarrefinado, verdiblanco	14	15
Pardo Bazán	término especializado	cabrifollo, genitriz, lamanisco, literalismo, psicalgia	9	11
Pardo Bazán	coloquial	hecho añicos, miquitrefe, pillete, pilluelo, piperete	7	13

Las tres novelas de Holmberg en el corpus incluyen 5213 lemas y 36 026 *tokens* de tipo sustantivo, verbo, adjetivo y adverbio, de los cuales 354 lemas (7 %) y 728 (2 %) *tokens* no se encuentran en el *DLE*. Las novelas de Pardo Bazán suman 12 484 lemas y 88 718 *tokens*. 1014 (8 %) de los lemas y 1932 de los *tokens* (2 %) no están registrados en el *DLE*. En el caso de Holmberg, algunos tipos de palabras que saltan a la vista son términos provenientes del latín, del francés e inglés como por ejemplo «aurum», «robe-de-chambre», «spleen». También en las novelas de Pardo Bazán hay extranjerismos provenientes de estas lenguas, además del «kummel» (licor de comín en alemán). La utilización de extranjerismos y, sobre todo, de palabras de moda del francés o el inglés es un rasgo conocido de las novelas de fin de siglo. Tanto Holmberg como Pardo Bazán utilizan las estrategias de derivación y composición para formar palabras. En Holmberg, algunos de esos términos están relacionados con la temática de las novelas de ciencia ficción, como por ejemplo «mediumnidad», «ante-marcial» o, en el caso de los neologismos, «etmóidues» y «paralelipípedo». En este autor también hay términos especializados como la «fuerza giratriz» o un olor «fosfóreo», que se deben al carácter científico de las obras. El vocabulario especializado de Pardo Bazán se puede interpretar en relación con las temáticas naturalistas, por ejemplo «genitriz» (en lugar de «generatriz») para designar a una madre, o «psicalgía» que se refiere a dolores de origen psicógeno. Por otra parte, Pardo Bazán utiliza muchos diminutivos, algunos aumentativos y palabras coloquiales que podemos interpretar como efecto de descripciones directas del entorno social. Algunas de las palabras que utiliza Pardo Bazán y que no se encuentran en el *DLE* tienen un origen gallego como «lamanisco»[13] y «cabrifollo»[14]. Mientras que el análisis global y cuantitativo de las palabras del corpus no encontradas en el *DLE* puso de manifiesto sobre todo las desviaciones ortográficas y gramaticales del corpus así como el funcionamiento de la lematización en FreeLing y en el *DLE*, el análisis del vocabulario excepcional de autores concretos sí contribuye a descubrir particularidades estilísticas de estos textos literarios.

13 «Tecido parecido ao damasco que se emprega para confeccionar manteis, toallas etc.», *DRAG* <http://academia.gal/dicionario/-/termo/busca/lamanisco> [último acceso: 15/10/2017].

14 «Planta da familia das caprifoliáceas que gabea polos muros e os troncos das árbores, de talo longo e fino, con ramas desde a base e flores amarelas que forman grupos e teñen un recendo intenso e agradable», *DRAG* <http://academia.gal/dicionario/-/termo/busca/cabrifollo> [último acceso: 15/10/2017].

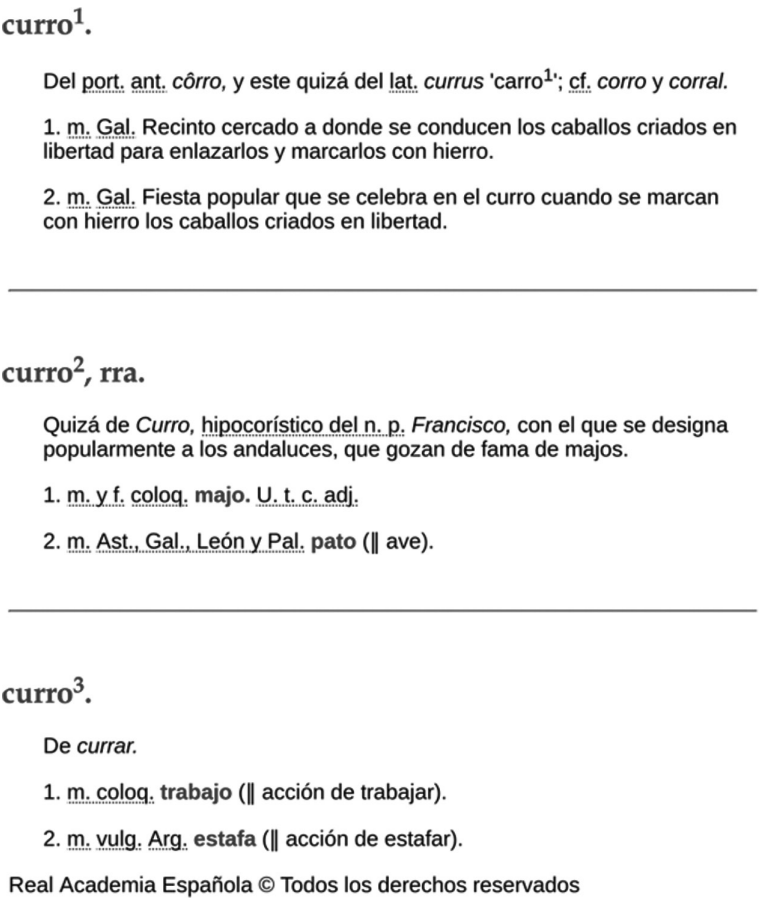
3.3.2 Palabras históricas, regionales, dialectales y coloquiales en el corpus y en el DLE

Pasamos a analizar las palabras del corpus que sí se encuentran en el *DLE*. ¿Cuáles y cuántas de las palabras están marcados como desusadas, y por tanto presentan una marca histórica? ¿Qué parte del léxico está marcado como proveniente de España o América, dialectal o coloquial?

Tabla 6: Palabras del corpus con formas en el *DLE* marcadas como desusadas, regionales, dialectales, coloquiales, americanismos, argentinismos o españolismos

Tipo de uso	Ejemplos	Lemas	%	*Tokens*	%
desusado	despabilar, marchante, melecina, tropelía, vaivenear	2023	6.50	228 148	22.76
dialectal	abastar, abonar, apegar, cunero, fierro	5	0.02	67	0.007
coloquial	asnal, baboso, curro, pajarraco, regordete	2477	7.96	220 732	22.02
España	achuchar, marrón, pirrar, talego, zagal	96	0.31	5041	0.50
América	botarate, bronca, caudillaje, gallero, milico	248	0.77	25 361	2.53
Argentina	endenantes, fifí, mate, palangana, viborear	739	2.35	38 930	3.88

La tabla 6 muestra los números absolutos y relativos de lemas y *tokens* clasificados según los diferentes tipos de uso. Podemos ver que más de una quinta parte de las palabras de las novelas en el corpus aparecen clasificadas como desusadas y otra quinta parte, como coloquial. El léxico de tipo dialectal es escaso. Los argentinismos alcanzan un mayor número que los americanismos en general y también son más numerosos que los españolismos, aunque los textos argentinos incluidos en el corpus tienen menos palabras que las novelas españolas (957 mil frente a 1.1 millones de *tokens*). La interpretación de estos números es problemática porque la categoría de uso no depende del lema sino de las acepciones diferentes. Por un lado, no podemos verificar el sentido en el que se utilizaron las palabras en los textos. Para este fin se necesitaría una anotación semántica detallada. Por otro lado, no nos es posible diferenciar entre la calificación de diferentes formas listadas en el *DLE* para un solo lema porque la estructura del diccionario en el *ebook* sigue el texto impreso y no es una estructura con marcado semántico (véase Imagen 8). Además, muchas palabras se pueden contar tanto como formas coloquiales o como regionales. Para poder interpretar estos resultados



curro[1].

Del port. ant. *côrro*, y este quizá del lat. *currus* 'carro[1]'; cf. *corro* y *corral*.

1. m. Gal. Recinto cercado a donde se conducen los caballos criados en libertad para enlazarlos y marcarlos con hierro.

2. m. Gal. Fiesta popular que se celebra en el curro cuando se marcan con hierro los caballos criados en libertad.

---

curro[2], rra.

Quizá de *Curro*, hipocorístico del n. p. *Francisco*, con el que se designa popularmente a los andaluces, que gozan de fama de majos.

1. m. y f. coloq. **majo.** U. t. c. adj.

2. m. Ast., Gal., León y Pal. **pato** (‖ ave).

---

curro[3].

De *currar*.

1. m. coloq. **trabajo** (‖ acción de trabajar).

2. m. vulg. Arg. **estafa** (‖ acción de estafar).

Imagen 8: Estructura de entradas homónimas en el *DLE*, con varias acepciones e indicaciones de tipos de uso[15]

mejor, este tipo de corpus de comparación sería útil, por ejemplo, para comparar la cantidad de formas desusadas en las novelas decimonónicas frente a las novelas modernas y, por tanto, apreciar la evolución histórica de esos términos.

15 Fuente: <http://dle.rae.es/srv/fetch?id=BkBBhTD|BkCutH5|BkCyiOI> [último acceso: 15/10/2017].

Para concluir, el *DLE* es una herramienta valiosa y útil no solo para consultas léxicas particulares sino también para análisis de tipo cuantitativo. Sin embargo, el valor del *DLE* para análisis léxicos digitales aumentaría notablemente con una versión accesible a técnicas de programación, por lo que sería bienvenida cualquier iniciativa en esa dirección. Para el corpus de trabajo utilizado aquí, se puede constatar que el *DLE* cubre la gran mayoría de los lemas (y *tokens*), aun cuando se trata de textos literarios del siglo XIX procedentes de diferentes países hispanohablantes. Por un lado, un análisis comparativo de textos literarios con un diccionario, como el que hemos realizado aquí, puede contribuir a entender mejor el léxico específico de los textos literarios. Por otro lado, una investigación de ese tipo también puede ayudar a descubrir candidatos para nuevas entradas y acepciones en el diccionario.

4 Conclusiones y perspectivas

Con *textbox* estamos dando los primeros pasos para la preparación de corpus históricos literarios en lenguas romances con métodos digitales de libre acceso. En concreto hemos centrado nuestra descripción y análisis en las colecciones de novelas españolas e hispanoamericanas. El formato TEI que utilizamos permite la reutilización de los corpus en numerosos contextos: para análisis cualitativos y cuantitativos, literarios y lingüísticos. Con la administración de datos mediante GitHub y Zenodo, nuestro proyecto ha encontrado también una solución viable para la publicación de esos textos.

Mostramos en este artículo que el análisis léxico es una de las muchas posibilidades de uso de *textbox*, con tres ejemplos concretos. Con pyzeta determinamos lemas característicos para las novelas provenientes de España frente a las de Hispanoamérica. La comparación de los resultados con *CORDE* muestra que los lemas identificados son relevantes para textos de esas variantes en general y no solo para los textos literarios aquí analizados. Utilizando las áreas léxicas «enfermedades» y «armas» del *NDHE*, constatamos que estas son significativamente más frecuentes en ciertos géneros novelescos (novela naturalista, gauchesca, histórica) que en otros. La comparación del léxico de novelas españolas y argentinas con el *DLE* mostró el gran alcance de esa obra de referencia y su utilidad para descubrir el léxico distintivo de autores individuales, pero también la necesidad de continuar desarrollando herramientas digitales relevantes para análisis léxicos y ofrecerlas de manera abierta.

Con todo, dentro y fuera de *textbox* todavía queda mucho por hacer. Para mejorar la representatividad de los resultados de análisis, se necesitan más textos en formato digital. También resulta imprescindible trabajar en la mejora de la

calidad de los textos que ya están disponibles. Hay que indicar las fuentes de los textos de manera fiable: ¿qué edición constituye la base del texto digital? ¿se respeta la ortografía original o está modernizada? ¿se han detectado errores en el texto y cuántos? Las herramientas para analizar las obras digitales representan otra área de trabajo. Son imprescindibles herramientas sensibles a las características de los textos históricos y también es imperativo asegurar un entrelazamiento razonable y eficaz de los diversos instrumentos.

La combinación de herramientas, ediciones críticas, corpus y recursos lingüísticos podría posibilitar y lanzar nuevas investigaciones digitales que por ahora resultan imposibles. Para ello se deben cumplir dos requisitos: la utilización de estándares y la publicación de los datos de manera abierta; es decir, permitir a otros investigadores un acceso completo a los datos, ya sea mediante descarga íntegra o mediante API. Esas dos características no solo representan maneras sólidas de abordar la investigación en la actualidad, sino que además estaremos poniendo las bases para que la investigación en español en entornos digitales pueda continuar desarrollándose durante las siguientes décadas.

Referencias bibliográficas

Agenjo, Xavier (2015): «Las bibliotecas virtuales españolas y el tratamiento textual de los recursos bibliográficos», *Ínsula: revista de letras y ciencias humanas* 822, 12–15.

Baquero Goyanes, Mariano (1986): *La novela naturalista española: Emilia Pardo Bazán*. Murcia: Universidad de Murcia.

Biblioteca Virtual Miguel de Cervantes (1999): Alicante: Universidad de Alicante. <www.cervantesvirtual.com> [último acceso: 15/10/2017].

Blei, David M. (2012): «Probabilistic Topic Models», *Communications of the ACM* 55, 4, 77.

Burnard, Lou/C. M. Sperberg-McQueen (2012): «TEI Lite: Encoding for Interchange: an introduction to the TEI. Final revised edition for TEI P5». <www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_lite.doc.html> [último acceso: 15/10/2017].

Burrows, John (2007): «All the Way Through: Testing for Authorship in Different Frequency Strata», *Literary and Linguistic Computing* 22, 1, 27–47.

CLiGS textbox <http://github.com/cligs/reference/> [último acceso: 15/10/2017].

Craig, Hugh/Arthur F. Kinney (eds.) (2009): *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

DLE = Real Academia Española/Asociación de Academias de la Lengua Española (2014[23]): *Diccionario de la lengua española*. Barcelona: Espasa Libros. <http://www.rae.es> [último acceso: 20/09/2017].

DRAG = Real Academia Galega (2012): *Diccionario da Real Academia Galega*. <https://academia.gal/dicionario/> [último acceso: 15/10/2017].

Eder, Maciej/Mike Kestemont/Jan Rybicki (2016): «Stylometry with R: A Package for Computational Text Analysis», *The R Journal* 16,1, 1–15.

ePubLibre (2013): ePubLibre. <www.epublibre.org> [último acceso: 15/10/2017].

Ertler, Klaus-Dieter (2013): *Moralische Wochenschriften*. Graz: Universität Graz. <www.gams.uni-graz.at/archive/objects/container:mws-es/methods/sdef:Context/get?locale=de&mode=&context=es> [último acceso: 15/10/2017].

Fellbaum, Christiane (ed.) (1998): *WordNet: an electronic lexical database*. Cambridge: MIT Press.

Fièvre, Paul (ed.) (2007–2017): «Théâtre classique». París: Université Paris-IV Sorbonne. <www.theatre-classique.fr> [último acceso: 15/10/2017].

Fradejas Rueda, José Manuel (2017): *7 Partidas Digital*. Valladolid: Universidad de Valladolid. <github.com/7PartidasDigital> [último acceso: 15/10/2017].

Gasparini, Sandra (2008): «De oradores, polémicas y distopías. La emergencia de la fantasía científica», *Anales Nueva Época* 11, 143–160.

GitHub <https://guides.github.com/introduction/git-handbook/> [último acceso: 15/10/2017].

Henny, Ulrike/Christof Schöch (2016): *How Good Are Our Texts, Really? Quality Assurance for Literary Texts from Various Sources*. Wurzburgo: *CLiGS*. <cligs.hypotheses.org/371> [último acceso: 15/10/2017].

Henny-Krahmer, Ulrike/Frederike Neuber (2017): «Criteria for Reviewing Digital Text Collections, version 1.0», *RIDE: A review journal for digital editions and resources*, 6. <www.i-d-e.de/publikationen/weitereschriften/criteria-text-collections-version-1-0/> [último acceso: 15/10/2017].

Hoover, David L. (2010): «Teasing out Authorship and Style with T-Tests and Zeta», *Digital Humanities Conference*, Londres. <http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-658.html> [último acceso: 15/10/2017].

Jannidis, Fotis/Hubertus Kohle/Malte Rehbein (eds.) (2017): *Digital Humanities: eine Einführung*. Stuttgart: Metzler Verlag.

Levine, Alex/Adriana Novoa (2012): *Darwinistas! The Construction of Evolutionary Thought in Nineteenth Century Argentina*. Leiden: Brill.

Miller, George A. (1995): «WordNet: A Lexical Database for English», *Communications of the ACM* 38, 11, 39–41.

Navarro-Colorado, Borja/María Ribes Lafoz/Noelia Sánchez (2015): *Corpus of Spanish Golden-Age Sonnets*. Alicante: University of Alicante. <github.com/bncolorado/CorpusSonetosSigloDeOro> [último acceso: 15/10/2017].

NDHE = Instituto de Investigación Rafael Lapesa de la Real Academia Española (2013–): *Nuevo diccionario histórico de la lengua española*. <http://web.frl.es/DH> [último acceso: 15/10/2017].

Nielsen, Lars Holm (2013): «ZENODO – An innovative service for sharing all research outputs». <https://doi.org/10.5281/zenodo.6815> [último acceso: 15/10/2017].

Padró, Lluís/Evgeny Stanislovsky (2012): «FreeLing 3.0: Towards Wider Multilinguality», *Proceedings of the Language Resources and Evaluation Conference*, Estambul. <nlp.lsi.upc.edu/publications/papers/padro12.pdf> [último acceso: 15/10/2017].

Pascual, José Antonio/Rafael García (2008): «Estado del *Nuevo diccionario histórico* de la Real Academia Española», en Pilar Garcés (ed.), *Diccionario histórico: nuevas perspectivas lingüísticas*. Madrid: Iberoamericana, 11–15.

Pascual, José Antonio/Mar Campos Souto (2014): «La morfología léxica en el *NDHE*», en Bruno Camus Bergareche (ed. lit.), *Morfología y diccionarios* (Anexos de la *Revista de Lexicografía*, 31). La Coruña: Universidade da Coruña, 123–148.

Rojas Castro, Antonio (2016): *Soledades de Luis de Góngora*. Barcelona: Universidad Pompeu Fabra. <www.soledadesediciondigital.com> [último acceso: 15/10/2017].

Ruiz, Pablo/Clara Martínez Cantón/José Calvo Tello (2017): *DISCO: Diachronic Spanish Sonnet Corpus*. Madrid: UNED. <github.com/pruizf/disco> [último acceso: 15/10/2017].

Schöch, Christof (2017): «Genre Analysis», en James O'Sullivan, *Digital Humanities for Literary Studies*, Pennsylvania: Pennsylvania State Univ. Press.

Schöch, Christof/Albin Zehe/José Calvo Tello (2018): «Burrows Zeta: Varianten und Evaluation», DHd 2018 Kritik der digitalen Vernunft. <dhd2018.uni-koeln.de/programm/> [último acceso: 15/10/2017].

Scott, Mike (1997): «PC Analysis of Key Words and Key Key Words», *System* 25, 2, 233–245.

TextGrid: Virtuelle Forschungsumgebung für die Geisteswissenschaften (2006): Gotinga: TextGrid Konsortium, <textgrid.de>. DOI: https://doi.org/10.5281/zenodo.597430> [último acceso: 15/10/2017].

II
Crítica de los recursos en línea: el desorden digital

Alejandro Fajardo

Lexicografía histórica con corpus y recursos digitales: aspectos metodológicos[1]

Resumen: Los recursos digitales han transformado la metodología de la investigación diacrónica del léxico, que está adaptándose a los cambios. La digitalización de sus fuentes aporta numerosas ventajas, pero también exige una reorganización del trabajo. Localizar los recursos que van surgiendo y aprender a utilizarlos es parte de las nuevas tareas lexicográficas, pero la innovación constante y la rápida obsolescencia causan dificultades. Por tanto, es necesaria una crítica permanente de las bases teóricas, de la configuración y de las funciones de los recursos. El objetivo de este estudio es analizar los de mayor utilidad y señalar las necesidades más inmediatas para la lexicografía histórica del español en este nuevo contexto.

Palabras clave: Lexicografía histórica, Lexicografía digital, Corpus lingüísticos, Humanidades digitales, Lingüística histórica

Abstract: Digital resources have changed the methodology of the lexicon's diachronic research, which is in the process of adaptation to this new situation. The digitalization of its sources brings many advantages, but it also requires work reorganization. Finding emerging resources and learning to use them are part of the new lexicographical tasks. However, constant innovation and rapid obsolescence cause difficulties. Therefore, it is necessary a permanent criticism of the theoretical bases as well as of the configuration and functions of the resources. The aim of this study is to analyse the most useful resources and to give an account of the most urgent needs for the historical lexicography of Spanish in this new context.

Keywords: Historical lexicography, Digital lexicography, Linguistic corpora, Digital humanities, Historical linguistics

1 Introducción

Desde los comienzos de la digitalización de textos, hace ya varias décadas, el incremento de información disponible ha aportado tal cantidad de datos para la historia del léxico que ha sido necesario replantear la metodología tradicional de investigación. La lexicografía histórica se ha visto especialmente afectada por

1 Este trabajo se enmarca en el proyecto de investigación FFI2016-76154-P, financiado por el Ministerio de Economía y Competitividad del Gobierno de España.

los cambios, ya que el acopio y procesamiento de los testimonios necesitó, en los proyectos de mayor envergadura, muchos años de trabajo antes de comenzar a publicar resultados.

En este estudio, se harán unas consideraciones críticas sobre la manera en que las humanidades digitales han afectado a la lexicografía diacrónica, entendiéndola como análisis y descripción histórica del léxico, pero no necesariamente orientada a producir diccionarios en sentido tradicional, es decir, sin una orientación *diccionarística*, pues el concepto de diccionario digital no se ha materializado aún en un producto de características tan claramente definibles como eran los diccionarios de papel. Por otro lado, la metalexicografía se ve especialmente afectada por las transformaciones que conlleva la nueva estructura informativa y el acceso a ella: aspectos de la reflexión teórica sobre la marcación, los límites entre definición lexicográfica y enciclopédica, el formato de los ejemplos, etc., deben ser vistos con una nueva perspectiva. En la actualidad, la constricción de espacio a que obligaba el papel ha sido sustituida por la presión del tiempo, que afecta tanto a quienes elaboran los diccionarios como a los usuarios que los consultan; en consecuencia, esta dimensión debe ser considerada como un aspecto decisivo al evaluar la calidad de los productos lexicográficos digitales.

A pesar de que los cambios se han generalizado, seguimos estando en un periodo de transición, por lo que es necesaria la crítica de los recursos disponibles para orientar a la informática hacia la solución de necesidades filológicas concretas. Los nuevos productos digitales no son una panacea para la lexicografía histórica; los corpus, p. ej., no se crearon pensando en ella y, aunque son indudablemente útiles, necesitan ajustes y complementos para adaptarse a las necesidades de la investigación. Por otra parte, el número de los recursos lingüísticos disponibles en español crece continuamente, lo que es positivo; sin embargo, su dispersión y la falta de estándares ralentizan y dificultan la obtención de resultados.

En las siguientes secciones, analizaremos las características, ventajas e inconvenientes para la lexicografía histórica de los corpus, archivos y ficheros, bases de datos y otros recursos como bibliotecas y hemerotecas digitales.

2 Los corpus frente a los archivos lexicográficos

2.1 La tradición lexicográfica del archivo

En el estado actual de la lexicografía digital se puede plantear la pregunta de si los corpus sustituyen a los archivos lexicográficos. El *archivo lexicográfico* (también llamado simplemente *fichero*, como los cajones o gavetas que tradicionalmente lo componían) procede de una tradición, de eficacia consolidada, orientada a

elaborar diccionarios; no debe confundirse, por tanto, con lo que en lingüística de corpus se denomina *archivo*, tal como lo define Rojo (2016: 1): «un conjunto heterogéneo de textos integrados en un recurso único como consecuencia de factores que no se vinculan al deseo de lograr una determinada composición general», donde el archivo no está concebido para una finalidad como redactar un diccionario.

La utilización que se haga de los ficheros lexicográficos y su relación con los nuevos recursos depende de las características del proyecto lexicográfico, que puede tener distintos puntos de partida:

a) El diccionario se basa en un archivo heredado de un proyecto precedente, reunido durante largo tiempo, cuya aceptación se hace irrenunciable por la calidad de sus materiales, pero también por motivos de tiempo y financiación; el archivo debe ser actualizado y se suelen ver los corpus como la mejor manera de completarlo o, incluso, de superarlo.
b) Para redactar el diccionario diacrónico se crea un corpus textual *ad hoc* completamente nuevo, esto implica la ruptura con la tradición y la renuncia a materiales legados.
c) Sobre la base de un fichero y otros materiales heredados se sobrepone un corpus creado *ad hoc,* es la opción de compromiso seguida por el *Nuevo diccionario histórico del español* (*NDHE*), que, sin renunciar al *Fichero general* de la RAE, ha creado un subcorpus específico para el diccionario histórico: el *Corpus del diccionario histórico-Nuclear* (*CDH-N*). El nuevo corpus ha sido ensamblado en una misma interfaz con otros corpus anteriores como son el *Corpus diacrónico del español* (*CORDE*) y el *Corpus de referencia del español actual* (*CREA*), dando lugar al *Corpus del nuevo diccionario histórico del español* (*CNDHE*), que tiene, por tanto, un carácter mixto.

Los proyectos lexicográficos de larga duración en español, como los desarrollados históricamente por la RAE en sus intentos de elaborar un diccionario histórico y, en Alemania, el *Diccionario del español medieval* (*DEM*), truncado en 2005 (Fajardo 2012), son *proyectos-legado* cuya dotación consiste, sobre todo, en enormes ficheros, a menudo acompañados de otros subficheros secundarios. En el caso de la RAE, su archivo, el *Fichero general*, con unos diez millones de papeletas, ha sido base del conjunto de las obras académicas, pero la incorporación de numerosas papeletas no pudo ser ajena al afán de la Academia por contar con un diccionario histórico (Campos 2017). En el caso del *DEM* de Müller (1987–2005), los 480 ficheros de su archivo contienen unas 850 000 papeletas (Arnold/Langenbacher 2018). Para constituir un archivo *ad hoc* como este se precisa una cuidadosa selección de textos relacionados con el objetivo que se persigue, que es describir la

historia del léxico de un periodo. En este aspecto, el archivo lexicográfico es homogéneo y su finalidad es específica, a diferencia de los corpus, que habitualmente se reúnen con fines mucho más generales (culturales, literarios, lingüísticos, etc.).

2.2 El fichero lexicográfico digital

El aprovechamiento de esos ricos materiales exige que sean digitalizados y puestos a disposición de los investigadores, proceso emprendido en los últimos años. Una muestra de esta nueva vida de los ficheros, como el legado por Bodo Müller, es el proyecto del *Diccionario del Español Medieval electrónico* (*DEMel*). La evolución digital revalorizará, sin duda, el concepto de archivo lexicográfico, menospreciado por la eclosión de los corpus digitales, porque no hay que olvidar que el punto de partida del trabajo del lexicógrafo histórico son las palabras y sus familias, por eso la papeleta o ficha (organizada en ficheros, con diversas agrupaciones más allá de lo alfabético, con sus remisiones internas y lematizaciones) ha sido la base de un método de trabajo cómodo y efectivo empleado durante tres siglos. En la ficha se pueden reunir distintas informaciones necesarias para estudiar la palabra:

— testimonio de aparición extraído de la base documental,
— fecha del documento base,
— referencia bibliográfica,
— registro en vocabularios, glosarios y diccionarios.

Pero también es posible incluir informaciones secundarias, producto de la elaboración filológica:

— lema,
— rasgos de significado o esbozos de definición,
— comentarios sobre la calidad de la edición,
— indicaciones sobre dataciones problemáticas,
— observaciones sobre la fecha de la edición y la del manuscrito,
— variantes gráficas del lema,
— remisiones a otros lemas, etc.

Como se comprueba, el fichero concentra diversas informaciones de interés que no se encuentran en los corpus porque, a diferencia de estos, contiene datos ya seleccionados y adaptados para elaborar un diccionario. Los corpus, en cambio, están diseñados con expectativas muy generales y sería un error considerarlos el resultado de la evolución del fichero. Un buen etiquetado en aspectos diacrónicos, diastráticos, diafásicos, dialectales, etc., puede mejorar la utilidad del

corpus, pero el lexicógrafo necesita, además, un tipo de información más compleja y detallada.

La nueva estructura digital de los datos léxicos debe permitir hacer visibles las relaciones entre ellos y facilitar el acceso simple y rápido a los textos que los atestiguan, pero también a bibliotecas y hemerotecas digitales, diccionarios y tesoros lexicográficos en línea. Es necesario, por tanto, replantear el concepto de *fichero lexicográfico,* que no debe quedarse en el retrato digital de las papeletas ni en un corpus adaptado para lexicógrafos, sino que el nuevo *fichero lexicográfico digital* debe basarse en la conexión simple y rápida de la *ficha digital* con todas las fuentes documentales y compartir la interfaz con el escritorio de redacción.

Sin embargo, en su estado presente, hay que tener en cuenta algunas limitaciones de los ficheros heredados, entre las principales:

a) En el caso de que los ficheros-legado se reutilicen para nuevos diccionarios, habrá discordancias entre el canon de textos papeletizados en ellos y los nuevos corpus o bases de datos que se implementen para la reelaboración, lo que producirá duplicaciones. El fichero, digitalizado o no, no es sustituible y es necesario el cotejo permanente con el nuevo recurso.
b) Los ficheros contienen otro tipo de informaciones secundarias (anotaciones del investigador, predefiniciones, remisiones, comentarios de diversa índole, etc.), que tampoco son sustituibles.
c) Las fichas no son «transparentes», en el sentido de que no están conectadas con las fuentes que en su momento fueron papeletizadas, dependen de una bibliografía que puede no ser ya accesible si no se ha conservado unida al archivo.
d) No se distingue entre la información útil y la obsoleta; p. ej., es frecuente que estén papeletizados textos con ediciones que han sido muy mejoradas por otras posteriores o que se den dataciones superadas ya por otras más tempranas.

En definitiva, puesto que será necesario seguir trabajando con esta herencia como complemento de nuevas fuentes y recursos, las fichas digitalizadas deberían incorporar informaciones y herramientas informáticas para paliar desajustes como los mencionados, y no limitarse a reproducir en otro soporte el fichero de papel. La base de datos con acceso directo a corpus y otros recursos de consulta y edición se perfila como la herramienta más útil para el trabajo lexicográfico, y la sucesora a la vez del fichero con papeletas y del diccionario de papel, mientras que el corpus sigue siendo un recurso imprescindible, pero auxiliar.

3 Características generales de los corpus diacrónicos

Desde que en 1978 el *Hispanic Seminary of Medieval Studies* comenzó a publicar en microfichas las concordancias de textos medievales (Kasten/Nitti/Anderson 1978), se fueron multiplicando los documentos que la lexicografía histórica debía procesar. Cuatro décadas después, las concordancias se presentan en corpus digitales que aportan múltiples informaciones añadidas a cada una de ellas (frecuencias, extensión, comparativas, etc.).

En la actualidad, contamos con unos pocos corpus generales de carácter diacrónico, que pretenden abarcar un amplio panorama de la historia de la lengua, y un número creciente de corpus especializados elaborados para desarrollar investigaciones sobre periodos o tipos de textos determinados. Los corpus diacrónicos se han convertido, así, en el principal recurso para la lexicografía, bien como base para elaborar un diccionario —el *CNDHE*—, bien para la consulta en investigaciones concretas. Sin embargo, ante la proliferación de este tipo de recursos, hay que tener en cuenta su tipología y características para aumentar la probabilidad de que su consulta aporte resultados, por lo que se analizará a continuación una selección de los de más amplio alcance.

3.1 Corpus generales

Los corpus generales han tenido en los últimos años un incremento grande en cantidad y calidad en lo que se refiere a las incorporaciones de textos de los periodos más recientes, pero el crecimiento es mucho más lento en los corpus diacrónicos —en el trabajo de Rojo (2016) puede comprobarse el abrumador desarrollo de los corpus dedicados al español de nuestros días—. El nuevo *Corpus del Español* (*CE*) tiene un tamaño de dos mil millones de palabras (Davies 2017), lo que supone que se ha multiplicado por 100 respecto a la parte del español del s. XX que había en el *Corpus del Español* original de 2002. Este gran crecimiento ha sido posible gracias a la captación masiva de textos de todo tipo de páginas de Internet, lo que significa que el aumento se ha producido casi exclusivamente en el español más reciente. Por otra parte, los materiales del *CREA* de la RAE han tenido una prolongación hasta la actualidad gracias al *Corpus del Español del Siglo XXI* (*CORPES XXI*). También en la Academia, el aumento de la parte diacrónica de los corpus ha sido más modesto, aunque ha habido avances en la formación del *CDH-N*, que ha incorporado nuevos textos (con unos 62 millones de palabras), a los que se han sumado los materiales procedentes del *CORDE* y el *CREA*, se han acoplado así tres «capas» (Campos Souto/Pascual 2012) que conforman el *CNDHE*.

3.2 Corpus especializados

La mayoría de los corpus marcan sus propios límites de especialización (cronológica, diatópica, textual, etc.), aunque algunos de ellos adquieren, por su riqueza, utilidad para estudios de tipo muy diverso. El *Corpus Diacrónico y Diatópico del Español de América* (*CORDIAM*) tiene un enfoque muy amplio, aunque no puede considerarse un corpus general mientras España no esté representada en él. Su interés para la historia del español es indudable debido, entre otras cualidades, a la amplia y bien estructurada tipología textual en que se basa, ya que ha añadido a los documentos archivísticos que le sirvieron de base dos nuevos subcorpus. En la medida en que siga aumentando en número de textos, su utilidad será creciente. En cuanto al *Corpus Hispánico y Americano en la Red: Textos Antiguos* (*CHARTA*), el otro gran corpus especializado, tiene una extensión geográfica panhispánica, pero sus límites, en este caso, vienen dados por la tipología documental, de naturaleza únicamente archivística. Ambos proporcionan una importante cantidad de documentos, que ya superan los cuatro millares en el *CORDIAM* y los dos millares en *CHARTA*, por lo que su consulta es irrenunciable para documentaciones de lexicografía histórica.

De menor amplitud son otros corpus vinculados con *CHARTA*, accesibles en forma de subcorpus integrados o, también, de manera autónoma como el *Corpus de documentos españoles anteriores a 1800* (*CODEA+*), el *Corpus diacrónico del español del Reino de Granada* (*CORDEREGRA*), el *Corpus Diacrónico de Documentación Malagueña* (*CODEMA*), etc. Por otra parte, el *Corpus Léxico de Inventarios* (*CorLexIn*) aporta léxico de interés de la vida cotidiana en el Siglo de Oro. Aún más específico es el corpus *Biblia medieval*, dedicado monográficamente a un tipo de textos: las traducciones bíblicas medievales; tiene el interés añadido de presentar un corpus paralelo, que permite cotejar el texto equivalente en las diferentes traducciones. En cuanto al proyecto *Post Scriptum*, está especializado en las cartas privadas escritas durante la Edad Moderna; es muy útil como fuente testimonial del uso cotidiano de la lengua y ofrece, además, dos formatos: una edición crítica digital y un corpus lingüísticamente anotado.

Dentro de los límites que se marcan, todos los corpus citados tienen una base documental amplia y cuidada, y comparten posibilidades de búsquedas simples o de coocurrencias, frecuencias, etc., pero presentan también grandes diferencias en cuanto a su estructura interna y a sus posibilidades de ser personalizados para fines concretos. Por este motivo, no conviene emprender el trabajo con ellos sin un análisis crítico previo de lo que cabe esperar de cada uno, como ha señalado Kabatek (2016: 7). En trabajos lexicográficos, en concreto, se deben valorar

las posibilidades de extraer y analizar los datos, los tipos de acceso, la simplicidad de la consulta y las posibilidades de exportar los resultados.

4 Análisis de funciones y configuración de los corpus

4.1 Tamaño total y tamaño por lapsos

El primer parámetro que hay que tener en cuenta es el tamaño del corpus, que está en relación directa con la mayor posibilidad de obtener de él datos significativos. En este aspecto, el *CNDHE* arroja resultados más cuantiosos, ya que suma a las aportaciones del *CORDE* y *CREA* el subcorpus *CDH-N* o «tercera capa». No obstante, dependiendo del corte diacrónico que interese, hay que relativizar los valores cuantitativos. Dado que la mayor parte de la «extensión diacrónica» del *CNDHE* está constituida por el *CORDE*, habrá que tener en cuenta la representatividad que cada época tiene en el conjunto. La cantidad de elementos en este corpus va aumentando a medida que se avanza en el tiempo, pero el «peso» que se otorga a toda la Edad Media, desde los orígenes hasta 1492, es relativamente pequeño, pues solo representa un 21 %, que aumenta al 28 % para los Siglos de Oro (1493–1713) y al 51 % para la Época Contemporánea desde 1724 hasta 1974.

Tras implementarse el *CDH-N*, la relación entre la mayor antigüedad y el menor número de datos no cambia demasiado, puesto que la distribución de la cantidad de registros por periodos temporales representa el 24 % en el corte de 1064–1500, el 19 % en 1501–1700, el 9 % en 1701–1800, el 11 % en 1801–1900 y el 37 % en 1901–2005.

En el caso del otro corpus general, el *CE*, cuenta con 18 millones de palabras del s. XIII al XV, 42 millones del s. XVI al XVIII y alrededor de 40 millones de los ss. XIX y XX; también aquí, y en mayor medida, se constata la escasez relativa de documentación de los primeros siglos.

En todo caso, la posibilidad de seleccionar el corte de estudio por años es una de las funciones básicas de los corpus, por lo que no hay problema para elegir un lapso determinado. Otra cuestión distinta es el análisis de las ocurrencias obtenidas de un lapso parcial en relación con el conjunto del corpus, porque en el caso de análisis cuantitativos no tiene el mismo valor comparar la frecuencia de las apariciones de una manera absoluta frente a todo el corpus o hacerlo relativizando la frecuencia en función de los cortes que sean significativos para nuestra investigación.

En cuanto a la relación entre el lapso parcial y el número de formas y documentos en los corpus específicos, *CHARTA* aporta detalladas estadísticas tanto

Tabla 1: Características generales de los corpus (dimensión, cronología, transparencia documental, interacción con el usuario)

	Tamaño total	Lapso total	Tamaño por lapsos parciales	Facsímil	Registro	Pago	*Feed-back*
CORDE	250 millones de palabras	• Origen-1974	• 21 % Orígenes-1492 • 28 % Siglos de Oro (1493–1713) • 51 % Época Contemporánea (1714–1974)	No	No	No	No
CNDHE	409 (62.2 del *CDH-N*: 38 de España y 24 de América; 22.04 del *CORDE*; 123.79 del *CREA*)	• Origen-2000	• 24 % 1064–1500 • 19 % 1501–1700 • 9 % 1701–1800 • 11 % en 1801–1900 • 37 % 1901–2005	No	No	No	Correo
CE	100 millones de palabras (18 XIII–XV, 42 XVI–XVIII, 40 XIX–XX)	• XIII–XX	• 18 % ss. XIII–XV • 42 ss. XVI–XVIII, • 40 % ss. XIX–XX	No	Opcional tras x consultas	Para consultas avanzadas	Correo
CORDIAM	4.5 millones de palabras	• 1494–1905	(en formación)	Sí	No	No	Formulario, correo
CHARTA	2076 documentos	• XII–XIX	• VIII: 0.77 % • IX: 0.05 % • X: 0.19 % • XI: 0.87 % • XII: 3.37 % • XIII: 53.71 % • XIV: 12.72 % • XV: 6.50 % • XVI: 9.34 % • XVII: 6.17 % • XVIII: 4.53 % • XIX: 1.78 %	Sí	No	No	Formulario

de frecuencias relativas como de número de documentos por siglos e incluso por años. En este caso se comprueba la preponderancia de la documentación del s. XIII, que representa un 53 % del total.

4.2 Taxonomía interna

Dado que cada corpus establece distinta taxonomía a partir de los textos que lo conforman, es posible distinguir una serie de niveles. En un primer nivel, se establecen agrupaciones mayores que abarcan un gran número de textos. El *CORDE* parte en su concepción original de una división interna entre «1. Ficción 2. No ficción», subdivididas a su vez en distintas modalidades o géneros en verso y prosa; desde el punto de vista del usuario, sin embargo, la búsqueda debe hacerse (además de las habituales por autor o año) por cuatro criterios: cronológico, medio, geográfico o tema.

En el caso del *CNDHE*, en un primer nivel la elección responde a los tres posibles orígenes de los materiales que reúnen en él: «1. Corpus del diccionario histórico-Nuclear (s. XII-2005), 2. s. XII-1975 [absorción del *CORDE*], 3. 1975–2000 [absorción del *CREA*]».

El *CE género/histórico* ofrece en un primer nivel las modalidades de resultados obtenibles, con varias formas de presentación: «1. Lista, 2. Gráfico, 3. Colocados, 4. Comparar, 5. PCEC (palabra clave en contexto)».

El *CORDIAM* parte de cuatro agrupaciones: «1. variación diatópica, 2. variación diacrónica, 3. variación autoral, 4. variación textual». La búsqueda, desde la perspectiva del usuario, ofrece múltiples opciones, pues es posible realizarla por metadatos: «1. Siglo, 2. País actual, 3. Tipo textual, 4. Autor, 5. Datos étnicos, 6. Periódico, 7. Fecha».

CHARTA ofrece distintos puntos de acceso a partir de la pantalla de presentación. En la modalidad «estadística», los datos se pueden obtener con las siguientes agrupaciones: «1. Cronológica, 2. Geográfica, 3. Tipología y archivo, 4. Grupos y subcorpus».

Los ejes de variación que parten de este primer nivel hacen posibles combinaciones muy diversas que, en cada corpus, obligan a tomar distintos caminos para llegar a resultados similares. Como vemos, estamos muy lejos de un estándar establecido, de manera que el lexicógrafo debe dedicar tiempo a familiarizarse con las distintas ramificaciones de cada corpus.

4.3 Género

En el diseño de los corpus, la organización del contenido en géneros textuales tiene una gran trascendencia (Bertolotti/Company 2014). Con frecuencia encontramos en una misma palabra diferencias semánticas relacionadas con su uso en textos de distinto tipo, puesto que dependiendo del género en que aparezca una forma pueden darse, p. ej., valores más metafóricos, más marcados estilísticamente, más restringidos a un nivel o registro, etc.

Tabla 2: Subcorpus por géneros

	Género (tipología textual)
CORDE	• Ficción/No Ficción • Verso: lírico/épico/dramático • Prosa: lírica/narrativa/dramática/didáctica/científica/de sociedad/periodística/publicitaria/religiosa/histórica/jurídica
CNDHE	• Artes (15) • Ciencias sociales; creencias, pensamiento (18) • Ciencia y tecnología (20) • Ficción (4) • Ocio, vida cotidiana (12) • Política, economía, comercio, finanzas (18) • Salud (8) • Tema desconocido
CE	• Hablado • Ficción • Prensa • Académico
CORDIAM	• Documentos: administrativos, cronísticos, entre particulares, cartas y otros, jurídicos • Literatura: narrativos, poéticos, prosa varia, teatro, textos cronísticos • Prensa: comentativos, informativos, publicitarios, anuncios varios
CHARTA	• Actas y declaraciones • Cartas de compraventa y contratos • Cartas privadas • Certificaciones • Estatutos • Informes y relaciones • Notas y breves • Otros • Recibí • S. tip. • Testamentos e inventarios • Textos legislativos

Cuanto mayor sea la posibilidad de hacer análisis de ocurrencias por géneros, subgéneros o combinaciones entre ellos, más útil será el corpus, porque se podrá enfocar mejor el objeto de la búsqueda. La restricción genérica puede funcionar, así, como un primer filtro semántico. Por otra parte, las coapariciones de palabras en uno u otro género resultan significativas para distintos tipos de estudios, no solo lingüísticos, sino también culturales. En textos de carácter científico,

p. ej., no se encontrarán las múltiples combinaciones fraseológicas en las que intervienen nombres de animales y que se han fijado en la lengua corriente; en transcripciones de textos orales, en cambio, serán más frecuentes palabras de un registro informal.

La comparación de apariciones de palabras o combinaciones en géneros diferentes es de interés desde el punto de vista semántico, pero, si se hace para obtener datos cuantitativos, el número de ocurrencias debe tomarse con mucha cautela, pues solo en géneros de similar tamaño serían válidas. Las comparaciones sí son fiables cuando el corpus proporciona para cada género no solo frecuencias absolutas, sino frecuencias normalizadas (en casos por millón). Esta función está disponible en el *CNDHE* y en el caso del *CE* solo para el s. XX, único periodo consultable por géneros. También *CHARTA* y *CORDIAM* cuentan con ella. Hay que señalar que, aunque el *CORDE* está integrado como «segunda capa» en el *CNDHE*, el acceso a través de géneros a los datos contenidos en *CORDE* es distinto si se hace directamente en él o si se hace una vez integrado en *CNDHE*, puesto que la clasificación en géneros no es la misma en uno y otro caso.

4.4 Frecuencia

Una de las funciones más características de todos los corpus es la posibilidad de proporcionar la frecuencia con que aparecen las palabras o combinaciones. Desde el punto de vista histórico, el análisis de la frecuencia normalizada de la forma (ocurrencias/millón de palabras) resulta de especial interés para el seguimiento de la evolución léxica, pero solo en la medida en que se disponga de fragmentos del corpus en los que poder contrastar las frecuencias normalizadas será posible obtener conclusiones de interés. Estos fragmentos pueden obedecer a distintos criterios taxonómicos y recibir diferentes denominaciones, pero las opciones que tiene el usuario para comparar las frecuencias normalizadas en los fragmentos de su elección son bastante limitadas en los corpus actuales.

El *CE* ofrece funciones interesantes que lo hacen muy flexible y avanzado para obtener frecuencias (aunque la comparación de la frecuencia por géneros está limitada al s. XX). Una opción muy útil, exclusiva de este corpus, es la posibilidad de hacer búsquedas partiendo no de la forma, sino de la frecuencia, lo que permite encontrar palabras que reúnan unas condiciones de frecuencia determinadas. Con esta función, se pueden combinar criterios diacrónicos para encontrar las *ocurrencias/lapso*, es decir, las formas que en un determinado periodo tengan cierto número de ocurrencias o, incluso, las que no aparecen nunca en un periodo. También es posible la búsqueda con criterios diatópicos a partir de frecuencias, lo que proporciona listas de palabras que cumplen determinado

rango de frecuencia en un país. Estas comparaciones, por otra parte, pueden hacerse bien de una forma concreta, bien de todas las formas en conjunto de ese periodo. Las búsquedas *frecuencia/forma* hacen posible la obtención de resultados seleccionando elementos diferentes (palabra, segmento, prefijo, colocación, etc.). La frecuencia por coocurrencia entre formas admite dos modalidades de comparaciones:

a) entre dos formas: *a…b,*
b) entre una forma *a* y un listado de todas las que concurran con ella ordenadas por frecuencia: 1º *a…b*, 2º *a…c*, 3º *a…d*, etc.

Así, a partir de criterios de *frecuencia-lapso/diatopía*, se puede obtener datos de mucho interés para el estudio de la evolución de combinaciones de cualquier tipo de elementos (morfología, colocaciones, fraseología, etc.).

El *CORDE* ha quedado obsoleto en varios aspectos, especialmente en su interfaz y en las limitaciones de su motor de búsqueda; la más molesta es el límite para obtener ocurrencias de formas de uso frecuente, pues a partir de los mil casos se nos informa de que «No se puede recuperar. Demasiados documentos». Afortunadamente, su absorción por el *CNDHE* ha solucionado este aspecto problemático, de manera que es posible realizar búsquedas sin límite de apariciones y aplicarles los filtros deseados de capa, género, subcorpus, etc.

4.5 Comparaciones

La posibilidad de realizar distintos tipos de comparaciones entre elementos, además de las de frecuencia, es otra de las funciones más útiles para aprovechar el contenido de los corpus. El acceso básico habitual permite encontrar las coapariciones entre formas (palabras, partes, etc.), estableciendo un límite de distancia entre la separación de una y otra. Sin embargo, una función que se encuentra en el *CE* permite un nivel de contraste mayor, pues hace posible comparar entre una forma *a* y listas de formas que crea el usuario; esto multiplica los resultados, que en lugar de limitarse al tipo *forma-forma* se organizan por frecuencia de *forma-lista* (1º *a…b*, 2º *a…c*, 3º *a…d*). En un tercer nivel, el *CE* posibilita incluso la comparación entre una lista de formas y todo el conjunto del corpus o una parte de él (el usuario puede crear un repertorio personal para cotejar, según sus intereses, un campo léxico predeterminado, una familia de palabras, una lista de sinónimos, un glosario, etc.).

Las posibles comparaciones dependen de los fragmentos de corpus con que se pueda operar (género, siglo, país…); lo ideal, sería poder hacerlas siempre operando con [*forma*/*lista*] en [*lapso*/*género*/*país*] en relación con [*forma*/*lista*]

en [*lapso*/*género*/*país*]. Si bien este tipo de comparaciones no se puede realizar plenamente en el conjunto del *CE* (ya que la fragmentación en géneros no es posible más que en uno de los cortes diacrónicos del corpus, como señalamos en 4.3), el concepto puede servir como modelo para el diseño de otros corpus, pues multiplica sus posibilidades.

4.6 Aportación de facsímiles y otras presentaciones

Un servicio de gran interés para el trabajo lexicográfico, y filológico en general, es el ofrecimiento al usuario de facsímiles de los documentos originales en que se basa el corpus (manuscritos, ediciones empleadas, etc.). La necesidad de consultar la fuente original ante cualquier duda que pueda surgir es habitual en la redacción lexicográfica, pero en los corpus generales, debido a su tamaño, no se aporta nunca, lamentablemente, este tipo de documentos.

En este aspecto, los corpus específicos son mucho más ricos. *CHARTA* aporta de forma sistemática estas valiosas informaciones por triplicado: en forma facsimilar, paleográfica y crítica; *CORDIAM* también está facilitando de forma progresiva el acceso a los facsimilares de los subcorpus de documentos y prensa (además de proporcionar la descarga en PDF de la transcripción del documento completo). En la medida en que los corpus incorporen *transparencia*, en el sentido de permitir el acceso a sus fuentes documentales, tendrán un valor añadido para ayudar a resolver las posibles dudas filológicas.

4.7 Relación con el usuario

La relación entre el usuario y «el corpus», entendido ahora como un ente difuso formado por filólogos y técnicos informáticos, podría compararse, en términos comerciales, con la «atención al cliente». Los motivos que llevan al usuario a intentar contactar con los responsables del diseño, mantenimiento o funcionamiento del corpus pueden ser variados; sus peticiones, reacciones, respuestas o *feedback* suelen tener la loable finalidad de comunicar errores o erratas para que se corrijan. Lo habitual es que se incluya en la interfaz de usuario de los corpus la posibilidad de contacto, generalmente por formulario o por correo electrónico (en el caso del *NDHE* incluso por correo postal). El corpus, por el hecho de estar en línea, debería contar, idealmente, con actualizaciones frecuentes, y no ser un producto cerrado; estos aspectos, sin embargo, afectan ya a la gestión, fuera de lo estrictamente filológico e informático.

Entre las actuaciones que debería incluir el mantenimiento filológico y técnico del corpus, se incluyen las correcciones de:

— erratas transmitidas desde las ediciones en que se basa el corpus,
— errores y confusiones en la datación de las ediciones,
— confusiones entre la fecha del original y de su traducción,
— modernizaciones del texto realizadas en distintas épocas,
— erratas en la digitalización debidas a malas lecturas del OCR, etc.

Otra cuestión de interés para el investigador es el acceso libre a los recursos. El único corpus con ciertas restricciones es el *CE*, que solicita un registro tras determinado número de consultas, y el pago de una suscripción para acceder a búsquedas más complejas ilimitadas.

4.8 Lematización y léxico oculto

La lematización automática, si bien ha avanzado, sigue presentando muchos errores, que aumentan a medida que se retrocede cronológicamente en los documentos del corpus. En estudios sobre textos medievales, p. ej., las variantes gráficas no pueden lematizarse de manera eficaz si no es con una intervención de los filólogos porque las posibilidades de variación no se limitan a las predecibles en la alternancia de las grafías, sino que suelen ir más allá.

En el caso del español, los arabismos medievales son especialmente problemáticos en este sentido, p. ej., la lematización de *alférez* debería recoger todas las variantes documentadas, al menos *alfférez, alphérez, alféreç, alferes; alférece; alferce, alferçe, alferze, alferse; alfériz; alfériç, alferis; alférit; alféraz; alféraci; alfiérez, alffiérez, alfiéreç; alfieres; alfiéret; alfiérece; alfierce, alfierze, alfierse; alfiéraz, alphiéraz; alfiérat; alfiáraz; alfrez; afáiret*, etc., formas extraídas del *DHLE* (1960–1996) y *DEM* (1987–2005). El registro de todas las variantes formales, en casos como este, es el resultado del trabajo lexicográfico y solo con muchas limitaciones la combinatoria del corpus podrá ayudar a localizarlas. En la mayor parte de los casos, solo será posible acceder a esas formas previendo antes su existencia y buscándolas luego; por el contrario, si se espera encontrarlas entre las ocurrencias del lema *alférez,* se perderá una parte de formas que son muy difíciles de localizar. Se trata de *léxico oculto*, que no es recuperable sin una lematización mucho más eficaz que la que tenemos ahora. En ocasiones, la lematización de los corpus históricos es posible que deba hacerse de forma inversa, a partir de los lemas y las variantes que se han determinado previamente en los ficheros y en los diccionarios.

Las mismas opciones de búsqueda múltiple que posibilitan la localización automática de variantes formales y su consiguiente conversión en lemas pueden ser un arma de doble filo y fuente de errores. P. ej., *chauz* (una especie de

alguacil) arroja entre las variantes formales que lematiza en *CNDHE chaucito*, fórmula de despedida afectuosa en algunas variedades americanas («¡chaucito amor!»); se comprueba que los factores semánticos que afectan a la lematización son más difíciles de automatizar que los estrictamente formales. Ante este tipo de problemas, se reafirma la consideración de los corpus como un complemento subordinado al fichero lexicográfico y a la base de datos léxicos lematizados (v. 2).

4.9 Léxico ausente

Otro problema es la escasa presencia en los corpus de cierto tipo de léxico poco usual, pero que, sin embargo, puede ser de interés en lexicografía histórica. P. ej., la búsqueda en los corpus de la voz *alecrín,* solo aporta las cinco ocurrencias del *CNDHE*, pero cuatro de ellas corresponden a un topónimo y la única relacionada con un tipo de escualo, que es la que interesa documentar históricamente, es muy reciente (1999). Sin embargo, es posible constatar su registro en diversos diccionarios americanos de los ss. XIX y XX y en textos ictiológicos recuperados de la *Biblioteca Digital Hispánica (BDH)*. En casos como este, se comprueba que no siempre la cantidad de datos, por sí sola, garantiza que se obtengan resultados: los 400 millones de palabras del corpus no son de interés para la redacción de esta entrada lexicográfica. Esto muestra que son necesarios otros tipos de recursos digitales para recuperar la información perdida en una galaxia de vocabularios, glosarios, diccionarios, polianteas, etc., que siguen siendo accesibles solo en papel o en digitalizaciones muy dispersas en Internet.

5 Otros recursos en línea

5.1 Bibliotecas, hemerotecas y directorios digitales

Las bibliotecas y hemerotecas digitales, imprescindibles para la lexicografía, están siendo desarrolladas por distintas instituciones que ponen a disposición de los investigadores recursos cada vez más útiles. Destaca la labor de la *BDH* de la Biblioteca Nacional de España, que está digitalizando materiales muy útiles para la búsqueda de recursos temáticos que se precisan en la redacción lexicográfica. Por otra parte, la *BDH* ha desarrollado un portal para integrar los recursos digitales de bibliotecas nacionales iberoamericanas: la *Biblioteca digital del Patrimonio Iberoamericano* (*BDPI*).

Con fines filológicos y, en concreto, lexicográficos, también es de interés la *Biblioteca Virtual de la Filología Española* (*BVFE*). Se trata, en realidad, de un directorio de enlaces que remite a las bibliotecas donde están digitalizadas las

obras, es decir, es un «recurso intermediario» que ayuda mucho a los lexicógrafos para paliar el problema de la dispersión de la información y ahorrar tiempo en búsquedas por un sinfín de catálogos de bibliotecas. Reúne los enlaces de acceso a más de 4500 diccionarios, vocabularios, glosarios, plantas de diccionarios, tratados de lexicografía, etc. Mediante la búsqueda avanzada es posible, además, aplicar filtros sobre aspectos de edición, fechas, lenguas y variedades lingüísticas, etc., lo que mejora sus resultados. El directorio cumple su cometido una vez puesta la obra a disposición del usuario, que no debe esperar, por tanto, poder hacer búsquedas dentro de los textos de un conjunto obras: el acceso se hace a un único ejemplar seleccionado, que puede ofrecer algún tipo de consulta o búsqueda en su interior dependiendo de cómo haya sido digitalizado por la biblioteca en que se ubique.

El formato habitual de la publicación en las bibliotecas digitales actualmente es PDF con OCR, por lo que la búsqueda de texto dentro de los documentos suele dar buen resultado (aun contando con errores de lectura en un porcentaje indeterminado, que depende, en gran medida, de la antigüedad de la digitalización). La mayor dificultad, no obstante, es la escasa estructuración digital de los datos léxicos, cuyo avance, en los próximos años, debe facilitar las investigaciones.

En cuanto a las hemerotecas digitales, son imprescindibles para la historia del léxico de los tres últimos siglos. Algunas de ellas informan de la cantidad de páginas escaneadas que contienen, lo que puede ser un indicio de su rendimiento para una investigación determinada. P. ej., la *Hemeroteca nacional digital de México* (*HNDM*) dispone de unos 9 millones de páginas, la *Biblioteca Virtual de Prensa Histórica* (*BVPH*) —una hemeroteca, a pesar de su nombre— 7.5 millones y *Jable: Archivo de prensa digital de Canarias* cuenta con unos 6 millones. Además de los problemas técnicos similares a los de las bibliotecas, tienen unas peculiaridades para el trabajo lexicográfico que deben considerarse:

a) Limitación cronológica: Son útiles sobre todo para los ss. XIX y XX, en este periodo sí pueden aportar primeras documentaciones y, además, contextos de uso, acompañados en ocasiones con fotografías o ilustraciones de interés enciclopédico.
b) Concentración geográfica: En el caso de la *Hemeroteca digital* (*HD*) de la BNE, la mayoría de los documentos pertenece a España, por lo que su utilidad desde una perspectiva panhispánica es limitada. La *HNDM* es otra valiosa fuente para América, pero aún faltan recursos como este en la mayoría de los países hispanohablantes; ante esta carencia, se hace necesario recurrir a las hemerotecas privadas de los principales diarios de estos países, que suelen estar accesibles en línea.

c) Dispersión: Incluso en lugares como España, donde se dispone de muchos materiales digitalizados, no es posible hacer consultas sin realizar repetidas y tediosas búsquedas, pues tampoco aquí se cuenta con un acceso centralizado; esto se debe a los varios niveles de orden político-administrativo de los que depende la digitalización de este tipo de recursos. La oferta de contenidos tiene una especialización territorial, aunque a menudo se superponen e interfieren los niveles locales, provinciales, autonómicos, etc., a los que se suman los que dependen de archivos, universidades, fundaciones, proyectos de investigación… Esta enorme dispersión es un obstáculo para realizar investigaciones, ya que se multiplica la necesidad de búsquedas (con el añadido de que este tipo de recursos, que no son específicos para lingüistas, no lematizan las formas que contienen). Sin embargo, hay que contar con que, en cualquier hemeroteca o biblioteca digital, pueden permanecer ocultas voces con dataciones tempranas o con usos de interés lexicográfico, alguna se podría encontrar, p. ej., en una relación de sucesos de la *Biblioteca digital del siglo de Oro* (*BIDISO*) o en cualquier otro sitio donde, si no se dispone de tiempo para múltiples búsquedas, quedará oculto. A pesar de estas dificultades, las biblio-hemerotecas digitales son de inestimable valor, sobre todo cuando el objeto de estudio es local (dialectología histórica, p. ej.). La interconexión de bases de datos es imprescindible para recopilar informaciones de interés lexicográfico, sin embargo, las posibilidades de búsqueda centralizada en prensa y revistas históricas son, por el momento, muy limitadas.

5.2 Portales y agregadores

Otro recurso de interés para el lexicógrafo es *HISPANA*, portal de acceso a las colecciones digitales de archivos, bibliotecas y museos españoles; es una fuente de información no solo bibliográfica, sino de todo tipo de documentos, incluidas imágenes. Su potencial utilidad consiste en lanzar consultas simultáneas sobre varios repositorios y bases de datos (permite también suscribirse a las fuentes de noticias, lo que resulta útil para estar al tanto, p. ej., de nuevos proyectos de digitalización). Entre las herramientas para simplificar la tarea del lexicógrafo, es necesario el desarrollo de sistemas como este, que sean capaces de localizar léxico de interés en distintos motores de búsqueda y mostrar combinados los datos léxicos que se extraigan de ellos. Algunos proyectos españoles de humanidades digitales se han agrupado en el portal *ARACNE*: *Red de Humanidades Digitales y Letras Hispánicas*, que incluye un metabuscador, marcando una pauta para evitar la dispersión y falta de estándares de que adolecen muchas de las nuevas investigaciones.

5.3 Bases de datos lexicográficas, tesoros lexicográficos y redes de diccionarios

Los diccionarios antiguos y otros repertorios léxicos de interés histórico son un recurso fundamental para los lexicógrafos, pues es posible encontrar en ellos valiosas documentaciones que escapan a otras fuentes: p. ej., palabras que por ser exotismos, indoamericanismos o por otros motivos aparecen tarde en los textos pero se documentan en diccionarios; en otros casos, lo que encontramos son constataciones de usos ya definitivamente fijados al adquirir estatus lexicográfico. Sin embargo, la digitalización de este tipo de fuentes no se ha desarrollado lo suficiente. A finales de los noventa, se hicieron algunas recopilaciones en CD-ROM que contenían diccionarios clásicos de España (Álvarez de Miranda 1998) y América (Haensch 2000), pero quedaron muy pronto obsoletas debido a su formato en imagen y a lo rudimentario de las búsquedas. Como recurso digital, se cuenta desde hace ya tiempo con el *Nuevo tesoro lexicográfico de la lengua española* (*NTLLE*); a pesar de su interfaz obsoleta, sigue siendo útil (p. ej., la búsqueda de la palabra *alecrín* —v. 4.9— realizada en todos los corpus, sin resultados de interés, se puede encontrar en 21 diccionarios gracias al *NTLLE*).

Son necesarias digitalizaciones eficaces para pasar de los viejos *tesoros lexicográficos* a un nuevo concepto de *redes de diccionarios*, estructurados e interconectados. Una muestra de las posibilidades de transformación es la base de datos *Tesoro.PR*, elaborada a partir del *Tesoro lexicográfico del español de Puerto Rico* publicado en papel (Vaquero/Morales 2005), que es accesible ahora en línea desde cualquiera de los campos habituales en la microestructura de un diccionario y está, además, etiquetada por campo temático en cada acepción; a esto se añade la conversión de las definiciones del *Tesoro* en un corpus en el que se pueden realizar búsquedas. La creación de nuevos formatos digitales a partir de los tesoros ya existentes abre vías para el estudio del léxico y muestra también el camino que deben seguir los proyectos que ya están en gestación. En el desarrollo de estos nuevos recursos sería deseable la búsqueda de estándares para facilitar el acceso a consultas múltiples.

5.4 Otros recursos mixtos

A todo lo anterior, hay que sumar una serie de recursos mixtos, entre la biblioteca digital y el corpus, que recopilan colecciones limitadas de textos y ofrecen algunos tipos de búsquedas útiles. Sus precedentes están en los CD-ROM que a finales de los noventa publicó el Hispanic Seminary of Medieval Studies de Madison, que estaban en texto plano y eran muy limitados. Ahora se encuentran accesibles en línea en forma de *Biblioteca Digital de Textos del Español Antiguo*

(*BIDTEA*) que, aunque no cuenta con las posibilidades de consulta que habitualmente ofrecen los corpus, sobrepasa las funciones que corresponden a una biblioteca digital, pues ofrece acceso interactivo a 345 textos con índices, concordancias y transcripciones semipaleográficas. Otro proyectzo inspirado en la misma institución americana dio lugar a los CD-ROM con textos medievales del *Archivo Digital de Manuscritos y Textos Españoles* (*ADMYTE*), que incorporó posibilidades de búsqueda y una limitada lematización; en su momento supuso un avance cuyos logros son consultables aún previa suscripción en línea.

6 Conclusiones generales

La proliferación de proyectos digitales de distintos tipos de textos pone a disposición de los lexicógrafos y de los filólogos en general nuevos materiales que permitirán aportar datos de interés para el conocimiento de la historia del léxico, especialmente en cuanto al adelanto de primeras documentaciones y a la atestiguación de voces en lugares fuera de España hasta ahora menos accesibles.

No obstante, la dispersión de los recursos y la falta de modelos comunes para los distintos proyectos de digitalización dificulta la búsqueda en las nuevas fuentes. Solo en la medida en que se vayan creando estándares, se dote de metadatos a los documentos y se perfeccionen los filtros de búsqueda se solucionarán los problemas actuales de sobreinformación y de dispendio de tiempo improductivo a los que se enfrenta el investigador de la historia del léxico.

Desde una perspectiva estrictamente lexicográfica, una necesidad urgente es la digitalización de diccionarios y repertorios léxicos de todo tipo, relegados en parte por el protagonismo de los corpus en los últimos años; es necesario, sobre todo, reformular digitalmente los tesoros lexicográficos y crear redes de diccionarios para superar la dependencia que aún existe de las obras en soporte papel y CD-ROM o de sus simples trasuntos en la web.

Mientras tanto, los corpus se han convertido en recurso básico para los lingüistas, pero su manejo debe ir precedido de un análisis de sus posibilidades y limitaciones para poder asegurar que aportan un rendimiento satisfactorio para el trabajo lexicográfico.

Por otra parte, es necesario desarrollar el concepto de archivo lexicográfico y entenderlo como base de datos con interfaz que incorpore herramientas de gestión y edición. En él se deben integrar tanto corpus *ad hoc* como enlaces de acceso a corpus externos y recursos en línea, especialmente a tesoros lexicográficos y a biblio-hemerotecas digitales que, a medida que se vayan haciendo más eficaces con metadatos y filtros, permitirán dar un nuevo paso adelante a la lexicografía histórica digital.

Referencias bibliográficas

ADMYTE = Archivo Digital de Manuscritos y Textos Españoles. <http://www.admyte.com/admyteonline/home.htm> [último acceso: 15/10/2017].

Álvarez de Miranda, Pedro (comp.) (1998): *Lexicografía española peninsular. Diccionarios clásicos*. [Recurso electrónico]. Madrid: Fundación Histórica Tavera.

ARACNE = Red de Humanidades Digitales y Letras Hispánicas. <http://www.red-aracne.es> [último acceso: 10/11/2017].

Arnold, Rafael/Jutta Langenbacher-Liebgott (2018): «El caudal léxico del español medieval y el nuevo proyecto DEM *electrónico* (DEM*el*)», en Roberto Antonelli *et al*. (eds.) *Atti del XXVIII Congresso Internazionale di Linguistica e Filologia Romanza (Roma, 18–23 Iuglio 2016)*, vol. 1. Estrasburgo: Éditions de Linguistique et de Philologie: 789–798.

BDH = Biblioteca Digital Hispánica. <http://www.bne.es/es/Catalogos/BibliotecaDigitalHispanica/Inicio/index.html>. [último acceso: 15/11/2017].

BDPI = Biblioteca digital del Patrimonio Iberoamericano. <http://www.iberoamericadigital.net/es/Inicio/> [último acceso: 15/11/2017].

Bertolotti, Virginia/Concepción Company (2014): «El *Corpus Diacrónico y Diatópico del Español de América* (*CORDIAM*). Una propuesta de tipología textual», *Cuadernos de la ALFAL* 6 (número especial: Claudia Parodi y Micaela Carrera de la Red (eds.), *El español en América. Corpus y textos*), 130–148. <http://www.mundoalfal.org/sites/default/files/revista/06_cuaderno_011.pdf [último acceso: 15/11/2017].

Biblia medieval <http://www.bibliamedieval.es/index.php> [último acceso: 15/11/2017].

BIDISO = Biblioteca digital del Siglo de Oro. <https://www.bidiso.es/index.htm> [último acceso: 10/11/2017].

BIDTEA = Biblioteca Digital de Textos del Español Antiguo. <http://www.hispanicseminary.org/textconc-es.htm> [último acceso: 10/10/2017].

BVFE = Alvar Ezquerra, Manuel (dir.): *Biblioteca virtual de la filología española*. Directorio bibliográfico de gramáticas, diccionarios, obras de ortografía, ortología, prosodia, métrica, diálogos e historia de la lengua. <https://www.bvfe.es/> [último acceso: 15/11/2017].

BVPH = Biblioteca Virtual de Prensa Histórica. <http://prensahistorica.mcu.es> [último acceso: 10/11/2017].

Campos Souto, Mar (2017): «Hacia una crónica del *Diccionario histórico* de la lengua española de 1933–1936: Los materiales del Archivo de la Real Academia Española», *BRAE* XCVII/CCCXV, 161–201.

Campos Souto, Mar/José A. Pascual (2012): «Lexicografía, filología e informática: una alianza imprescindible. A propósito de la situación del *NDHE*», en Dolores Corbella *et al.* (eds.), *Lexicografía hispánica del siglo XXI: Nuevos proyectos y perspectivas. Homenaje al profesor Cristóbal Corrales Zumbado*. Madrid: Arco/Libros, 151–170.

CDH-N = Instituto de Investigación Rafael Lapesa de la Real Academia Española (2013): *Corpus del Nuevo diccionario histórico nuclear*. <http://web.frl.es/CNDHE> [último acceso: 15/11/2017].

CE = *Corpus del español* <www.corpusdelespanol.org> [último acceso: 01/11/2017].

CHARTA = *Corpus Hispánico y Americano en la Red: Textos Antiguos*. <http://www.corpuscharta.es> [último acceso: 15/11/2017].

CNDHE = Instituto de Investigación Rafael Lapesa de la Real Academia Española (2013): *Corpus del Nuevo diccionario histórico*. <http://web.frl.es/CNDHE> [último acceso: 15/11/2017].

CODEA+ = GITHE (Grupo de Investigación Textos para la Historia del Español): *CODEA+ 2015* (*Corpus de documentos españoles anteriores a 1800*). <http://corpuscodea.es> [último acceso: 15/11/2017].

CODEMA = *Corpus Diacrónico de Documentación Malagueña*. <http://www.arinta.uma.es/contenidos/codema_buscador.action> [último acceso: 10/10/2017].

CORDE = Real Academia Española: Banco de datos (CORDE). *Corpus diacrónico del español*. <http://www.rae.es> [último acceso: 15/11/2017].

CORDEREGRA = Calderón Campos, Miguel/M.ª Teresa García Godoy (dirs.): *Corpus diacrónico del español del Reino de Granada. 1492–1833*. <http://www.corderegra.es> [último acceso: 15/11/2017].

CORDIAM = Academia Mexicana de la Lengua. *Corpus Diacrónico y Diatópico del Español de América*. <www.cordiam.org> [último acceso: 10/11/2017].

CorLexIn = Morala Rodríguez, José R. (dir.): *Corpus Léxico de Inventarios (CorLexIn)*. <http://web.frl.es/CORLEXIN.html> [último acceso: 15/11/2017].

CORPES XXI: Real Academia Española: Banco de datos (*CORPES XXI*). *Corpus del español del siglo XXI*. <http://www.rae.es> [último acceso: 01/10/2017].

CREA = Real Academia Española: Banco de datos (*CREA*). *Corpus de referencia del español actual*. <http://www.rae.es> [último acceso: 15/11/2017].

Davies, Mark (2017): «Compare to other corpora: for historical research (compared to *CORDE*)». <http://www.corpusdelespanol.org/compare.asp> [último acceso: 01/11/2017].

DEM = Müller, Bodo (1987–2005): *Diccionario del español medieval*, vol. 1, fascículos 1–10, vol. 2, fascículos 11–20, vol. 3, fascículos 21–26. Heidelberg: Winter.

DEMel = *Diccionario del Español Medieval electrónico*. <http://go.upb.de/DEMel> y <http://www.romanistik.uni-rostock.de/forschung/sprachwissenschaft/DEMel/> [último acceso: 20/11/2017].

DHLE = Real Academia Española (1960–1996): *Diccionario histórico de la lengua española*. Madrid: Imp. Aguirre.

Fajardo, Alejandro (2012): «La lexicografía histórica actual: tradición y nuevas perspectivas», en Mar Campos *et al*. (eds.), *Assi como es de suso dicho. Estudios de morfología y léxico en homenaje a Jesús Pena*. San Millán de la Cogolla: Cilengua, 195–211.

Haensch, Günther (comp.) (2000): *Textos clásicos sobre la historia de la lexicografía del español en América* [Recurso electrónico]. Madrid: Fundación Histórica Tavera.

HD = *Hemeroteca digital*. <http://www.bne.es/es/Inicio/> [último acceso: 15/10/2017].

HISPANA = <http://hispana.mcu.es> [último acceso: 15/11/2017].

HNDM = *Hemeroteca nacional digital de México*. <http://www.hndm.unam.mx> [último acceso: 15/11/2017].

Jable = *Jable. Archivo de prensa digital de Canarias*. <https://biblioteca.ulpgc.es/jable> [último acceso: 15/11/2017].

Kabatek, Johannes (2016): «Un nuevo capítulo en la lingüística histórica iberorrománica: el trabajo crítico con los corpus. Introducción a este volumen», en Johannes Kabatek (ed.), *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín/Boston: De Gruyter, 1–17.

Kasten, Lloyd A./John J. Nitti/Jean Anderson (1978): *Concordances and texts of the Royal Scriptorium manuscripts of Alfonso X, El Sabio*. [Microfichas]. Madison: Hispanic Seminary of Medieval Studies.

NDHE = Real Academia Española (2013–): *Nuevo diccionario histórico del español*. <http://frl.es/DH> [último acceso: 15/11/2017].

NTLLE = Real Academia Española (2001): *Nuevo tesoro lexicográfico de la lengua española*. <http://ntlle.rae.es/ntlle/SrvltGUILoginNtlle> [último acceso: 15/11/2017].

Post Scriptum = CLUL (ed.) (2014): *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. <http://ps.clul.ul.pt.> [último acceso: 15/11/2017].

Rojo, Guillermo (2016): «Corpus textuales del español». <https://gramatica.usc.es/~grojo/Publicaciones/corpus_textuales_espanol_borrador_final.pdf> [último acceso: 05/09/2017].

Tesoro.PR = Tesoro lexicográfico del español de Puerto Rico. <https://tesoro.pr> [último acceso: 15/11/2017].

Vaquero, María/Amparo Morales (2005): *Tesoro lexicográfico del español de Puerto Rico*. Puerto Rico: Academia Puertorriqueña de la Lengua Española.

Francisco Javier Herrero Ruiz de Loizaga

Algunos problemas en la aplicación de los corpus informatizados al estudio de la diacronía del español, con especial atención a los procesos de gramaticalización[1]

Resumen: En los estudios diacrónicos sobre la lengua ha sido siempre necesario partir del análisis de corpus. Con la llegada de los corpus informatizados, el modo de trabajar en algunas parcelas de los estudios lingüísticos ha experimentado un inevitable cambio, que afecta también a los estudios diacrónicos. Analizamos y valoramos distintos corpus útiles para el estudio diacrónico del español, especialmente en relación con los procesos de gramaticalización. Aunque tenemos en cuenta también otros corpus especializados, nos centramos en los corpus diacrónicos de carácter general, *CE* de Davies y *CORDE*, *CREA* y *CORPES* académicos. Partiendo de su innegable utilidad, son instrumentos de los que no podemos prescindir para los estudios históricos sobre el español, llamamos también la atención sobre los problemas que se plantean y las precauciones que hay que tener para su empleo.

Palabras clave: Corpus informatizados, Lingüística diacrónica, Gramaticalización

Abstract: In the diachronic studies about language it has always been necessary to start from the analysis of the corpora. After the arrival of computerized corpora, the way of working in some fields of the linguistic studies has undergone an inevitable change, which also affects the diachronic studies. We analyze and evaluate different corpora that can be used for the diachronic study of Spanish, especially in relation to the grammaticalization processes, and although we also take into account other specialized corpora, we focus on the diachronic corpora of a general nature, *CE* of Davies and *CORDE*, *CREA* and *CORPES* of the Real Academia. Although they are very useful instruments that we can't ignore for historical studies in Spanish, we call attention to the problems that arise in its use and the precautions that must be taken.

Keywords: Computerized corpus, Diachronic Linguistics, Grammaticalization

---

1 Este trabajo se inscribe en el marco del proyecto de investigación FFI2015-64080-P, *Procesos de gramaticalización en la historia del español (V): gramaticalización, lexicalización y análisis del discurso desde una perspectiva histórica*, del Ministerio de Economía y Competitividad.

1 Introducción. La utilización de corpus en los estudios de lingüística histórica

El estudio de cualquier lengua en su perspectiva diacrónica está claramente ligado a la utilización de corpus. Esto es algo metodológicamente necesario, puesto que solo mediante el estudio de textos podemos llegar a tener conocimientos más o menos fiables de los hechos lingüísticos de épocas pasadas. De este modo, en el surgimiento de los modernos estudios diacrónicos sobre el español, la utilización de corpus, de diversas características y extensión, ha sido una constante. Por citar algún ejemplo señalado, podemos ver cómo la *Gramática del Cid*, de Menéndez Pidal (1976[5]), analiza exhaustivamente el texto del *Cantar* para realizar una detallada exposición de los diversos aspectos, fonéticos, morfológicos y sintácticos de la obra, aunque tenga también en cuenta otros textos anteriores y posteriores.

Esta línea de trabajos sobre diversos aspectos de la diacronía del español, basados en la lectura y despojo de un corpus, que puede estar constituido por una sola obra o un conjunto más o menos amplio de ellas, en función del interés de la investigación y los fenómenos estudiados, se mantiene a lo largo de la mayor parte del siglo XX. Pero con el desarrollo de los medios informáticos en las últimas décadas del pasado siglo, comenzaron también a realizarse corpus informatizados, que tendrán una importante repercusión en los estudios sobre las diversas lenguas. Concretamente, algunos de estos corpus ofrecerán un conjunto de datos de gran interés para el estudio de la diacronía del español. El primer corpus informatizado de importancia en los estudios diacrónicos del español fue el *ADMYTE* (Admyte 0, 1990 y Admyte 1, 1991), llevado a cabo por Francisco Marcos Marín, Charles Faulhaber, Ángel Gómez Moreno y Antonio Cortijo Ocaña, en el que se transcribe y edita en CD-ROM una amplia serie de textos medievales españoles, junto con un facsímil de los mismos (en *ADMYTE* I, no en *ADMYTE* II). Posteriormente se habilitó su acceso en red bajo suscripción.

La aparición de corpus informatizados que reúnen un volumen notable de textos en castellano supuso sin duda un cambio importante en el modo de enfrentarse a los estudios diacrónicos del español: no hay duda de que el recurso a la utilización de los corpus sigue siendo completamente indispensable en el estudio de la diacronía de cualquier lengua, pero la posibilidad de acceder con pulsar una tecla a una cantidad de datos que anteriormente podría haber supuesto un tiempo muy prolongado de lectura (a veces incluso imposible de realizar en un período razonable por un único investigador) llevará en muchos casos a un retroceso (a veces incluso abandono) en el recurso a la lectura directa y despojo manual del texto, en favor de la utilización del recurso electrónico.

No obstante, hay que señalar que existen distintos tipos de repertorios informatizados con distintas capacidades, y su utilidad será muy diferente en función de su extensión, especialización, tipo de lematización y motor de búsqueda.

2 Distintos corpus informáticos y su rendimiento en los estudios diacrónicos

En las últimas décadas se han venido elaborando una serie de corpus[2] que recogen un conjunto amplio, pero limitado, de textos de una determinada diacronía y tipo. La elaboración de estos corpus es muy cuidadosa, y los resultados son textos que proporcionan un grado muy alto de fiabilidad en las muestras que ofrecen cuando el investigador realiza una búsqueda determinada. A este tipo de corpus pertenecen la *Biblioteca Digital de Textos del Español Antiguo* del Medieval Seminary of Hispanic Studies de la Universidad de Madison[3], el corpus *Biblia medieval*, desarrollado por Andrés Enrique-Arias[4], el *CODEA+*[5], *Corpus de documentos españoles anteriores a 1800*, desarrollado por el equipo que dirige Pedro Sánchez-Prieto Borja, o el *CORDIAM*, *Corpus diacrónico y diatópico del español de América*, cuya elaboración ha sido dirigida por Concepción Company y Virginia Bertolotti. Estos corpus ofrecen conjuntos cuidados de textos, pero no muy amplios, a veces incluso divididos en diversos corpus menores, como sucede en el caso de la *Biblioteca Digital de Textos del Español Antiguo*, que comenzó ofreciendo en línea desde 2011 los 20 archivos que componen el corpus de la *Obra en prosa de Alfonso X el Sabio*, al que siguieron otros como el de textos médicos españoles, textos cronísticos españoles, textos navarro-aragoneses, hasta un total de 8 por el momento, que, según declaran sus creadores, serán ampliados en el futuro. Las transcripciones que ofrecen estos corpus son

2 Sobre corpus informatizados para el estudio del español, tanto de carácter sincrónico como diacrónico, v. Rojo (2016).

3 <http://www.hispanicseminary.org/textconc-es.htm>.

4 Sobre las posibilidades de explotación del corpus *Biblia medieval* para los estudios diacrónicos, v. Enrique-Arias (2009 y 2012).

5 El *CODEA+* forma parte del corpus *CHARTA* (*Corpus Hispánico y Americano en la Red. Textos Antiguos*), en el que se integran, además de *CODEA*, una serie de subcorpus como *CORHEN*, *COTAGAL*, *CORDEREGRA*, etc. (v. <http://www.corpuscharta.es/grupos.html>). De esta serie de subcorpus nos referiremos al *CODEA*, el más amplio y con su propio portal en el que en estos momentos puede accederse a un número mayor de documentos que el que es accesible a través del portal de *CHARTA*. Sobre el *CODEA*, v. Sánchez-Prieto (2012) y Sánchez-Prieto *et al.* (2009).

de carácter semipaleográfico, como en la *Biblioteca Digital de Textos del Español Antiguo*[6] o el *CORDIAM*[7], o paleográfico, como en el corpus *Biblia medieval* o en el *CODEA*, como explícitamente se declara en ellos. Tanto la *Biblia medieval* como el *CODEA* ofrecen además facsímiles de los textos que editan, lo que permite que el investigador, si lo necesita, pueda comprobar la fidelidad del texto reproducido[8]. Los motores de búsqueda son semejantes, y en general útiles para la búsqueda léxica, pero menos para la búsqueda destinada a la investigación de carácter gramatical. La *Biblioteca Digital de Textos del Español Antiguo* solo permite hacer búsquedas por palabras, prefijos y sufijos, las posibilidades de búsqueda en *Post Scriptum* son similares: palabras, comienzo y terminación de palabras o secuencias en el interior de una palabra; el corpus *Biblia medieval*[9], el *CODEA* o el *CORDIAM* permiten también búsquedas por secuencias de palabras. El *CODEA* (y en general *CHARTA*) y el *CORDIAM* permiten además buscar por colocaciones. La fiabilidad de estos corpus es elevada, y su utilidad manifiesta, pero limitada a estudios específicos sobre un determinado espacio geográfico, cronológico, o un determinado registro.

Pero indudablemente, los corpus que tienen mayor importancia para el estudio diacrónico, de los que puede obtenerse una mayor rentabilidad, son los corpus

6 En Mackenzie (1997[5]: viii) se declara explícitamente: «However the transcription, though close, is not paleographic».

7 En el *CORDIAM* no se dice explícitamente el tipo de transcripción que se realiza. Parece haber cierta libertad por parte de los diversos colaboradores a la hora de transcribir los textos, aunque hay después un proceso de sistematización: «Los corpus fueron subidos a *CORDIAM* tal como fueron autorizados y enviados por los colaboradores (ver Colaboradores y Corpus en esta página de inicio). No obstante, para facilitar las búsquedas y la interpretación de los documentos, fue realizada una mínima sistematización ecdótica y filológica que atañe, en lo esencial, a abreviaturas, separación de palabras e intervenciones en el texto, y que atañe en menor medida a la información contenida en los metadatos» (<http://www.cordiam.org/doc/edicion-y-sistematizacion.html>).

8 También entre los materiales del *CORDIAM* hay facsímiles de muchos de los documentos que digitaliza, pero no de todos. Depende de que los hayan proporcionado o no los colaboradores. En los metadatos de cada documento se especifica si hay o no facsímil disponible, aunque por el momento no hay acceso a él. Sobre las características y desarrollo de *CORDIAM*, v. Bertolotti/Company (2014).

9 El corpus *Biblia medieval* está integrado también dentro de los que ofrece la *Biblioteca Digital de Textos del Español Antiguo*, pero tiene además un portal independiente con un motor de búsqueda que permite búsquedas más afinadas y ofrece en columnas paralelas el texto de las distintas versiones bíblicas correspondientes al pasaje seleccionado.

generales, que recogen textos de todas las épocas del idioma y de distintos registros y tipos muy variados de documentos. Entre ellos están el *Corpus del Español* de Davies (*CE*) y los corpus académicos como *CORDE* y *CDH*, a los que nos referiremos con más detalle.

3 El *Corpus del español* y los corpus académicos

3.1 Caracterización

El *Corpus del español* de Davies (*CE*), en el subcorpus género/histórico[10], el que nos interesa para los estudios diacrónicos, reúne un volumen de textos que suma unos 100 millones de palabras. El *CORDE* académico recoge 236 709 914 en textos que van desde el siglo VIII a 1974. Ambos corpus presentan, frente a los anteriores, un volumen muy superior de textos fichados, un registro temporal mucho más amplio desde el español medieval al moderno, y una variedad de tipos de textos también mucho mayor (el desplegable del *CORDE*, en el apartado tema, da una lista de tipos muy variados de textos. Cada uno de estos tipos puede seleccionarse si desea hacerse una búsqueda específica). El *CDH* (*Corpus del Nuevo diccionario histórico del español*)[11], reúne 355 740 238 registros que básicamente corresponden a los textos fichados tanto en *CORDE* como en *CREA* entre el siglo XII y el año 2000. Como complemento de estos corpus, son también muy útiles para los estudios históricos los corpus académicos *CREA* y *CORPES XXI*, que recogen textos de los periodos cronológicos comprendidos entre 1975 y 2005 el primero, y entre 2000 y 2015 el segundo, y el subcorpus web/dialectos del *Corpus del español* de Davies. En el CREA hay algo más de 160 millones de formas y el material se distribuye, aproximadamente, al 50 % entre textos españoles e hispanoamericanos.[12] Respecto al *CORPES XXI*, como la propia Academia indica, «La distribución general del *CORPES* asigna un 30 % del total a formas procedentes de España y un 70 % a formas procedentes de América»[13]. En su estado actual (desde junio de 2016), están incorporados a este corpus 237 678 documentos con unos 225 millones de formas[14]. El *Corpus del español* de Mark Davies (CE), en su nueva versión accesible en red desde 2016, recoge para el español actual,

10 <http://www.corpusdelespanol.org/hist-gen/>.
11 <http://www.rae.es/recursos/banco-de-datos/cdh>.
12 <http://www.rae.es/recursos/banco-de-datos/crea>.
13 *Corpus del español del siglo XXI (CORPES) Descripción del sistema de codificación*, <http://www.rae.es/sites/default/files/CORPES_Sistema_de_codificacion_12_2015.pdf>.
14 <http://www.rae.es/recursos/banco-de-datos/corpes-xxi>.

en el subcorpus web/dialectos[15], 1985 000 000 palabras de los 21 países hispanohablantes, lo que le convierte en el mayor corpus del español disponible en la actualidad, con la limitación de estar constituido exclusivamente por textos recogidos de la web en 2013–2014 (aunque, según se dice en su propia página web, está previsto para el 2018 aumentar el número de textos y el período cronológico de los mismos a 2010–2018)[16], y la de no permitir hacer una selección por géneros o temas.

Estos grandes corpus, mucho más extensos en su alcance geográfico, cronológico y con mayor variedad de registros, resultan fundamentales para el estudio de la historia del español, ya que permiten tener una visión más global y abarcadora del estado de lengua y su evolución a lo largo del tiempo que los corpus parciales anteriormente mencionados.

3.2 Algunas cuestiones relativas a la lematización

Estos corpus diacrónicos amplios presentan además un motor de búsqueda que permite búsquedas más complejas que la de los corpus más especializados, algunos de los cuales se limitaban a facilitar la posibilidad de búsquedas léxicas, unas veces permitiendo cadenas de palabras y otras solo palabras aisladas. Tanto el *CORDE* como el *CE* permiten no solo la búsqueda de palabras y secuencias de palabras, sino también la búsqueda por comienzos o terminaciones, la búsqueda de palabras en las que se ofrecen una o más letras no especificadas en su interior, o la búsqueda de secuencias colocadas a una determinada distancia unas de otras[17]. Estas mayores posibilidades de los motores de búsqueda permiten también una mayor utilidad para su empleo en estudios de carácter sintáctico. Así, por ejemplo, pueden estudiarse las correlaciones del tipo *tanto … que*, *tanto … como* o *apenas … cuando*, para lo que es preciso recurrir al criterio de distancia entre los elementos que las conforman. El *CE* presenta también una lematización de las palabras que aparecen en el corpus en función de su categoría gramatical: sustantivos, adjetivos (con posibilidad de variación de género y número), pronombres, determinantes, verbos (con sus diversas posibilidades de flexión), adverbios, conjunciones, preposiciones, interjecciones, etc.; permite la búsqueda por términos exactos o lemas, con todas las variaciones formales que

15 <http://www.corpusdelespanol.org/web.dial/>.

16 V. «Comparación con otros corpus», <http://www.corpusdelespanol.org/compare_corpes_s.asp>.

17 Estas posibilidades, de un modo más limitado, las ofrece también el *CODEA+*.

corresponderían a un sustantivo, adjetivo o verbo, lo que se indica poniendo el término que queramos buscar entre corchetes o en mayúsculas. Por ejemplo, si pedimos que busque [haber] o HABER nos daría todas las distintas formas del verbo. Por otra parte, también podemos pedir que busque una palabra o lema colocado de un modo concreto respecto a determinadas categorías gramaticales. Esto permite, obviamente, obtener un rendimiento muy alto para la búsqueda de determinadas construcciones, y lo convierte en una herramienta poderosa para el estudio de determinados hechos sintácticos[18]. No hay una lematización semejante en el *CORDE*, que por tanto presenta menos utilidad para algunos estudios históricos de carácter gramatical. No obstante, el corpus académico presenta otras ventajas como su mayor volumen de datos, la posibilidad de ofrecer los datos por países o la posibilidad de parcelar los períodos cronológicos por años, frente a la posibilidad de hacerlo solamente por siglos que ofrece el *CE*, o una parcelación más detallada por tipos de texto[19]. Estas posibilidades de búsqueda más afinada para determinadas cuestiones gramaticales han propiciado que diversos estudios diacrónicos se realicen sobre el *CE*, a pesar del menor volumen de textos de este corpus frente a *CORDE*. Así, por ejemplo, el estudio de Fernández (2015) sobre perífrasis verbales en el español clásico utiliza como corpus electrónico prioritariamente el de Mark Davies, dado que permite encontrar cualquier forma de un verbo más un infinitivo, gerundio o participio, y, en su caso, la intercalación de una preposición, algo

18 Estos corpus son indudablemente útiles para el estudio de muchos fenómenos gramaticales, y entre ellos muchos de los que atañen a la gramaticalización, donde seguimos el comportamiento de una palabra o grupo de palabras (v. Buenafuentes/Sánchez 2012 para la aplicación de los corpus académicos *CORDE* y *CREA* al estudio de la gramaticalización y la lexicalización, con una aplicación a algunos casos concretos, como la gramaticalización de *esquina a* y la lexicalización de *sepan cuantos*). No resultan, sin embargo, tan útiles para estudiar de un modo satisfactorio otros fenómenos sintácticos, como la evolución del orden de palabras o los diversos constituyentes de los márgenes oracionales; o para estudios de carácter pragmático, como las formas de cortesía, saludos, despedidas, maneras de dar órdenes, etc., aunque a veces sí puedan servir como instrumento auxiliar. A algunas de estas limitaciones se ha referido Enrique-Arias (2012: 422). Rojo (2012: 435–36) señala algunas ventajas y desventajas del método tradicional frente al trabajo con corpus.

19 Davies (2002) compara las características del *CE* y del *CORDE*, y en (2005, 2008 y 2009) compara *CE* con *CORDE* y *CREA*, señalando la mayor flexibilidad y rapidez de búsqueda de su corpus frente a los corpus académicos. Por su parte Rojo (2010) realiza también una comparación entre *CORDE* y *CREA* y el *CE* para señalar, en cambio, las ventajas que ofrecen los corpus académicos frente al de Davies en la investigación histórica.

que no podríamos realizar con *CORDE*[20]; también el estudio de Herrero (2016b) sobre el conector consecutivo *de ahí (que)*, parte del corpus de Davies, pues el *CE* permite con más facilidad buscar las secuencias del tipo *de ahí/aquí/allí* + verbos que presentan una inferencia de lo anteriormente dicho, a partir de las cuales se inició la gramaticalización de la secuencia *de aquí/ahí* como conector. El estudio de Granvik (2018) sobre las completivas de sustantivo sigue el procedimiento de emplear ambos corpus, *CE* y *CORDE*: en una primera fase extrajo «500 casos de la secuencia N *de que* por siglo del *Corpus del español* (Davies 2002) usando la fórmula [NN* *de que*], en la que [NN*] hace referencia a cualquier elemento nominal», y en una segunda utilizó el *CORDE* para el examen de los 34 sustantivos seleccionados a partir de la extracción anterior para su análisis específico.

El *CDH* sí incorpora una lematización que incluye las características gramaticales de las palabras incluidas en el corpus, y aunque no tiene el mismo nivel de detalle que el *CE*, y en consecuencia la misma flexibilidad, sí aumenta la capacidad de este corpus académico para las búsquedas gramaticales. En el *CDH*, hay una etiquetación por categorías gramaticales, pero sin tener en cuenta las variaciones formales de esas categorías. De este modo podemos pedir, por ejemplo, que busque el verbo *andar* como lema, lo que implica la búsqueda de todas las formas del verbo *andar* (posibilidad que no tiene el *CORDE*) seguido de la categoría gramatical verbo, sin posibilidad de especificar la forma verbal. Es claro que en la mayoría de los casos serán gerundios o participios lo que aparezcan, pero habrá que revisar cada caso para ver a qué tipo de perífrasis corresponde. El *CORPES XXI* y la versión anotada de *CREA* presentan una interfaz semejante a la de *CDH*, pero el motor de búsqueda es más afinado y permite distinguir las distintas formas de los verbos, sustantivos o adjetivos, con lo que llegan a unas prestaciones próximas a las de *CE* para la búsqueda gramatical. Por otra parte, la etiquetación de las palabras del *CDH* (también del *CORPES XXI* y de la versión anotada de *CREA*) no es todo lo precisa que le gustaría al investigador para la realización de las búsquedas de tipo gramatical. De este modo, las formas que fuera de un contexto determinado pueden pertenecer a diversas categorías gramaticales, en las que sería especialmente útil la etiquetación precisa de la categoría gramatical, tienen en muchos de los ejemplos del corpus una múltiple lematización, y aparecerán tanto cuando se solicita una categoría gramatical como cuando se solicita otra. En los textos lematizados por el Departamento de Tecnología de la Real Academia Española dentro de lo que la Real Academia

20 Garachana/Artigas (2012) señalan explícitamente la poca adecuación del motor de búsqueda del *CORDE* para el estudio histórico de las perífrasis verbales.

denomina corpus nuclear del *CDH*, la lematización es más precisa, y así, por ejemplo, si le pedimos que busque el lema *son* con categoría gramatical de sustantivo, nos devolverá los casos de *son, sones* sustantivo, y no la tercera persona del plural del presente de indicativo de *ser*. En cambio, los textos que quedan fuera del corpus nuclear, como las extensiones del corpus entre el siglo XII y 1975 y entre 1975 y 2000, con preanotación morfosintáctica, realizada con herramientas de *software* libre en que las formas tienen etiquetación múltiple, devuelven siempre la forma *son*, aunque esté funcionando como verbo. También sucede esto con la forma *a*, que en el corpus nuclear está etiquetada como preposición o como verbo y en la extensión siglo XII-1975 de ambos modos. Esto tiene algunas repercusiones indeseadas en la búsqueda de estructuras gramaticales, como sucede en el caso de las perífrasis a las que antes nos hemos referido. Así, por ejemplo, podemos pedirle al *CDH* que busque el verbo *ir* como lema seguido de la preposición *a* una posición a su derecha y de un verbo en un intervalo de dos posiciones a su derecha; pero dado que la forma *a* está etiquetada como preposición solo en el *CDH nuclear*, pero no en la extensión del corpus entre el siglo XII y 1975, en que la forma *a* viene etiquetada como preposición y verbo (puesto que se tienen en cuenta grafías antiguas del verbo *haber*, y *a* es representación gráfica habitual de la tercera persona del presente de indicativo en español medieval), devuelve correctamente los ejemplos que pertenecen al *CDH nuclear*, pero no los de la extensión, de la que devuelve todos, pues el intervalo de dos posiciones a la derecha incluye también esa primera posición en la que *a* está etiquetado como verbo. Obviamente, si limitamos la búsqueda a los ejemplos del *CDH nuclear*, que es una posibilidad que ofrece el corpus, podemos obtener búsquedas gramaticales más precisas, aunque sobre un corpus menor (algo más de 62 000 000 de palabras), y en consecuencia con un número de ejemplos más reducido. Pero incluso en el *CDH* nuclear, la lematización múltiple es frecuente[21].

Y algo semejante sucede con la búsqueda de *y* como adverbio en textos medievales. El *CDH* nuclear etiqueta separadamente los usos como conjunción y como adverbio, y la extensión no. Esto nos permite, por ejemplo, pedir al *CDH* nuclear que busque los ejemplos de *haber* como lema y el adverbio *y* inmediatamente a la izquierda o a la derecha de la forma verbal, lo que puede resultar útil para estudiar el proceso de creación y extensión de la forma impersonal *hay*. No obstante, incluso el *CDH* nuclear etiqueta como adverbio aquellos casos en que lleva

21 Por ejemplo, si pedimos a *CDH* que busque la forma *so*, incluso en el *CDH* nuclear puede aparecer etiquetada en un texto medieval como verbo (en general la forma verbal aparece como *só* en los textos fichados en *CDH*), como posesivo, como preposición, como adverbio y como interjección.

tilde, *ý*, pero no aquellos en que no la lleva. Al etiquetar exclusivamente como adverbio los casos en que lleva tilde, aunque son la mayoría de los fichados en textos medievales, dejan de aparecer secuencias como «Alli sobre mare Galilee es El Guiacob, e *a y*.i. castiello que assi a nonbre: El Gué Jacob» (*La fazienda de Ultra Mar*, c. 1200, *apud CDH*).

De cualquier manera, la lematización del *CE* para estos casos tampoco mejora los resultados que nos da el *CDH*. En principio, el *CE* también permite precisar la categoría gramatical de una palabra, pero los resultados no son siempre adecuados. Con el propio ejemplo que nos da en su página web el *CE*: «trabajo_n o trabajo.[n*] limitará el resultado a *trabajo* como sustantivo, y trabajo.[v*] o trabajo_v lo limitará a *trabajo* como forma verbal»[22]. Sin embargo, lo cierto es que la lematización no es muy perfecta. Si vemos los resultados que devuelve al pedirle los casos de *trabajo* como forma verbal, podemos observar que muchos ejemplos son de *trabajo* como sustantivo, como podemos ver en la imagen 1 donde aparecen los primeros registros que recoge:

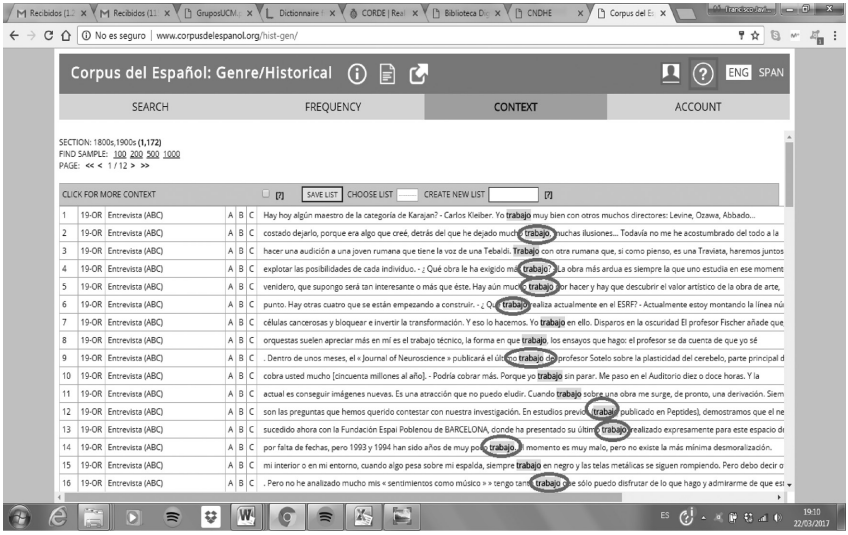


Imagen 1: Primeros ejemplos de la lista del *CE* al pedirle que busque trabajo_v o trabajo.[v*] (formas en que *trabajo* es verbo) en los siglos XIX y XX. Rodeados los casos en que es sustantivo.

22 <http://www.corpusdelespanol.org/hist-gen/9>.

Y por otra parte no deja de sorprender que, aplicado a algunas otras palabras, este recurso no funcione. Por ejemplo, si pedimos que nos dé las distintas categorías gramaticales con que está etiquetado *son*, podemos ver que solo aparece como verbo, tercera persona del plural del presente de indicativo, pero nunca como sustantivo. Por eso precisamente, si pedimos que nos dé todas las formas de *son*, poniendo en la casilla de búsqueda [son] o SON, es decir, cualquier forma de *son* sustantivo, solo devuelve las formas del plural. Si pedimos las distintas categorías gramaticales de la forma *vivo* nos clasifica sus ocurrencias como sustantivo masculino singular, adjetivo masculino singular y nombre propio, pero en ningún caso como verbo. Pero si abrimos la lista de ejemplos que devuelve el *CE* cuando *vivo* está clasificado como adjetivo, inmediatamente comenzamos a ver muchos empleos verbales:

Imagen 2: Primeros ejemplos de la lista del *CE* al pedirle que busque vivo_j o vivo.[j*] (formas en que *vivo* es adjetivo) en los siglos XIX y XX. Rodeados los casos en que es verbo.

Y como consecuencia de esta etiquetación, si pedimos las distintas formas del verbo *vivir*, faltan todas las correspondientes a la primera persona del singular del presente de indicativo.

Y en cuanto al adverbio locativo medieval *y* (*hi, i*) no aparece lematizado como adverbio. *Y* se identifica siempre como conjunción, con lo que no se puede pedir

al *CE* que nos busque [haber] seguido o precedido del adverbio *y*. Tendremos que pedirle la forma [haber] junto a la forma *y* y descartar aquellos casos en que *y* sea conjunción. Por otra parte, la forma *a* no aparece nunca etiquetada como forma del verbo *haber*, con lo que secuencias como *a y* o *y a* debemos pedirlas específicamente, puesto que no las devolvería la petición [haber] o HABER con la forma *y* colocada una posición delante o detrás de la forma verbal.

3.3 Selección de textos y datos bibliográficos

A diferencia de los corpus más pequeños y especializados a los que antes nos hemos referido, con una transcripción muy cuidadosa y fidedigna de los textos utilizados, en corpus como *CORDE*, *CDH* o *CE* los textos no están transcritos a partir de originales ni mucho menos podemos recurrir a la ayuda de un facsímil incluido en el propio corpus para verificar la fidelidad de los textos a los originales. Lógicamente, esto no es posible con un volumen de textos tan amplio, y ha de recurrirse al uso de ediciones previas de muy diverso tipo. Además, cada uno de estos corpus acompaña los ejemplos que facilita con una serie de datos respecto a los textos utilizados, que pueden ser más o menos completos y fidedignos.

Respecto a los datos bibliográficos, los que proporciona el *CE* son más bien escasos. En la mayor parte de los casos se trata de una dirección URL de una página web. Muchos de los textos literarios proceden de la Biblioteca virtual Miguel de Cervantes, aunque las fuentes son muy diversas. En algunas ocasiones la indicación no es muy precisa. Por ejemplo, en las entrevistas procedentes del diario *ABC* se da la dirección electrónica de este diario, pero si vamos a esa dirección se abre la edición actual de *ABC*, no la entrevista concreta que queramos buscar. En otros casos en los que no hay una dirección web la indicación tampoco es precisa, como cuando se da como fuente «Habla culta». A partir del título podemos saber de qué ciudad es el habla culta (Madrid, Lima, Buenos Aires, etc.). Se trata de las publicaciones que recogen muestras del habla culta de diversas ciudades de España e Hispanoamérica, desarrollados dentro del Proyecto de Estudio de la norma lingüística culta de las principales ciudades de Iberoamérica y de la península ibérica, coordinado por Juan M. Lope Blanch, pero en este, como en otros casos, no se da la referencia bibliográfica precisa ni se indican las páginas del texto a que corresponden los ejemplos. Los diversos textos de los que se extraen los ejemplos van acompañados de una datación, necesaria para cualquier tipo de estudio histórico. Pero en muchos falta la indicación de la fecha de los testimonios, especialmente en los más recientes, correspondientes por ejemplo al registro oral o periodístico del siglo XX. Y en otros muchos casos

la datación no es correcta. Algunas veces salta a la vista que hay un error en la fecha: por ejemplo, si al hacer una determinada consulta el corpus nos devuelve un ejemplo del *Quijote* fechado en 1582 (cuando evidentemente ha de ser de 1605 si corresponde a la primera parte y 1615 si corresponde a la segunda)[23], como podemos ver en esta captura de pantalla:

Imagen 3: Inicio del primer capítulo de la primera parte del *Quijote*. Señalado error en la fecha y salto de un fragmento a otro (del capítulo I al XVII).

En ejemplos como este, es fácil darse cuenta del error y subsanarlo, pero obviamente, en otros muchos casos no es así. Por eso hay que tener mucho cuidado con las fechas proporcionadas por el *CE*, que si no son revisadas pueden llevar a interpretaciones erróneas en cuanto a la cronología de la evolución de determinados fenómenos. Por ejemplo, al estudiar el desarrollo de las locuciones de posterioridad inmediata (Herrero en prensa), obtuve en el *CE* una serie de ejemplos de *en cuanto* fechados hacia mediados del siglo XVIII, en 1757. Todos ellos procedían de la *Relación histórica de la vida del Venerable Padre Fray Junípero Serra*, escrita por Francisco Palou. Pero realmente la primera edición de esta obra es de 1787 (México: Imprenta de Don Felipe de Zúñiga), por lo que hay que retrasarlos a finales del XVIII, que es cuando la locución *en cuanto* aparece (o reaparece, si tenemos en cuenta algunos escasos ejemplos medievales) con el valor moderno de posterioridad inmediata.

Vemos además que a veces el texto no está bien transcrito. Así sucede en el ejemplo anterior de Cervantes, en el que se pasa de las primeras líneas del primer capítulo del *Quijote* al capítulo XVII sin ninguna advertencia: tras las

23 Esta mala datación ya había sido advertida por Rojo (2010: 39–40).

bien conocidas palabras del inicio del primer capítulo del Quijote, «En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no» no sigue «ha mucho tiempo que vivía un hidalgo», sino que pasamos a «manera que el molimiento de las estacas fue tortas y pan pintado», que no da sentido alguno, puesto que se salta bruscamente a un fragmento del capítulo XVII. Y esto sucede también en otras ocasiones. Al estudiar la formación del conector consecutivo *de ahí (que)*, y sus variantes *de aquí (que)*, etc. y pedir que nos busque ejemplos de esta última (Herrero 2016b: 577), el *CE* ofrece el siguiente pasaje de Jovellanos:

> Hasta la época que citamos nuestra población fue muy escasa y, digan lo que quieran otros calculistas, la abundancia de pastos, bosques y términos incultos, la falta sean mal pagados, la decoración ridícula y mal servida, el vestuario impropio e indecente, el alumbrado escaso, la música miserable y el baile pésimo o nada. De aquí que los poetas, los artistas, los compositores que trabajan para la escena sean ruinmente recompensados y, por lo mismo, que solamente se vean en ella las heces del ingenio. (Jovellanos, *Espectáculos y diversiones públicas*, 1796, *apud CE*).

Donde obviamente hay una incongruencia debida a la mezcla de dos pasajes: hasta «la falta» se trata de un pasaje, que corresponde al epígrafe «caza» de la primera parte de la obra; y desde «sean mal pagados» se trata de otro, que corresponde al epígrafe «arbitrios para costear esta reforma» de la segunda parte.

El *CORDE* sí fecha generalmente bien las obras citadas, independientemente de que tome partido por una u otra propuesta cuando se trata de obras cuya fecha de composición es discutida. Así, fecha el *Cantar de mio Cid* en 1140, siguiendo las tesis pidalianas, y la *Fazienda de Ultramar* c. 1200, fecha muy posterior al intervalo 1126–1142, en que data esta obra su editor Moshé Lazar (Almerich 1965: 12), pero anterior a la que sugiere Rafael Lapesa (1981: 233–234), ya cumplido el primer tercio del siglo XIII. El *CDH* ofrece dos posibilidades de ordenación: por la fecha de composición, o por la fecha del manuscrito o edición utilizada, pero solo para los textos incluidos en el corpus nuclear. Así, por ejemplo, para el *Cantar de mio Cid* ofrece la fecha de 1140 para la redacción, y siglo XIV —la datación más generalmente admitida— como fecha del manuscrito conservado. Esta doble datación es importante, sobre todo en el caso de textos medievales conservados en copias tardías, porque muchas veces pueden explicar por qué encontramos hechos lingüísticos que parecen más modernos de lo que la fecha de composición invitaría a suponer[24]. No obstante, ocasionalmente hay también algún error en la fecha atribuida a algunas obras. Por ejemplo, *La niña*

24 A este problema en el *CORDE* se refieren Nieuwenhuijsen (2009) y Garachana/Artigas (2012: 51–57). Queda en parte solucionado en el *CDH*.

de los embustes, Teresa de Manzanares de Castillo Solórzano, de 1632, aparece fechada en el *CORDE* en 1692, con más de medio siglo de desfase. Sí aparece la fecha correcta en el *CDH*. Sin embargo, en algunos casos hay errores en el *CDH* que no sé bien a qué causa obedecen, como sucede cuando los ejemplos correspondientes a un texto aparecen fichados como correspondientes a otro texto, que tiene lógicamente una cronología distinta. Así, algunos ejemplos de la obra de Pemán *Mis almuerzos con gente importante*, del siglo XX, nos aparecen en la lista de ejemplos fechada en 1578. Al abrir el ejemplo, sin embargo, aparecen datos bibliográficos correspondientes a San Juan de la Cruz, e incrustados en ellos, los de Pemán, que son los que realmente le corresponden, de la siguiente manera: «1578–84 SAN JUAN DE LA CRUZ, *fragmento* (1970 Pemán, José María, *Mis almuerzos con gente importante*) [España] [Barcelona, Dopesa, 1970] Astrología y ciencias ocultas»[25].

Otro dato interesante que ofrecen los corpus académicos *CORDE, CREA, CDH* y *CORPES XXI* es el de la procedencia de los ejemplos. Además de permitir su interfaz la selección por países —*CDH, CORPES XXI* y la versión anotada de *CREA* también por grandes áreas geográficas de América— a la hora de pedir que recupere las palabras o secuencias de estas, los ejemplos recuperados siempre dan indicación del país de procedencia. Esto es muy útil cuando se estudia la evolución y uso de cualquier elemento, léxico o gramatical, pues permite ver su existencia, grado de difusión e intensidad de empleo en distintas áreas. Así, al estudiar la gramaticalización de las secuencias formadas por *según* + interrogativo, especialmente *según qué*, podemos ver que se trata de un fenómeno claramente español, y especialmente característico del centro y norte de la península (Herrero 2015a). Del mismo modo, podemos advertir también que es un uso característicamente español el empleo de la secuencia ponderativa *que pa(ra) qué*, que funciona como adjetivo o adverbio incidiendo sobre un sustantivo o un verbo («me dio un susto que para qué») (Herrero 2016d: 358–360). Y la distribución geográfica de los ejemplos nos permite analizar qué locuciones de posterioridad inmediata se usan o son más frecuentes en las distintas áreas hispanohablantes (Herrero en prensa), o la existencia de algunas áreas hispanoamericanas (Paraguay y Bolivia) donde el conector consecutivo *de ahí que* se construye con frecuencia con el verbo en indicativo. En algunas ocasiones, no obstante, hay que tener cierto cuidado al interpretar los datos geográficos de

25 Los ejemplos de esta obra de Pemán sí aparecen correctamente fechados y referidos a la edición de 1970 de *Mis almuerzos con gente importante* en *CORDE*.

procedencia indicados en estos corpus. Generalmente no porque la procedencia de un determinado autor del que se incluye un texto en alguno de estos corpus esté mal indicada[26], sino porque algunos de ellos residen o han residido mucho tiempo fuera de su país de origen, y adquieren rasgos de otra área donde han pasado gran parte de su vida. Así no nos puede extrañar, a pesar de que en el *CORPES XXI* se dé como indicación geográfica Uruguay, que un autor como Gervasio Posadas utilice la expresión característicamente española *que para qué*, puesto que este autor de origen uruguayo lleva muchos años residiendo en Madrid, o que use *según qué* como indefinido un argentino como Andrés Ehrenhaus, en cuyos textos el *CORPES XXI* da como indicación geográfica Argentina, puesto que reside en Barcelona, y este es de hecho un empleo muy frecuente en Cataluña.

3.4 Problemas en la interpretación y transcripción de los ejemplos

En algunas ocasiones surgen problemas respecto a la fiabilidad de los datos que pueden ser debidas a diversos factores que señalamos a continuación.

3.4.1 Digitalización incorrecta de una palabra o pasaje

A veces estos problemas pueden estar relacionados con la incorrecta transcripción de los textos o con las características de las ediciones utilizadas. Respecto a las transcripciones, podemos encontrarnos con que se ha digitalizado incorrectamente alguna palabra o pasaje de la edición que ha servido de base para *CORDE* y *CDH*. Por ejemplo, muchos de los casos que devuelve el *CORDE* en textos del XVIII en que *incluso* aparece formalmente inmovilizado en la forma del masculino singular, como corresponde al marcador escalar de foco (uso que se consolida en el XIX), corresponden a malas transcripciones. Así, entre otros casos, el *CORDE* recoge el siguiente texto de Forner: «Toda religión (*incluso* la Cristiana) es invención política» (Juan Pablo Forner, *Discursos filosóficos sobre el hombre*, 1787, Biblioteca Virtual Miguel de Cervantes, Universidad de Alicante (Alicante), 2002, *apud CORDE*), pero en la edición de la Biblioteca Virtual Miguel de Cervantes, y en el texto original[27] se lee «Toda religion (*inclusa* la Christiana) es invencion política» (Herrero 2014a: 196 n. 17). Desde el siglo XVI, encontramos *al igual de* + SN en construcciones parentéticas modales, y desde la

26 Hay algún caso en que la clasificación del país no está bien efectuada, pero son pocos. Por ejemplo, el texto «Peregrinaje en tierras del poeta» de Roberto Bennett está catalogado en *CORPES XXI* como «Venezuela» cuando realmente Bennett es uruguayo.

27 Facsímil accesible en línea: <http://adrastea.ugr.es/tmp/_webpac2_1106239.95851>.

segunda mitad del XVIII aparece también la variante *al igual que* seguida de SN o de una oración completa. En el *CORDE* aparece un ejemplo del siglo XVII de *al igual que* en un texto de Juan de Piña, *Epítome de las fábulas de la antigüedad*, que ficha a partir de la primera edición de 1635: «que viessen los nauegantes en el mar de la voluntad que las pechugas dirían con los pies y no con las alas de las Sirenas, sino *al igual que* el aue indiana, que de veinte y quatro pies cada pluma buela por los ayres en ins valientes vñas el mayor elefante hasta el más encumbrado monte», pero al cotejarlo con el facsímil del original[28] verificamos que se trata de una errata por *al igual de* (Herrero 2018: 278, n. 19).

3.4.2 Erratas en las ediciones digitalizadas

Pero en otras ocasiones la errata o la mala transcripción está en la edición que ha servido de base a la digitalización del *CORDE*. Y no siempre es fácil darse cuenta de ello. Así, por ejemplo, la forma *pronto* es en principio un cultismo introducido como adjetivo, y que no comenzará a tener usos adverbiales hasta finales del siglo XVIII (Herrero 2016c: 111); por eso llama la atención que el *CORDE* proporcione dos ejemplos del uso adverbial de *pronto* del siglo XV, uno de la *Crónica incompleta de los Reyes Católicos* (1469–1476) y otro del *Libro de Acuerdos del Concejo Madrileño* (1498–1501). Y efectivamente los dos ejemplos están en las ediciones fichadas por este corpus. Sin embargo, no corresponden a lo escrito en los originales del XV. En ambos casos se halla escrito en ellos *presto*, que sí tenía entonces con frecuencia uso adverbial, abreviado como *psto*, con un trazo consistente en una pequeña línea sobre la *p*, en el caso de la *Crónica incompleta de los Reyes Católicos*, y sobre la *p* y la *s* en el caso del *Libro de Acuerdos del Concejo Madrileño* (Herrero 2016c: 108 n. 12 y 13). La semejanza formal entre el comienzo y el final de *pr[es]to* y *pr[on]to* y el uso actual de *pronto* (y no de *presto*) como adverbio de tiempo explican estos errores en la transcripción por parte de los editores modernos[29].

---

28 <http://alfama.sim.ucm.es/dioscorides/consulta_libro.asp?ref=B19777863&idioma=0>.

29 El aceptar los ejemplos tempranos de *pronto* del *CORDE* sin comprobar su autenticidad, lleva a Azofra (2014: 397) a afirmar que «ya entrado el siglo XV aparece el adverbio [*pronto*] con el significado actual, aunque de forma muy escasa», lo que, como hemos visto, no es cierto. De hecho, dentro del corpus que ella misma examina no lo encuentra hasta finales del XVIII, en *La comedia Nueva* de Moratín, lo que, concuerda con la época en que en verdad comienza a tener este uso. También Espinosa (2010: 103) acepta el ejemplo del *Libro de los acuerdos del Concejo madrileño* como una de las primeras manifestaciones conservadas del adverbio *pronto*, pero ya sabemos que el ejemplo no corresponde a la lectura del original.

Entre los ejemplos de *pronto* en el siglo XVII el *CORDE* recoge el siguiente de Castillo Solórzano:

> Sírvase vuestra merced de hacer presentación a su señora destos servicios, para que *pronto*, en su tribunal, alcancen el premio que merecen (Alonso de Castillo Solórzano, *La niña de los embustes, Teresa de Manzanares*, 1632[30], *apud CORDE*),

que efectivamente figura así en la edición fichada de Antonio Rey Hazas[31], pero, consultado el original en la edición digital de la Biblioteca Nacional[32], en este pasaje se lee:

> Siruase v.m. de hazer presentacion à su señora destos seruicios, para que *puestos* en su tribunal alcancen el premio que merecen,

donde no hay ningún adverbio de tiempo (Herrero 2016c: 109).

En la *Historia del Nuevo Mundo* de Bernabé Cobo (1653) aparecen cuatro ejemplos del adverbio *pronto*, que efectivamente se hallan en la edición de Marcos Jiménez de la Espada (Sevilla, Sociedad de Bibliófilos Andaluces, 1890–1893), fichada en el corpus académico. Sin embargo, ninguno de ellos figura en la primera edición de 1653, digitalizada por la universidad de Sevilla[33], en la que figura *presto* o *de presto* (Herrero 2016c: 109–110).

También en el caso de la gramaticalización de la secuencia *desde luego* como confirmativo y marcador de evidencia, observa González (2012: 97 n. 9) un uso sorprendentemente temprano registrado en el *CORDE*, ya en el siglo XIII:

> Todas estas cosas dessuso dichas uos damos, que las ayades del dia que esta carta fue fecha fasta veynte anos conplidos. Et porque nos dades uos todos los heredamientos que comprastes de Pero Yuanes Baraganna e de donna Maria, sso muger, en Sancta Maria de Valdecespedes e en so termino. Et por dos mille moruedis de los dineros blancos de la primera querra, que nos *desde luego* adelantradamientre, para prouesion del dicho monesterio, a tiempo que los auemos mucho mester *(Documentos del Monasterio de Santa María de Trianos*, 1299, ed. de Guillermo Castán Lanaspa y Javier Castán Lanaspa, Universidad de Salamanca, 1992, *apud CORDE*).

Espinosa (2010: 147) acepta esta fecha como primer testimonio del uso de esta locución aún con el valor de complemento temporal, 'desde ahora', desde el que

30 Corregimos la datación. Como ya dijimos, el *CORDE* fecha erróneamente el texto en 1692.

31 Alonso de Castillo Solórzano (1632 [1986]), *La niña de los embustes, Teresa de Manzanares*, en *Picaresca femenina*, Antonio Rey Hazas (ed.), Barcelona: Plaza y Janés, 246.

32 <http://bdh-rd.bne.es/viewer.vm?id=0000078414&page=1>.

33 <http://fondosdigitales.us.es/fondos/libros/2423/11/historia-del-nuevo-mundo-por-el-padre-bernabe-cobo-de-la-compania-de-jesus/>.

se deslizará después a locución adverbial de afirmación y marcador de evidencia; sin embargo, como señala González (2012: 97 n. 9), a pesar de que esa es efectivamente la lectura que encontramos en la edición de Guillermo y Javier Castán Lanaspa de la documentación de Santa María de Trianos, y aunque no hemos accedido al documento original, podemos estar casi seguros de que es una errata, y de que en el texto original habría una forma del verbo *dar* en segunda persona del plural. Fijémonos en lo fácil que resultaría una alteración en el orden de las grafías de *dedes*, presente de subjuntivo del verbo *dar*, en *desde,* más teniendo en cuenta que la terminación *-des* de la segunda persona del plural ya no existe en la lengua moderna, lo que pudo facilitar una fácil confusión del editor o impresor. La razón más evidente para suponer que se trata de una errata es que *desde luego* no proporciona una lectura con sentido en el contexto en que aparece, mientras que *dedes* sí lo hace. Además, en el contexto inmediato, en el que se habla de donaciones al monasterio, aparece en varias ocasiones el verbo *dar*. Por último, podemos añadir que el propio Guillermo Castán Lanaspa, en otro trabajo suyo, cita explícitamente este texto y este pasaje y transcribe *dades luego*, no *desde luego*[34].

3.4.3 Uso de ediciones modernizadas

En algunas ocasiones, el *CORDE* usa ediciones que modernizan el texto original. En estos casos, podemos encontrarnos con el empleo de términos o construcciones inexistentes en la época en que se escribió el texto, pero que sí tienen uso en la época del editor moderno, que a veces modifica el texto sustituyendo lo que le parece anticuado, e introduce formas que le parecen más familiares para un lector moderno. Por ejemplo, al buscar en *CORDE* materiales para fechar el proceso de gramaticalización de *sin embargo*, primero como locución prepositiva y luego como conector contraargumentativo, uso este último al que parece llegar a finales del XVII, encontramos diversos ejemplos fechados en la primera mitad del XVI. En algunos casos, esto se debe simplemente a que se está utilizando una edición modernizada, que no refleja fielmente el texto del original. Así, vemos un ejemplo de *sin embargo* conector contraargumentativo, que el *CORDE* (y *CDH*, que sigue las mismas ediciones) fecha en 1528 en la traducción hecha por Juan

34 «*Tríanos,* carp. 984-13, de 1299: *…dos mille moruedis…que nos dades luego adelantradamientre, para prouesion del dicho monesterio a tiempo que lo auemos mucho mester*» (Castán Lanaspa 1983: 78 n. 62). La *consecutio temporum* pediría mejor *dedes* que *dades*, al estar en una oración subordinada sustantiva que depende del verbo *querer* en futuro. No obstante esto solo lo aclarará la consulta directa del documento.

Justiniano de la obra de Juan Luis Vives, *Instrucción de la mujer christiana*, y que el corpus, académico da de la siguiente manera:

> Crísipo, filósofo acutísimo, mandaba que las amas de leche fuesen cuerdas y buenas, lo cual nosotros debemos seguir y avisarlo a las madres que no podrán o no les será así lícito criar a sus hijos con su propia leche; en lo cual, *sin embargo*, no es tan necesario que se ponga tanta diligencia en los hijos como en las hijas.

Pero realmente en el texto de la traducción de Justiniano no aparece *sin embargo*, sino *pero*, utilizado como conector contraargumentativo, lo que aún era posible en el español clásico, y encontramos con alguna frecuencia en autores en los que, como Justiniano o Garcilaso, se da un influjo italianizante:

> Cryſippo philosopho acutiſſimo mandaua las amas de leche fueſſen cuerdas/y buenas: lo qual nosotros deuemos ſeguir/y auisallo alas madres que non podran o noles ſera aſſi licito criar a ſus hijos cõſu propia leche: enloqual *pero* no es tan neceſſario que ſe ponga tanta diligẽcia enlos hijos/como enlas hijas (Juan Luis Vives, *Libro llamado Instrucción de la muger christiana […] traduzido ahora nuevamente d[el] latin en romance por Juan Justiniano,* Valencia: Jorge Costilla, 1528, fol i, v°, col. a)[35].

La Academia ficha una edición de Elisabeth Teresa Howe que, como ella misma dice en la introducción, «es una revisión modernizada», y por lo tanto no es válida para ejemplificar la lengua de la época (Herrero 2015b: 864–865).

3.4.4 Ediciones con doble versión, literal y modernizada

Algunas ediciones de textos antiguos contienen dos versiones: una que mantiene fielmente el texto original, y otra que moderniza el texto para el lector no especializado. Lógicamente, los corpus pensados para el estudio histórico de la lengua, como *CORDE* y *CDH*, deberían incluir solamente el texto que reproduce la transcripción del original, y no la versión modernizada. Sorprendentemente, podemos encontrarnos con el caso contrario: *CORDE* y *CDH* recogen solo la versión modernizada de María Nieves Sánchez del *Regimiento contra la peste* de Fernando Álvarez, que fechan c. 1501, o la del *Tratado nuevo* de Diego Álvarez Chanca, de 1506, y no la transcripción del original. Consecuencia de ello es que encontramos una muestra de lengua del siglo XX etiquetada como si fuese de principios del XVI. Por dar algunos ejemplos de usos que no corresponden a la época, y naturalmente no están en los originales, hallamos en el *Regimiento contra la peste* el uso de *sin embargo* como conector contraargumentativo o de *a*

35 He consultado para esta referencia el ejemplar de la Biblioteca Nacional, con signatura R/1289.

pesar de que (locución que comienza a tener uso en el XVIII), dados como ejemplos de lengua del XVI, donde en el texto original, como puede verse en la transcripción que aparece en la misma edición, se utiliza *empero* y *no embargante* respectivamente (Taranta *et al*. 1993: 163). Y en ambos vemos el empleo de la locución temporal *de inmediato*, que comienza a documentarse muy débilmente en el XVII, donde en el texto original aparecía *muy prestamente* en el caso de Álvarez y *luego* en el caso de Chanca (Taranta *et al*. 1993: 164 y 199).

3.4.5 Intervenciones del editor moderno en el texto

En otros casos, aunque la edición no modernice generalmente la lengua del autor, hay adiciones o modificaciones del editor en algunos pasajes, por lo que esos pasajes no son tampoco representativos de la lengua de la época del original, a la que aparecen adscritas en *CORDE*. Por ejemplo, al estudiar la formación del adverbio de inmediatez *enseguida* (Herrero 2017), el *CORDE* da un ejemplo de este adverbio en la *Traducción de «Orlando furioso» de Ludovico Ariosto* de Jerónimo de Urrea, de 1549, siguiendo la edición de Francisco José de Alcántara. El ejemplo viene dado en un fragmento que está entre corchetes:

> al que de algún aprieto la libraba,/entre rosados labios se lo encierra:/[*en seguida*, invisible, se ha perdido,/dejando a todos locos, sin sentido]./Y así como le vino de primero/gana de ir con Roldán o Sacripante/para volver al reino verdadero/de Galafrón, al último Levante,/así ahora con desdén muy por entero/mudó la voluntad (Jerónimo de Urrea, *Traducción de «Orlando furioso» de Ludovico Ariosto*, Francisco José Alcántara, Planeta (Barcelona), 1988, *apud CORDE*).

Aunque en el *CORDE* no se avisa de ello, los fragmentos que en el texto fichado por la Academia aparecen entre corchetes corresponden a los que aparecen en cursiva en la edición de Alcántara. Y en el prólogo de esta edición, se nos dice que «mi intento ha sido —en los límites de lo posible y respetando el texto del traductor— acercarlo al original italiano cuando Urrea se aleja innecesariamente: a veces simples frases; a veces versos y aun estrofas enteras. En estos casos, mi aportación personal va en cursiva». En consecuencia, este ejemplo corresponde a un uso de *en seguida* en la lengua del siglo XX, no del XVI. Algo semejante encontramos en los ejemplos que proceden de las *Leyendas de Alejandro Magno*, texto aljamiado de la segunda mitad del XVI, que el *CORDE* ficha siguiendo la edición de Guillén Robles, en la que a veces se modernizan elementos del texto original. Vemos así que aparecen en él seis casos de *enseguida*, pero ninguno corresponde al original del XVI, sino a una modernización del editor del siglo XIX.

3.4.6 Encabezamiento y titulillos puestos por un editor posterior

Relacionado con el caso anterior está el de aquellos textos en que el *CORDE* digitaliza una edición que respeta el texto original, pero introduce entradillas de diverso tipo, como encabezamientos de capítulos, apartados o informaciones sobre el contenido de documentos editados, que aparecen también reproducidos sin indicación de que se trata de elementos añadidos en una cronología distinta. Por ejemplo, el *CORDE* recoge un ejemplo de *a pesar de que* como conector contraargumentativo en un texto fechado en el XV, cuando este conector no se utiliza hasta finales del XVIII, pero realmente aparece en el encabezamiento, correspondiente al editor moderno que da entrada a uno de los documentos reproducidos:

> Los Reyes Católicos mandan a Abrahán Seneor, juez mayor de las aljamas, que se reúna con un hombre sabio, tal como manda la ley de los judíos, para revisar la sentencia por la que determinaba que Jacob Leví, vecino de Madrigal, debía pechar con los judíos de Olmedo, al considerar que allí había vivido y todavía tenía ciertos bienes, *a pesar de que* contribuía con los judíos de Madrigal (*Documentación medieval abulense en el Registro General del Sello*, 1485–1488, ed. de Gregorio del Ser Quijano, Ávila, Institución Gran Duque de Alba/Obra Cultural de la Caja de Ahorros de Ávila, 1995, *apud CORDE*).

En la *Historia de Las Indias* de Fray Bartolomé de las Casas (c. 1527–1561), los editores del texto fichado[36] en *CORDE* dicen que «Se han redactado los sumarios de los capítulos que el padre Las Casas dejó en blanco, procurando seguir la pauta marcada por los que él redactó […]. Estos sumarios complementarios introducidos en la presente edición van entre corchetes». Pero esta aclaración no está disponible para el usuario del *CORDE*. Solo en estos sumarios, no escritos por el propio Las Casas, volvemos a encontrar en un caso *enseguida*:

> […] Cómo los vasallos que tenía por allí Moctezuma, rey de la ciudad de México, le informaron[37] *enseguida* de cuántos eran los españoles, etc.] (Fray Bartolomé de las Casas, *Historia de las Indias*, c. 1527–1561, *apud CORDE*),

y se trata de nuevo de un ejemplo de lengua del siglo XX, no del XVI como se desprendería de la indicación del *CORDE*.

---

36 Los editores de esta obra son Miguel Ángel Medina, Jesús Ángel Barreda e Isacio Pérez Fernández, quienes editan la *Historia de las Indias* como volúmenes 3–5 de las *Obras completas* de Fray Bartolomé de las Casas (Madrid, Alianza), no Paulino Castañeda Delgado, como indica el *CORDE*, quien es editor del volumen 2 de dichas *Obras completas, De unico vocationis modo*.

37 El *CORDE* transcribe *infamaron*, evidente errata corregida a la vista de la edición citada en el corpus académico.

3.4.7 Ediciones posteriores a la fecha de composición del texto

En muchos casos, un texto se ha conservado en copia o edición posterior a la fecha de la composición original del texto. A esta cuestión nos hemos referido anteriormente, es algo bastante habitual en los textos medievales y que, como vimos, en parte soluciona el corpus nuclear del *CDH*. Es bien sabido que los manuscritos que han conservado obras como el *Cantar de mio Cid,* las obras de Berceo*, Libro de Apolonio, Alexandre, Libro del caballero Zifar*, etc. son posteriores a su fecha de composición. Y en cualquier caso, constituye una precaución elemental en textos medievales indagar la fecha de la edición o manuscritos que nos han transmitido los textos. Pero en otros casos, con obras más tardías y textos menos conocidos, los corpus académicos proporcionan sin ninguna otra indicación una fecha, la que se cree de composición de la obra, pero la obra ha sido transmitida por una copia o edición muy posterior, y a ello se pueden deber algunas o muchas de las características lingüísticas del texto, tal y como nos ha llegado.

Al estudiar el empleo de *lo mismo que* como locución modal (Herrero 2016a), he observado que este empleo se documenta con una muy baja frecuencia en el siglo XVII en el *CORDE*. Uno de los ejemplos que da procede del *Libro de las Medicinas Caseras*, que *CORDE* y *CDH* fechan en 1611–1650. Sin embargo, el original de este libro no se ha conservado, y sus editores modernos editan una copia tardía de c. 1870, que probablemente modifica el original. De hecho, advierten (Guerra y Sánchez 1984: XXXV–XXXVI) lo siguiente:

> el amanuense realizó la transcripción con los errores propios de un indígena filipino de limitada educación gramatical, e introdujo glosas y materiales ajenos al texto primitivo de Fr. Blas de la Madre de Dios […]. En un caso, la calamba, se hace mención de la quina, dato incompatible con la fecha de 1611 en que Fr. Blas de la madre de Dios escribió el *Libro de las Medicinas Caseras*, pues las virtudes curativas de la quina en las tercianas palúdicas se descubrieron en Perú hacia 1630 y el uso de esta corteza solo se hizo universal hasta medio siglo después. Por ello, esta copia del *Libro de las Medicinas Caseras* plantea el dilema heurístico de haber adulterado el texto de Fr. Blas de la Madre de Dios o tratarse de un manuscrito apócrifo.

Por tanto, no es claro que este texto pueda tenerse en cuenta como representativo de la lengua del XVII.

La locución *enseguida* se consolida como adverbio de inmediatez en el siglo XIX. Ya a finales del siglo XVIII hay algunos ejemplos en los que, aunque sigue siendo posible entenderla de acuerdo con su valor primitivo de 'a continuación', podría entenderse también con el valor de 'inmediatamente'. Hay no obstante en el *CORDE* un ejemplo de mediados del XVIII en que *en seguida* presenta claramente el valor de 'inmediatamente':

> y habiendo yo salido con una armadilla de embarcaciones que junté para ahuyentarlos, se fugaron *en seguida* (Juan José Delgado, *Historia general sacro-profana, política y natural de las islas del Poniente llamadas Filipinas*, 1754, Juan Atayde, Imp. de El Eco de Filipinas (Manila), 1892, *apud CORDE*),

pero es dudoso porque el texto de Delgado, aunque fue escrito a mediados del XVIII, no se publicó hasta 1892 y su editor comenta (p. XIV) que la edición no está realizada sobre el original, sino sobre una copia «mal sacada», en la que hubo que hacer correcciones, suplir vacíos e interpretar frases ininteligibles. Por tanto, al menos parte del texto está reelaborado a finales del XIX, por lo que no siempre puede tenerse como texto representativo de mediados del XVIII.

3.4.8 Traducciones

Es evidente que las traducciones son textos que han de ser tenidos en cuenta en un corpus que tenga un enfoque histórico. Algunas de ellas son de hecho piezas bien conocidas de la literatura y la cultura españolas. De este modo, está plenamente justificada la inclusión dentro de los corpus académicos de obras como las traducciones bíblicas medievales, la traducción hecha por Boscán de *El Cortesano* de Castiglione, o las traducciones de textos bíblicos y clásicos de Fray Luis, entre otras. Sin embargo, lo que no es admisible es que aparezcan en el *CORDE* y *CDH* traducciones modernas de diversos textos que se presentan con la fecha de la obra original y sin ninguna indicación de que se trate de una traducción. A todas luces, lo que sucede es que se ha tomado la traducción por el texto original, sin darse cuenta de que este estaba en otra lengua. Podemos citar algunos de estos textos catalogados como escritos españoles del siglo XVI. Al estudiar diversos procesos de gramaticalización y lexicalización encontramos con frecuencia documentaciones tempranas en el *Viaje a la Tercera hecha por el comendador del Chaste* (1583). Vemos, por ejemplo, el primer registro de uso de *lo mismo que* con valor modal 'como' («Pedro desembarcó fácilmente, pues los portugueses, portándose *lo mismo que* en la Tercera, huyeron á los montes»), no hay ningún otro en el siglo XVI (Herrero 2016a: 340 n. 25); también encontramos casos del empleo de *en seguida* como adverbio temporal de inmediatez, uso que comienza a darse a finales del XVIII y principios del XIX (Herrero 2017), y dos ejemplos del uso de *sin embargo* (Herrero 2015b: 865), que comienza a asentarse en la transición del XVI al XVII; igualmente encontramos en el *Diario de Erich Lassota von Steblau*, fechado en 1583, tres ejemplos de *sin embargo* y otros dos de *en seguida* que no corresponden a usos de la época. El propio Cesáreo

Fernández Duro, editor de estas obras[38] nos informa de que el texto de Lassota von Steblau, escrito en alemán, permaneció inédito hasta 1866, y lo que publica es una traducción hecha a finales del siglo XIX (firmada por F. R.). El *Viaje a la Tercera hecha por el comendador del Chaste*, corresponde a un original escrito en francés, como advierte también Cesáreo Fernández, traducido también a finales del XIX. Estos textos, por tanto, no pueden utilizarse como ejemplos de la lengua de la segunda mitad del XVI, en todo caso podrían presentarse como traducciones del XIX, aunque lo mejor sería eliminarlos del corpus. Podemos agregar otro caso semejante, correspondiente a la *Carta de Antonio Brito al Rey de Portugal sobre algunos sucesos en la India*, de 1523. Al estudiar el proceso de adverbialización del adjetivo *pronto* (Herrero 2016c) y del sintagma *en seguida* (Herrero 2017) encontramos ejemplos del uso de ambos como adverbios en la carta mencionada, fechada en 1523, muy anteriores a la fijación de estas unidades como adverbios de tiempo. Pero lo cierto es que no tiene ningún sentido que el capitán portugués Antonio de Brito escriba al rey Juan III de Portugal una carta en español, conservada en el Archivo de la Torre do Tombo de Lisboa. Y en efecto, lo que sucede es que el *CORDE* ficha el extracto de dicha carta hecho por el historiador Juan Bautista Muñoz (1745–1799) y publicado por Martín Fernández de Navarrete en la *Colección de los viajes y descubrimientos que hicieron los españoles desde fines del siglo XV*[39]. La lengua de esta traducción es por lo tanto la de la segunda mitad del XVIII. La carta original en portugués está publicada íntegra en las páginas 464–476 del volumen titulado *Alguns documentos do Archivo Nacional da Torre do Tombo ácerca das navegações e conquistas portuguezas publicadas por ordem do governo de Sua Magestade Fidelissima ao celebrar-se a commemoração quadricentenaria do Descobrimento da America* (Lisboa: Imprensa Nacional, 1892)[40]. Una consecuencia de estos errores, es que pueden pasar a la redacción del *Nuevo Diccionario Histórico del Español*, que en definitiva toma el *CDH* como fuente de datos y ejemplos. Entre los nombres de diversas armas que incluye, figura el probable portuguesismo *verso* como 'pieza de artillería' y da como primera documentación castellana de esta voz la carta de Brito: «Se documenta por primera vez, con la acepción de 'arma de artillería cuyo tamaño y calibre corresponde a la mitad de la culebrina', en 1523, en la

38 *La conquista de las Azores en 1583 descrita por el capitán de navío Cesáreo Fernández Duro* (Madrid, Sucesores de Rivadeneyra, 1886).

39 Tomo IV, Madrid, Imprenta Nacional, 1837: 305–311.

40 Puede consultarse en la BDH el facsímil digitalizado por la Biblioteca Nacional (<bdh-rd.bne.es/viewer.vm?id=0000006089&page=1>); y también el facsímil digitalizado por la Biblioteca Nacional de Portugal (<purl.pt/26221/1/index.html#/1/html>).

Carta de Antonio Brito al Rey de Portugal sobre algunos sucesos de la India»[41]. Pero, como acabamos de ver, esta documentación no es válida, pues realmente el texto castellano es traducción del portugués hecha en el siglo XVIII.

También en el *CE* nos encontramos con el problema de que aparecen fichadas traducciones tardías con la fecha del texto original, como si este estuviera escrito en castellano. Limitándonos a un ejemplo, podemos señalar que el *CE* recoge el marcador consecutivo *de aquí* + SN en la *Predicación del evangelio en las Indias* (1590) de José de Acosta; también aparece *de aquí que* en las *Antigüedades de la Nueva España* de Francisco Hernández, y asimismo recoge *de aquí que* y *de ahí que* en el siglo XVI en el epistolario de Juan Ginés de Sepúlveda, y diversos ejemplos de *de ahí que* en el *Tratado sobre el préstamo y la usura* de Luis de Molina, todos ellos textos del XVI. Sin embargo, en todos los casos, el original del XVI está escrito en latín, y lo que se fichan son traducciones del siglo XX (Herrero 2016b: 574).

3.4.9 Fragmentos escritos en otras lenguas

Relacionado con el apartado anterior está el de aquellas obras que están en su mayor parte redactadas en español, pero incluyen algunos pasajes o fragmentos en otras lenguas. En algunas ediciones modernas, estos pasajes o fragmentos han sido traducidos al castellano. Y el *CORDE* y el *CDH* han fichado estas ediciones sin advertir que la lengua de esa parte del texto es moderna. En más de una ocasión, construcciones que se han gramaticalizado como conjunciones o marcadores discursivos, se recogen con una datación temprana precisamente en estos textos, pero no corresponden a la lengua del XVI o del XVII, sino a la época de la traducción posterior. El *CE* recoge un ejemplo del conector consecutivo *de aquí que* en la *Libra astronómica y filosófica* y uno de *de aquí que* y otro de *de ahí que* en el *Teatro de virtudes políticas que constituyen a un príncipe*, ambas obras de Sigüenza y Góngora de finales del XVII. Pero en ambos casos corresponden a traducciones modernas de citas que Sigüenza y Góngora daba en latín[42]. También el *CORDE* recoge dos usos modernos de *de ahí que* en el *Tratado sobre los indios que han sido hechos esclavos* (1552) de Fray Bartolomé de las Casas, pero

41 <web.frl.es/DH/org/login/Inicio.view>, s. v. *verso*.

42 Puede comprobarse con la consulta de las primeras ediciones de ambas obras que están digitalizadas en línea: <http://www.cervantesvirtual.com/obra-visor/teatro-de-virtudes-politicas-que-constituyen-a-un-principe-advertidas-en-los-monarcas-antiguos-del-0/html/b9749496-625a-41c6-a948-0f747ca27c03_78.html, http://books.google.com.au/books?hl=es&id=98-I7iboOp0C&q=211.#v=onepage&q=211.&f=false>.

los dos se encuentran en los corolarios del tratado, escritos, a diferencia del resto del texto, en latín. Los ejemplos de *de ahí que* corresponden también a una traducción moderna[43].

3.5 Uso de los resultados de los corpus como indicios para la comprobación de ejemplos de los mismos corpus o de otras fuentes

La búsqueda de un término o de una construcción en corpus históricos de carácter general, como *CORDE* o *CE*, produce una serie de resultados que muchas veces permite ver un camino en la introducción del término, o el proceso de gramaticalización de determinados elementos o secuencias. De este modo, cuando un elemento previamente no se usaba en español, o un elemento o construcción no se usaban en determinados contextos y con determinados valores, es normal que comencemos a ver primero un uso escaso de ellos en la época primitiva en que aparecen o alcanzan un determinado valor, y que después se incremente su uso. Por eso mismo, cuando encontramos una documentación aislada, o un número reducido de ellas en un determinado período, o cuando son escasos los empleos que corresponderían al período más antiguo documentado de su uso, siendo conscientes de los errores que a veces se dan en el fichado en estos corpus, debemos revisarlas para asegurarnos de la exactitud de estas documentaciones. Esto en unos casos llevará a la verificación de los registros, en otros a descartarlos por tratarse de muestras erróneas. Así, por ejemplo, podemos constatar que son correctos los testimonios que recoge el *CORDE*, procedentes de la *General Estoria* alfonsí, de la correlación temporal *apenas…cuando*, e incluso de *apenas* gramaticalizado ya como conjunción sin estar en correlación con *cuando*, lo que antedata en varios siglos la fecha que había venido dándose de la aparición de estos usos[44] (Herrero 2016d: 360–361 y en prensa). Y por el contrario, hemos podido comprobar también que otros ejemplos sorprendentemente tempranos,

43 De nuevo hemos podido comprobarlo mediante el facsímil de la edición de 1552 digitalizado en la Biblioteca Digital Hispánica, <http://bibliotecadigitalhispanica.bne.es/view/action/singleViewer.do?dvs=1397901238321~612&locale=es&VIEWER_URL=/view/action/singleViewer.do?&DELIVERY_RULE_ID=10&frameId=1&usePid1=true&usePid2=true>.

44 Eberenz (1982: 317) señala la aparición de *apenas…cuando* en la segunda mitad del XVI, y yo mismo (Herrero 2005: 222–223) había hecho referencia a algunos ejemplos desde finales del XIV o principios del XV. Más tardía es la gramaticalización de *apenas* como conjunción temporal. Eberenz (1982: 318–19) la situaba a mediados del XVIII, y Herrero (2005: 224) había señalado un testimonio de finales del siglo XVII. Sorprende hallar ejemplos cuatro siglos anteriores, pero hemos verificado la exactitud del ejemplo alfonsí recogido en *CORDE*: «e por los grandes averes que levavan tomó muy grand

como el de *desde luego* del siglo XIII o los ejemplos del siglo XV de *pronto* con valor adverbial (de hecho todos los anteriores a la segunda mitad del XVIII) no son correctos.

Y las orientaciones temporales que nos dan los ejemplos recogidos en *CORDE* y *CDH* también pueden servir para ponernos sobre aviso de posibles malas dataciones en otras fuentes. De este modo, al estudiar la formación y evolución de las interjecciones de negación y rechazo *ca*, *quia* y *qué va* (Herrero 2014b) observamos, teniendo en cuenta los datos de *CORDE*, que se comienzan a encontrar las interjecciones de negación y rechazo *quia* y *ca*, surgidas de un proceso de elipsis a partir de oraciones interrogativas del tipo *que ha de* + infinitivo, desde el segundo cuarto del XIX. Por ello, resulta sumamente extraño que aparezca ya un ejemplo del XVII de la interjección *ca* con este valor, como señala Corominas (*DECH*, I, *s. v. ¡ca!*) en Quiñones de Benavente. Más aún teniendo en cuenta que hasta el siglo XVIII el *CORDE* ni siquiera recoge el empleo de *qué* + presente de indicativo de *haber* + *de* con valor de negación o rechazo. Si acudimos a la edición de Cotarelo y Mori de los entremeses de Quiñones de Benavente de donde toma el ejemplo Corominas[45], comprobamos que, efectivamente, aparece en ella esta interjección: «VARRETA. ¿No es mejor matallos?/JARRETE. ¡Ca, ca! Déjese regir por mí». Sin embargo, mediante la consulta de las dos ediciones del siglo XVII en que se contiene el *Entremés de las nueces*[46], de donde procede este ejemplo, y del manuscrito de dicho entremés, ca. 1649, cuyo facsímil es accesible a través de la BDH, comprobamos que la lectura correcta es *ea*, interjección corriente en el XVII (Herrero 2014b: 242–243).

4 Conclusiones

Los corpus informatizados para el estudio del español son hoy en día una herramienta sumamente útil para la investigación. Dentro de los análisis de carácter

cobdicia a Faraón, e *apenas* acabaron de soterrar sos muertos cogiós con mil vezes mill omnes d'armas e fue empós ellos» (Alfonso X, *General Estoria. Primera parte*, ca. 1275, ed. de Pedro Sánchez Prieto-Borja, Universidad de Alcalá de Henares (Alcalá de Henares), 2002, *apud CORDE*) a través del facsímil del manuscrito de la Biblioteca Nacional (signatura Mss/816) digitalizado en la *BDH*. El ejemplo está en el fol. 163 v, <http://bdh-rd.bne.es/viewer.vm?id=0000131513&page=1>.

45 Cotarelo y Mori (1911: 817a).

46 *Entremeses nuevos de diversos autores*, Alcalá de Henares, Fra[n]cisco Ropero, 1643. PR Real Biblioteca, Sign. IX/5054; *Laurel de entremeses varios: Repartido en diez y nueve entremeses nuevos. Escogidos de los mejores ingenios de España*, Zaragoza: Juan de Ibar, 1660.

diacrónico resultan imprescindibles en el estudio del léxico, y también en el de muchos aspectos de la evolución gramatical, incluyendo muchos de los trabajos sobre gramaticalización. De los diversos corpus en línea, aunque cada uno de ellos puede resultar útil para fines específicos, los más importantes para el estudio de los procesos evolutivos en la historia del español son los grandes corpus de carácter general, especialmente *CE*, *CORDE* y *CDH*. No obstante, y aun siendo irrenunciable su consulta, han de ser utilizados con bastante precaución, y desde luego hay que revisar muchos de los ejemplos que ofrecen para asegurarnos de su autenticidad y fidelidad al original y a la lengua del período estudiado, como se ha señalado en los apartados anteriores.

Sería sumamente deseable que *CORDE*, *CDH* y *CE* realizaran una detenida revisión de las ediciones digitalizadas, que, además de subsanar errores de copia, a veces muy importantes para la datación de los fenómenos estudiados y la comprensión de sus estados evolutivos, depurase los datos bibliográficos ofrecidos, y sobre todo eliminase el empleo de ediciones descuidadas o modernizadas total o parcialmente de textos antiguos, sustituyéndolas por ediciones fiables y, en la medida de lo posible, recurriendo a la digitalización de manuscritos y ediciones originales. Por supuesto, deberían también eliminarse las traducciones modernas de textos de otras lenguas que aparecen fichadas como si fuesen los originales y estos estuvieran escritos en español.

Sería también muy deseable que los grandes corpus de carácter histórico aumentasen el volumen de textos fichados, lo que permitiría afinar más en los estudios diacrónicos. Por otra parte, si nos fijamos en los distintos periodos de la historia del español, podemos ver que hay también un desequilibrio en los materiales fichados para las distintas épocas. Así, por ejemplo, en el *CORDE* se recogen 14 490 011 palabras en 3179 documentos del sigo XVIII, frente a 50 620 521 palabras en 6005 documentos en el XVI; 36 386 678 palabras en el XVII o 43 398 647 palabras en el XIX[47]. Es evidente que el siglo XVIII, importante, entre otros aspectos, para el estudio de muchos de los cambios gramaticales y de los procesos de gramaticalización en la historia del español, está muy poco representado, y se debería ampliar notablemente su peso para obtener una

47 Aunque el número de textos fichados para el XVIII es prácticamente igual al que se ficha en el XIX y solo ligeramente inferior al fichado en el XVII, su extensión es obviamente mucho menor, lo que se refleja en un número total de palabras que no llega a la mitad del que se recoge para el XVII o el XIX. Por otra parte, como señala Octavio de Toledo (2016: 63), el peso de unos pocos autores (Feijoo, Torres Villarroel, Mayans y Luzán) supone el 38 % del total de los documentos fichados para la primera mitad de este siglo.

representación semejante a la que hay para otras etapas de la historia del español. En los últimos años, los grandes corpus de referencia han incrementado extraordinariamente su volumen por lo que se refiere al español actual, como ha sucedido con el *CE* que ha aumentado su corpus hasta casi 2000 millones de palabras con textos recogidos de la web en 2013–2014, o se han creado nuevos corpus, como el *CORPES XXI*, que complementa al *CREA* añadiendo varios cientos de millones de palabras de la época más reciente del español, pero no se ha incrementado el volumen de texto fichado en el subcorpus histórico del *CE* o en el *CORDE*. Por otra parte, sería también muy útil dotar al *CORDE* de un motor de búsqueda similar al del *CORPES XXI*, como se ha hecho con el *CREA*. Es cierto que el *CDH*, que recoge la mayoría de los materiales del *CORDE* y gran parte del *CREA*, tiene un motor de búsqueda semejante al de estos corpus, y que la interfaz de consulta del *CDH* ha servido para desarrollar la de *CORPES XXI*, pero este último es más potente que el que actualmente tiene el *CDH*. Sería también muy útil desarrollar la lematización del *CDH*, extendiendo y mejorando la del corpus nuclear al conjunto del corpus[48].

Somos conscientes de la dificultad de una tarea de esta envergadura, tanto por lo que a recursos humanos como económicos se refiere, pero sin duda el aumento del volumen de datos y la mejora de la etiquetación de los mismos supondrían una importantísima ayuda para el desarrollo de muchos estudios sobre la diacronía del español.

Referencias bibliográficas

ADMYTE = Archivo Digital de Manuscritos y Textos Españoles. Madrid: Micronet, Ministerio de Cultura, Biblioteca Nacional, 1992 (CD-ROM); ADMYTE II, Madrid: Dirección General de Archivos y Bibliotecas, 1999 (2 CD-ROM). En línea. <http://www.admyte.com/admyteonline/home.htm> [último acceso: 15/05/2017].

Almerich (1965): *La fazienda de ultra mar. Biblia romanceada et itinéraire biblique en prose castillane du XII^e siècle*. Moshé Lazar (ed.), Salamanca: Acta Salmanticensia.

Azofra Sierra, Elena (2014): «Adverbios de tiempo. Demostrativos, comparativos y modo-temporales», en Concepción Company (dir.), *Sintaxis*

48 De hecho, la Real Academia señala explícitamente el proyecto de mejorar la lematización del *CDH* (Real Academia Española: *Corpus del Nuevo diccionario histórico del español* (*CDH*). *Manual de consulta (v 3.1)* p. 9, n. 1. <http://web.frl.es/CNDHE/org/publico/pages/ayuda/ayuda.view>).

histórica de la lengua española. Tercera Parte: Preposiciones, adverbios y conjunciones. Relaciones interoracionales, vol. I, México: FCE/UNAM, 313–410.

BDH = Biblioteca Digital Hispánica. <http://www.bne.es/es/Catalogos/BibliotecaDigitalHispanica/Inicio/index.html> [último acceso: 15/05/2017].

Bertolotti, Virginia/Concepción Company Company (2014): «El *Corpus diacrónico y diatópico del español de América (CORDIAM)*. Propuesta de tipología textual», *Cuadernos de la ALFAL* 6, 130–148.

Biblia medieval = Andrés Enrique-Arias y F. Javier Pueyo Mena (2008-). <http://www.bibliamedieval.es/index.php> [último acceso: 15/05/2017].

Biblioteca Digital de Textos del Español Antiguo = <http://www.hispanicseminary.org/textconc-es.htm> [último acceso: 15/05/2017].

Buenafuentes de la Mata, Cristina/Carlos Sánchez Lancis (2012): «Procesos de gramaticalización y lexicalización a la luz de los corpus académicos», en Tomás Jiménez Juliá *et al.* (eds.), *Cum corde et in nova grammatica. Estudios ofrecidos a Guillermo Rojo*. Santiago de Compostela: Universidade de Santiago de Compostela, 153–165.

Castán Lanaspa, Guillermo (1983): «Créditos, deudas y pagos en el área rural castellano-leonesa (siglos XI–XIV)», *Studia Historica. Historia medieval* 1, 67–85.

Castán Lanaspa, Guillermo/Javier Castán Lanaspa (1992): *Documentos del Monasterio de Santa María de Trianos (Siglos XII–XIII)*. Salamanca: Universidad de Salamanca.

CDH = Instituto de Investigación Rafael Lapesa de la Real Academia Española (2013): *Corpus del Nuevo diccionario histórico*. <http://web.frl.es/CNDHE> [último acceso: 15/05/2017].

CE = Davies, Mark (2002-), *Corpus del español*. <www.corpusdelespanol.org> [último acceso: 15/05/2017].

CHARTA = *Corpus Hispánico y Americano en la Red: Textos Antiguos*. <http://www.corpuscharta.es/> [último acceso: 15/05/2017].

CODEA+ = *Corpus de documentos españoles anteriores a 1800*. <http://corpuscodea.es/> [último acceso: 15/05/2017].

CORDE = Real Academia Española: *Corpus diacrónico del español*. <http://www.rae.es> [último acceso: 15/05/2017].

CORDIAM = Academia Mexicana de la Lengua: *Corpus Diacrónico y Diatópico del Español de América*. <www.cordiam.org> [último acceso: 15/05/2017].

CORPES XXI = Real Academia Española: *Corpus del español del siglo XXI*. <http://www.rae.es> [último acceso: 15/05/2017].

Cotarelo y Mori, Emilio (1911): *Colección de entremeses, loas, bailes, jácaras y mojigangas desde fines del siglo XVI á mediados del XVIII*. Madrid: Casa Editorial Bailly-Bailliére (NBAE, 18).

CREA: Real Academia Española, *Corpus de referencia del español actual*. <http://www.rae.es> [último acceso: 15/05/2017].

Davies, Mark (2002): «Un corpus anotado de 100.000.000 palabras del español histórico y moderno», *Procesamiento del lenguaje natural* 29, 21–27.

Davies, Mark (2005): «Advanced research on syntactic and semantic change with the *Corpus del Español*», en Claus D. Pusch, Johannes Kabatek y Wolfgang Raible (eds.), *Romance Corpus Linguistics II: Corpora and Diachronic Linguistics*. Tubinga: Günter Narr, 203–214.

Davies, Mark (2008): «Spanish and Portuguese Corpus Linguistics», *Studies in Hispanic and Lusophone Linguistics* 1, 149–186.

Davies, Mark (2009): «Creating useful historical corpora: a comparison of *CORDE*, the *Corpus del español*, and the *Corpus do português*», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Madrid/Fráncfort: Iberoamericana / Vervuert, 137–166.

DECH = Corominas, Joan/José Antonio Pascual (1980–1991): *Diccionario crítico etimológico castellano e hispánico*, 6 vols. Madrid: Gredos.

Delgado, Juan José (1892): *Historia general sacro-profana, política y natural de las islas del Poniente llamadas Filipinas*. Manila: Juan Atayde, Imp. de El Eco de Filipinas.

Eberenz, Rolf (1982): «Las conjunciones temporales del español», *BRAE* 62, 289–385.

Enrique-Arias, Andrés (2009): «Ventajas e inconvenientes del uso de *Biblia medieval* (un corpus paralelo y alineado de textos bíblicos) para la investigación en lingüística histórica del español», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Madrid/Fráncfort: Iberoamericana/Vervuert, 269–283.

Enrique-Arias, Andrés (2012): «*Biblia medieval*: diseño y aplicaciones de un corpus paralelo y alineado del español medieval», en Emilio Montero Cartelle (ed.), *Actas del VIII Congreso Internacional de Historia de la Lengua Española*. [Santiago de Compostela]: Meubook, 421–431.

Espinosa Elorza, Rosa María (2010): *Procesos de formación y cambio en las llamadas «palabras gramaticales»*. San Millán de la Cogolla: Cilengua.

Fernández Martín, Patricia (2015): *Perífrasis verbales en español clásico (1519-1656): novela picaresca, género epistolar y crónicas de Indias*. Madrid: Universidad Complutense de Madrid, Tesis doctoral.

Garachana, Mar/Esther Artigas (2012): «Corpus digitalizados y palabras gramaticales», *Scriptum Digital* 1, 37–65.

González Pérez, Rosario (2012): «Sobre la historia de algunos marcadores confirmativos en español: la creación de *por supuesto* y su contraste con *desde luego*», en María Marta García Negroni (coord.), *Actas del II Coloquio Internacional Marcadores del discurso en lenguas románicas: un enfoque contrastivo*. Buenos Aires: Editorial de la Facultad de Filosofía y Letras, 89–102. E-book. <http://il.institutos.filo.uba.ar/sites/il.institutos.filo.uba.ar/files/MARCADORES%202020.pdf> [último acceso: 15/05/2017].

Granvik, Anton (2018): «Variación y cambio sintáctico en las oraciones completivas de sustantivos en el español clásico: N *que* vs. N *de que*», en José Luis Girón Alconchel, Francisco Javier Herrero Ruiz de Loizaga y Daniel M. Sáez Rivera (eds.), *Procesos de gramaticalización y textualización en la historia del español*. Madrid/Fráncfort: Iberoamericana/Vervuert, 189–229.

Guerra, Francisco/María del Carmen Sánchez Téllez (1984): *El libro de las medicinas caseras de fray Blas de la Madre de Dios. Manila 1611*. Madrid: Ediciones Cultura Hispánica.

Herrero Ruiz de Loizaga, Francisco Javier (2005): *Sintaxis histórica de la oración compuesta en español*. Madrid: Gredos.

Herrero Ruiz de Loizaga, Francisco Javier (2014a): «Los operadores escalares de foco *aun, hasta, incluso* e *y todo*. Historia y uso», *Vox Romanica* 73, 179–217.

Herrero Ruiz de Loizaga, Francisco Javier (2014b): «*Quia, ca, qué va*. Elipsis y gramaticalización de elementos interjectivos de negación», en José Luis Girón Alconchel y Daniel M. Sáez Rivera (eds.), *Procesos de gramaticalización en la historia del español*. Madrid/Fráncfort: Iberoamericana/Vervuert, 233–262.

Herrero Ruiz de Loizaga, Francisco Javier (2015a): «*Según* + interrogativo con valor indefinido», *Verba* 42, 239–268.

Herrero Ruiz de Loizaga, Francisco Javier (2015b): «*Sin embargo de*. Creación y pérdida de una locución preposicional concesiva», en José María García Martín (dir.), Teresa Bastardín Candón y Manuel Rivas Zancarrón (coords.), *Actas del IX Congreso Internacional de Historia de la lengua Española*, tomo I. Madrid/Fráncfort: Iberoamericana/Vervuert, 857–877.

Herrero Ruiz de Loizaga, Francisco Javier (2016a): «'Lo mismo que te quiero te quisiera'. Formación de la locución comparativa *lo mismo que* en el español clásico», en Marta Fernández Alcaide, Elena Leal Abad y Álvaro S. Octavio de Toledo y Huerta (eds.), *En la estela del Quijote. Cambio lingüístico, normas y tradiciones discursivas en el siglo XVII*, Fráncfort: Peter Lang, 327–350.

Herrero Ruiz de Loizaga, Francisco Javier (2016b): «La formación del conector consecutivo *de ahí (que)*», en Araceli López Serena y Antonio Narbona (coords.), *El español a través del tiempo. Estudios ofrecidos a Rafael Cano Aguilar*. Sevilla: Universidad de Sevilla, 581–608.

Herrero Ruiz de Loizaga, Francisco Javier (2016c): «Historia y usos del adverbio *pronto*», *Estudios Filológicos* 57, 103–122.

Herrero Ruiz de Loizaga, Francisco Javier (2016d): «La elisión en la formación de conjunciones y marcadores discursivos», en Benjamín García Hernández y Mª Azucena Penas Ibáñez (eds.), *Semántica latina y románica. Unidades de significado conceptual y procedimental*. Berna: Peter Lang, 351–386.

Herrero Ruiz de Loizaga, Francisco Javier (2017): «*Enseguida* e *inmediatamente*. Creación de adverbios temporales de inmediatez en el español moderno», *BRAE* 97, 467–513.

Herrero Ruiz de Loizaga, Francisco Javier (2018): «*Igual que* e *igual de…que* en construcciones modales y comparativas. Estudio histórico», en José Luis Girón Alconchel, Francisco Javier Herrero Ruiz de Loizaga y Daniel M. Sáez Rivera (eds.), *Procesos de gramaticalización y textualización en la historia del español*. Madrid/Fráncfort: Iberoamericana/Vervuert, 257–298.

Herrero Ruiz de Loizaga, Francisco Javier (en prensa): «La expresión de la posterioridad inmediata: mantenimiento, pérdida y renovación de nexos y variación diatópica», *Actas del X Congreso Internacional de Historia de la Lengua Española*.

Lapesa, Rafael (1981[9]): *Historia de la lengua española*. Madrid: Gredos.

Mackenzie, David (1997[5]): *A Manual of Manuscript Transcription for the Dictionary of the Old Spanish Language*, Fifth Edition Revised and Expanded by Ray Harris-Northall. Madison: The Hispanic Seminary of Medieval Studies. Accesible en línea: <http://www.hispanicseminary.org/manual-es.htm>.

Menéndez Pidal, Ramón (1976[5]): *Cantar de mio Cid. Texto, gramática y vocabulario. Primera parte. Crítica del texto, Gramática*. Madrid: Espasa Calpe.

Nieuwenhuijsen, Dorien (2009): «El rastreo del desarrollo de algunos pronombres personales en español: (im)posibilidades de los corpus diacrónicos digitales», en Andrés Enrique-Arias (coord.), *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus*. Madrid/Fráncfort: Iberoamericana/Vervuert, 365–384.

Octavio de Toledo y Huerta, Álvaro S. (2016): «Aprovechamiento del *CORDE* para el estudio sintáctico del primer español moderno (ca. 1675–1825)», en Johannes Kabatek (ed.), *Lingüística de corpus y lingüística histórica*

iberorrománica. Berlín/Boston: de Gruyter Mouton (Beihefte zur *Zeitschrift für romanische Philologie* 405), 57–89.

Post Scriptum = CLUL (Ed.) (2014): *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. <http://ps.clul.ul.pt> [último acceso: 15/05/2017].

Rojo, Guillermo (2010): «Sobre codificación y explotación de corpus textuales: Otra comparación del *Corpus del español* con el *CORDE* y el *CREA*», *Lingüística* 24, 11–50.

Rojo, Guillermo (2012): «El papel de los corpus en el estudio de la historia del español», en Emilio Montero Cartelle (ed.), *Actas del VIII Congreso Internacional de Historia de la Lengua Española*. [Santiago de Compostela]: Meubook, 433–444.

Rojo, Guillermo (2016): «Corpus textuales del español», en Javier Gutiérrez Rexach (coord.), *Enciclopedia de Lingüística Hispánica*, vol. 2. Londres / Nueva York: Routledge, 285–296.

Sánchez-Prieto Borja, Pedro (2012): «Un corpus para el estudio integral de fuentes documentales (*CODEA*)», en Emilio Montero Cartelle (ed.), *Actas del VIII Congreso Internacional de Historia de la Lengua Española*. [Santiago de Compostela]: Meubook, 445–466.

Sánchez-Prieto Borja, Pedro *et al*. (2009): «El corpus de documentos españoles anteriores a 1700 (*CODEA*)», en Andrés Enrique-Arias (ed.), *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid/Fráncfort: Iberoamericana/Vervuert, 25–38.

Taranta, Velasco de/Licenciado Flores/Fernando Álvarez/Diego Álvarez de Chanca (1993): *Tratados de la peste*. María Nieves Sánchez (ed.), Madrid: Arco/Libros.

III
Del corpus a los estudios léxicos

Miguel Calderón Campos

Andalucismos en el corpus del reino de Granada[1]

Resumen: En este artículo se estudian seis andalucismos presentes en el *Corpus diacrónico del español del reino de Granada. 1492–1833* (*CORDEREGRA*): cuatro de ellos (*tiradizo*, *medianillo*, *acijado* y *orón*) proceden de inventarios de bienes del siglo XVII y muestran una frecuencia alta de aparición en el corpus de Granada, si la comparamos con la que tienen en el *Corpus Léxico de Inventarios* (*CorLexIn*). Ninguno de los cuatro suele aparecer en los repertorios lexicográficos como voces andaluzas. Los datos de frecuencia se han obtenido mediante un test de significancia estadística (*log-likelihood*). Las otras dos palabras estudiadas (*molle(d)o* y *lagarto*) son designaciones antiguas del bíceps del brazo, que han sobrevivido dialectalmente en Andalucía, Canarias o partes de América.

Palabras clave: Corpus lingüísticos, Lexicografía histórica, Inventarios de bienes, Declaraciones de cirujanos, Andalucismos

Abstract: This article presents a study of six andalusianisms found in the *Corpus diacrónico del español del reino de Granada. 1492–1833* (*CORDEREGRA*) *(Diachronic Corpus of Spanish in the Kingdom of Granada)*: four of them (*tiradizo*, *medianillo*, *acijado* and *orón*) come from 17th century inventories of goods, and frequently appear in the Granada corpus, when compared with their appearance in the *Corpus Léxico de Inventarios* (*CorLexIn*) (*Lexical Corpus of Inventories*). None of the four tends to appear in the lexicographical repertories as Andalusian voices. The frequency data is based on a statistical significance test (*log-likelihood*). The other two words under study (*molle(d)o* and *lagarto*) are old designations for biceps that have dialectically survived in Andalusia, the Canary Islands and in parts of the Americas.

Keywords: Linguistic corpora, Historical lexicography, Inventories of goods, Surgeons' statements, Andalusianisms

1 Introducción

El *Corpus diacrónico del español del reino de Granada* (*CORDEREGRA*), que estamos elaborando en la Universidad de Granada, encaja en lo que podríamos

1 Para la realización de este trabajo se ha contado con la financiación del Ministerio de Economía y Competitividad al proyecto de referencia FFI2013-46207-P.

llamar, tomando prestada una designación de la geografía lingüística, corpus «de pequeño dominio». Se trata de un corpus histórico y dialectal de una región que constituyó una demarcación administrativa entre 1492 y 1833. El reino venía a ocupar, *grosso modo*, las actuales provincias de Málaga, Granada y Almería, el antiguo reino nazarí, que pasó a formar parte de la corona castellana después de 1492.

En el momento actual, el corpus cuenta con unas 100 000 palabras para el siglo XVI, 127 000 para el XVII, 255 000 para el XVIII y 15 000 para el XIX (Calderón Campos 2015: 11). La mayor parte de la documentación pertenece a dos tipologías textuales: inventarios de bienes y declaraciones de testigos en juicios criminales. Los manuscritos están siendo transcritos siguiendo la metodología de la red internacional *CHARTA* (Sánchez-Prieto Borja 2011).

Estos corpus de pequeño dominio, muy reducidos en número de palabras, si los comparamos con corpus generales como el *CORDE* o el *Corpus del español* (*CE*), tienen la ventaja del rigor filológico con el que se seleccionan y transcriben los documentos, así como de la rigurosa datación y localización geográfica (Morala 2014: 6). Igualmente, los documentos se aproximan más a la oralidad, puesto que se redactan sin intención creativa ni reflexividad formal y con el apresuramiento de su finalidad utilitaria (Company 2012: 262). Son, en fin, documentos de carácter práctico, redactados por escribanos locales, cuyo único fin es tomar lo más fielmente posible las palabras de un testigo o hacer una lista de bienes legados en un testamento, entregados en una carta de dote y arras o decomisados en un embargo. Textos de esta naturaleza no sufren deturpaciones por ediciones o correcciones posteriores y son proclives a mostrar rasgos dialectales (fonéticos, morfosintácticos y léxicos) que se corregirían a favor de variantes neutras o cultas en obras literarias, periodísticas o científicas, las que, muy mayoritariamente, constituyen los grandes corpus generales.

Campos Souto se refiere precisamente a esta dicotomía entre corpus generales de gran tamaño y corpus pequeños, resaltando la necesidad de convivencia de unos y otros y advirtiendo del peligro de las ediciones mal datadas o mal editadas que se cuelan en los primeros: «Quizá se debe tomar conciencia de que la apuesta por el paradigma *Big Data* no puede hacerse a costa del abandono del —llamémoslo así— modelo *High Quality Data*» (Campos Souto 2016: 56). Los corpus de pequeño dominio se inscriben en este paradigma filológico, más centrado en la calidad de los datos que en la cantidad, pero sin excluir la representatividad cuantitativa basada en una buena selección textual.

Vamos a centrar nuestro estudio en dos contribuciones de un corpus de pequeño dominio como el *CORDEREGRA* al estudio de la historia del léxico español. En el primer caso se estudiarán algunos andalucismos del siglo XVII

(*tiradizo*, *medianillo*, *acijado* y *orón*) presentes en el corpus granadino, tomando el *CorLexIn* como corpus de control. En el segundo, se analizarán dos designaciones anatómicas empleadas por los cirujanos granadinos del siglo XVIII cuando tenían que dictar certificaciones de heridos en altercados y peleas: *molle(d)o* y *lagarto*. En el primer enfoque, centraremos nuestra atención en los inventarios de bienes del corpus granadino: el foco estará en lo cuantitativo, con la comparación de este corpus con los datos del *CorLexIn*, empleando el programa de concordancias *AntConc*, de Laurence Anthony; en el segundo estudio, nos centraremos más en aspectos cualitativos procedentes de las declaraciones de cirujanos en los expedientes judiciales de los archivos.

2 Andalucismos en los inventarios del *CORDEREGRA*

Alvar Ezquerra señalaba en 1999 los frecuentes errores de localización de muchos andalucismos, que los diccionarios registran difusamente como provinciales o generales, cuando pueden no ser ni una cosa ni la otra. Señala, igualmente, la falta de documentación histórica de muchos de ellos:

> Tras los cotejos realizados con los andalucismos que registra la Academia y nuestro *Tesoro* veíamos que casi un millar de ellos no quedaban documentados. Luego hemos visto cómo algunos aparecían entre los datos del *ALEA* […] o entre las columnas del *Vocabulario andaluz* de Alcalá Venceslada, pero no son sino una mínima parte, mientras que para el resto la única información que seguimos teniendo es la del repertorio académico (Alvar Ezquerra 1999: 67).

Presenta *almadraqueja* como uno de estos casos no documentados. Efectivamente, en el *DHLE* (1960–1996) figura como andalucismo[2], con la acepción de 'colchoneta, colchón delgado'. Pero el único ejemplo que puede aportar es uno literario de Cela: «vagabundo […] en una *almadraqueja* de pajones» (*Primer viaje andaluz*, 171, *DHLE*: s. v.). El resto es información lexicográfica no validada con ejemplos: aparece en Zerolo (1895)[3], en el *Dicc. Tecn. Hispanoamer.* (1930) y en Alcalá Venceslada (1980).

La consulta del *CORDEREGRA* sitúa la primera documentación de *almadraqueja* en 1514, en una carta de dote de Baza: «una *almadraqueja* de amarillo y colorado»; para el siglo XVII, el *CorLexIn* proporciona tres ejemplos granadinos

2 Como voz andaluza también en *DRAE* 1992, última edición que recoge esta voz.

3 Zerolo es el primero en indicar que se emplea «en algunos puntos de Andalucía», para designar «un colchoncillo largo y estrecho que usan los que tienen que dormir al raso guardando eras».

(Montefrío 1661)[4], uno murciano (Moratalla, 1632) y otro segoviano (Revenga, 1655). Por último, el corpus del reino de Granada arroja dos ejemplos inequívocos de *almadraqueja* 'colchón', que sirven para certificar la pervivencia del arabismo en el siglo XVIII: «el qual [cirujano] habiendo llegado a él lo halló vivo y quejándose; y habiéndolo desnudado en una *almadraqueja* donde estaba acostado vio tenía […] una herida» (El Salar, Granada, 1700); «tres sábanas de tiraiso […] tres almoadas […] tres colchas de giñapos […] dos *almadraquejas* de lienzo» (Frigiliana, Málaga, 1797).

Es evidente que los corpus archivísticos regionales son una fuente primaria de enorme utilidad para la lexicografía histórica, pues recogen, como se ha dicho, formas léxicas alejadas de la norma, que no suelen aparecer en los corpus generales. Gracias al *CorLexIn*, Morala (2015: 154–162) ha podido documentar fehacientemente andalucismos gaditanos registrados en el *DRAE*, el *TLHA* o el *ALEA*[5], o ausentes de los repertorios lexicográficos[6].

En las líneas que siguen vamos a localizar y estudiar algunos andalucismos presentes en el corpus del reino de Granada. Para ello, hemos creado específicamente un subcorpus de inventarios de bienes del siglo XVII de la provincia de Granada, que contrastaremos con los datos del *CorLexIn* de inventarios no andaluces de la misma centuria. En concreto, vamos a comparar un corpus granadino de 28 659 palabras[7] con el *CorLexIn* sin los inventarios de Andalucía, que suma un total de 592 937 palabras[8]. Este corpus ofrece datos de todas las provincias españolas, salvo las de Galicia, Cataluña y Baleares. De la comunidad valenciana solo recoge inventarios de Alicante; y de Tenerife en Canarias. Además, el *CorLexIn* ha digitalizado inventarios de Bolivia, Chile, Colombia, Guatemala y México, y en menor medida de Panamá, Perú, Puerto Rico y El Salvador.

---

4 Los tres con el valor inequívoco de 'colchón': «una *almadraqueja* de quatro baras, con su jenchimiento de lana» (Montefrío, Granada, 1661, *CorLexIn*).

5 Entre otras, las voces *cobra*, *tusón*, *alverjón*, *tacho*, *falsera*, *pajecillo*, etc. Las dos últimas están presentes también en el *CORDEREGRA*: «un *pajezillo* de la palancana charolado en veinte reales» (Baza, Granada, 1787), «una *falsera* de cama de lienzo tiradizo» (Murtas, Granada, 1725).

6 Como *pulseros*, *borne*, *estacón*, etc. *Pulseros* figura también en el corpus granadino, especialmente en la provincia de Málaga (6 ejemplos): «otros *pulseros* de coral» (Málaga, 1704); «unos *pulseros* de aljófar» (Baza, Granada, 1787).

7 Este corpus de estudio está formado por los inventarios granadinos del *CorLexIn* (21 836 palabras), al que se han sumado los inventarios de la provincia de Granada del *CORDEREGRA* (6823 palabras).

8 El total de palabras del *CorLexIn* (s. XVII) en el momento de la consulta (29/04/2017) era de 738 160.

El contraste entre ambos corpus se puede realizar automáticamente mediante el programa de concordancias *AntConc*, que permite la elaboración de una lista de palabras clave (*Keyword list*) del corpus de estudio, en nuestro caso, de los inventarios granadinos del XVII. En la lingüística de corpus, una *palabra clave* es aquella que presenta en un corpus dado una frecuencia inusualmente alta en comparación con la que esa misma palabra tiene en un corpus de referencia. *AntConc* calcula la relevancia de las palabras del corpus mediante un test de *log-likelihood* o de significancia estadística (que puede modificarse, si se desea, por una prueba de *chi cuadrado*). En cualquier caso, el programa asigna a cada palabra del corpus un resultado numérico denominado *keyness*, que podríamos traducir libremente como «índice de relevancia». Cuanto más alto sea, mayor es la relevancia de un determinado ítem, esto es, más inusualmente frecuente es ese ítem en el subcorpus que se está investigando en relación con el de control. Por pura lógica, los andalucismos del corpus granadino del siglo XVII deben figurar en la lista con un índice *keyness* alto, prueba de su mayor frecuencia relativa en este corpus respecto del *CorLexIn*.

De esta operación, se obtiene la lista que figura en la tabla 1, en la que se han incluido las diez palabras con un *keyness* más alto (LL en la tabla 1). Analizaremos cuatro de ellas, por haber sido insuficientemente estudiadas y sobre todo porque no suelen considerarse andalucismos en la mayor parte de los repertorios lexicográficos: *tiradizo* y *medianillo*, dos voces semiespecializadas que designaban en Andalucía oriental un tipo de tela; *acijado*, un nombre de color de origen árabe y de significado controvertido; y *orón*, generalmente interpretado como murcianismo[9].

2.1 *Tiradizo*

Procede del sintagma *lienzo tiradizo* y se aplica al lienzo casero o del lugar. Entre los ejemplos, destaco los siguientes:

> Dos camisones de lienço de lino *tiradiso* e dos pares de pañetes (1465 Córdoba, Navarro Gavilán 2014: 449); Tres sáuanas de *tiradizo* nueuas. Otra sáuana destopa (1648 Alcalá la Real, Jaén, *CorLexIn*); Otras dos almohadas de lienço *tiradiço* nuevo con encajes (1653 Cádiar, Granada, *CORDEREGRA*); Otra camisa de crea, faldas de *tiradiço*, labrada con seda negra (1661 Torre Cardela, Granada, *CorLexIn*); Otra camisa, cuerpo y mangas de crea y ruedo de *tiradizo* (1686 Cabra, Córdoba, *CorLexIn*); Otra toalla de *tiradiço* con flueques (1673 Bailén, Jaén, *CorLexIn*); Unos calzones de lienço *tiradizo* sin estrenar

9 Para *zaragüelles* se han calculado los datos sin tener en cuenta la variante *zarafuelles*, que encontramos en León y Burgos. Si la incluyéramos, su LL en Granada sería 11,92.

Tabla 1: Andalucismos más frecuentes en Granada (s. XVII)

Ítem	FOGr	FRGr	FOCor	FRCor	LL	Bayes
Tiradizo	48	1675	8	14	250,06	236,72
Henchimiento	26	907	27	46	89,02	75,68
Candiota	10	348	0	0	61,51	-
Flueque	10	348	1	2	54,90	41,56
Acijado	8	279	6	10	30,65	17,31
Alfarjía	7	244	4	7	29,01	15,67
Orón	6	209	2	3	28,10	14,76
Zaragüelles	5	174	3	5	20,45	7,11
Medianillo	3	104	0	0	18,45	-
Tinajuela	5	174	6	10	16,16	2,82

FOGr = Frecuencia observada en el corpus de Granada

FRGr = Frecuencia por millón en el corpus de Granada

FOCor = Frecuencia observada en el *CorLexIn* (sin Andalucía)

FRCor = Frecuencia por millón en el *CorLexIn* (sin Andalucía)

LL = Log-likelihood. Cuanto mayor sea el valor de LL (*log-likelihood*), más significativa es la diferencia entre ambos corpus. Los resultados por encima de 15,13 equivalen a un valor $p < 0.0001$ (más información sobre los tests de significancia estadística pueden verse en la web http://ucrel.lancs.ac.uk/llwizard.html, elaborada por Paul Rayson).

Bayes = Factor Bayes. El factor *Bayes* informa del grado de certidumbre en contra o a favor de la hipótesis nula (H0). Cuando es mayor de 10 significa que los datos aportan una «muy fuerte evidencia» contra H0 (para más detalles, véase la web http://ucrel.lancs.ac.uk/llwizard.html). Entre 2–6 implica «evidencia positiva contra H0»; y entre 6–10, «fuerte evidencia contra H0». H0 en todos nuestros ejemplos significa que la frecuencia del ítem evaluado en Granada es similar a la del mismo ítem fuera de Andalucía.

(1701 Murtas, Granada, *CORDEREGRA*); Dos zábanas de lienso *tiradiso* (1702 Frigiliana, Málaga, *CORDEREGRA*); Una falsera de cama de lienzo *tiradizo* (1752 Murtas, Granada, *CORDEREGRA*).

De *tiradizo* contamos con muy poca información lexicográfica (Calderón Campos 2011: 147–148), a pesar de tener una presencia significativa en los inventarios de bienes. En Andalucía oriental, se llamaba *tiradizo*[10] al lienzo casero, es decir, el que se tejía en el entorno próximo para el abasto ordinario de las casas. El reformista jiennense Moya Torres y Velasco lo vincula con los llamados

10 En seis inventarios granadinos aparece el grupo nominal completo: «diez varas de *lienço tiradiço*, lino y estopa» (1642 Cádiar, Granada, *CORDEREGRA*).

Mapa 1: *Tiradizo* y *medianillo* (XVII–XVIII)

«paños bastos y bayetas» del uso cotidiano de su región: «los *lienzos de tiradizos* y demás cosas de basto que la gente trabajadora gasta» (Moya Torres 1730: 73). Resultaba más grueso (ideal para el trabajo en el campo) y más barato que el lienzo medianillo y la crea. De *tiradizo* se hacían sábanas, almohadas, camisas, servilletas, colchones, toallas, manteles, etc., como se aprecia en los ejemplos citados. El adjetivo *tiradizo* referido al lienzo puede aludir al aspecto estirado y tirante de este (Morala 2017: 152).

En el siglo XVII, la designación *tiradizo* para el lienzo casero se localizaba sobre todo en Granada, Córdoba, Jaén y Almería[11]. De los ciento cuatro ejemplos incluidos en el *CorLexIn*, solo ocho se recogen fuera de esta parte de Andalucía, concretamente en un inventario toledano de Illescas, de 1626.

Las primeras documentaciones de que disponemos proceden del Archivo Histórico Provincial de Córdoba, de la segunda mitad del siglo XV (Navarro Gavilán 2014: 450). El *CORDE* recoge un ejemplo temprano y aislado (1463) en un inventario cántabro. La designación sigue plenamente viva en el siglo XVIII, tanto en Granada como en Jaén (Torres Martínez 2013). En este siglo, el *CORDEREGRA* ofrece también ejemplos malagueños.

La equivalencia de *lienzo tiradizo* con *lienzo casero* se constata en los comentarios de algunos eruditos de la época, que reflexionan sobre la decadencia de los telares locales, de los que había cuatro o cinco en cada pueblo y que iban

11 Para ejemplos almerienses, véase Morala (2017: 152). El *CorLexIn* registra cuarenta y dos ejemplos en Granada, veintisiete en Córdoba y veintisiete en Jaén.

desapareciendo por el empuje de telas traídas de otras partes de España o del extranjero, más baratas y de mayor calidad. Miñano y Bedoya, en su *Diccionario geográfico* (1826–1829), describe la decadente situación de la industria local en el pueblo alpujarreno de Pórtugos. En el ejemplo se constata la equivalencia entre lienzo casero y tiradizo:

> No hay fábricas ni talleres donde los vecinos puedan ganar un jornal, esceptuando 4 o 5 que han establecido armonas (sic) de jabón blando, con el que surten el pueblo y seis o siete de las inmediaciones; y los cuatro o seis telares de lienzo casero llamado *tiradizo*, apenas ocupan todo el año, y bastan para tejer los lienzos y su mantelería gruesa (Miñano y Bedoya 1826–1829: 97).

En el mismo tono preocupado, un miembro de la Sociedad Patriótica de Jaén, vinculado con la Sociedad Económica de Amigos del País, se lamentaba a finales del siglo XVIII de que se estuviera abandonando el cultivo del lino en la región, con lo que las telas locales (de *tiradizo* y *medianillo*) estaban siendo sustituidas por otras foráneas de mejor precio:

> Los lienzos hacen más falta que todo; y sin embargo de que podía haber muy buenos linares en esta Vega […] apenas se deja lugar para ellos, porque los hortelanos tienen cosechas más seguras y más útiles en su hortaliza, en el trigo y el maíz […]. Por eso, y porque sin embargo de los grandes derechos que pagan […] los lienzos extrangeros, aún se venden estos con más conveniencia, se va aniquilando este ramo de industrias ¿Quién ha de comprar una vara de *tiradizo* común y estrecho por ocho y nueve rs. si la de crea cuesta menos, siendo más blanca, más fina y más ancha? (Martínez de Mazas 1794: 443–444).

La designación de *(lienzo) tiradizo* alterna en los inventarios del XVII con *lienzo casero* (169 ejemplos)[12] y *lienzo de la tierra* (51 ejemplos): «dos colchones de lienzo casero llenos de lana» (1666 Arcos de la Frontera, Cádiz, *CorLexIn*), «dos almuadas de lienço casero, nuebas» (1648 Mardid, *CorLexIn*), «quatro serbilletas de mano, de lienço de la tierra» (1633 Tolosa, Guipúzcoa, *CorLexIn*).

2.2 *Medianillo*

Procede de la nominalización del adjetivo en el sintagma *lienzo medianillo*. Designa al lienzo casero o del lugar, más fino que el *tiradizo*:

> Otra sábana de *medianillo* con su randa (1628 Antequera, Málaga, *CorLexIn*); Dos sábanas de lienço *medianillo* a medio traer (1659 Almería, *CorLexIn*); Otros calçones de *medianillo*, con sus encaxes y puntas (1661 Guadahortuna, Granada, *CorLexIn*); Una

12 A veces también el adjetivo se sustantiva, como en «una almoada de *casero* usada» (1651 Dalías, Almería, *CorLexIn*), «sáuana de *casero* demediada» (1676 Trigueros, Huelva, *CorLexIn*).

camisa de mujer, el cuerpo de *medianillo* (1664 Cabra, Córdoba, *CorLexIn*); Vna toalla de *medianillo* con puntas (1673 Bailén, Jaén, *CorLexIn*); Otras dos almohadas de *medianillo* mediadas (1701 Murtas, Granada, *CORDEREGRA*); Dos pedasos de lienso de *medianillo* (1705 Málaga, *CORDEREGRA*); Una docena de camisas, la una de ellas de crea, otra de *medianillo* (1795 Ugíjar, Granada, *CORDEREGRA*).

El diminutivo *medianillo* se empleaba en Andalucía oriental para designar al lienzo fino manufacturado en el entorno próximo: «unos lienzos más delgados, que dicen *medianillos*, haciendo la hilaza muy delgada» (Moya Torres 1730: 74). El mecanismo de derivación es el mismo que el empleado en otros casos, frecuentes en el *CorLexIn*, donde las lexicalizaciones de nombres de telas se realizan mediante el sufijo *-illo*, para indicar una tela más fina o más basta: «un manto de cañamillo» (1646 Alfaro, La Rioja); «otro bestido de picotillo de seda» (1646 Medina de Rioseco, Valladolid); «una capa de pardillo» (1647 Villacalbiel, León); «unas naguas de damasquillo açul» (1649 Fiñana, Almería); «un guardapiés de chamelotillo» (1676 Santander); «huna sáuana de yladillo» (1697 Narila, Granada), etc.

Es importante destacar que la distribución geográfica de *medianillo* coincide con la de *tiradizo* (ver mapa 1): los treinta y siete ejemplos del *CorLexIn* proceden de Almería, Córdoba, Granada, Jaén y Málaga. De *medianillo* se hacían, al igual que de *tiradizo*, almohadas, sábanas, colchas, toallas, camisas, calzones, peinadores, paños de manos, etc.

En el siglo XVIII, el corpus de Granada recoge catorce ejemplos de *medianillo*, en inventarios de Granada capital (1748), Alpujarras (1701, 1790), Baza (1703) y Málaga (1705).

2.3 *Acijado*

Acijado es un derivado de *acije* 'sulfato de hierro, cobre, cinc o cobalto', procedente del árabe andalusí *azzáj*, y este del clásico *zāj* (Corriente 1999: s. v. *aceche*). En los inventarios del siglo XVII equivale muy probablemente a un color parecido al rojo cobrizo.

Un frutero de red *açijada*, con puntas blancas y *açijadas* (1599 Granada, *CORDEREGRA*); Siete camisas […] labradas de blanco, azul y *azixado* (1659 Serón, Almería, *CorLexIn*); Dos almoadas de ruan, labradas de hilo *azijado* y asul (1660 Rociana del Condado, Huelva, *CorLexIn*); Otro par de almohadas de olanda labradas con hilo *azijado* (1665 Andújar, Jaén, *CorLexIn*); Almoadas labradas de hilo *cexado* (1671 Baza, Granada, *CorLexIn*); Vn paño de manos de medianillo labrado con hilo *açijado* (1673 Bailén, Jaén, *CorLexIn*).

Es muy significativa la distribución geográfica de *acijado* y sus variantes ortográficas (*acijado*, *azijado*, *açijado*, *asijado*, *acixado*, *azixado*, *cejado*, *çexado*, etc.),

Tabla 2: *Acijado* en el siglo XVII

	H	AL	GR	CA	J	SE	CO	MA	Resto	Total
nº de casos	12	11	8	8	5	2	2	1	6	55

circunscrita casi exclusivamente a Andalucía, región en la que se localizan cuarenta y nueve de los cincuenta y cinco ejemplos extraídos del *CorLexIn* y *CORDEREGRA* (s. XVII) (ver tabla 2). Los seis ejemplos no andaluces proceden de México (dos casos), Toledo, Ciudad Real, Álava y Valladolid.

Resulta difícil determinar el color que designa el adjetivo *acijado*. Los ejemplos del corpus no dan apenas pistas: se aplica casi exclusivamente al hilo («hilo acijado»), con el que se labran o bordan almohadas, cojines, toallas, camisas, paños, etc., o bien al hilo con el que se confeccionan los flecos y puntas de estas telas. El color con el que más se combina es el azul («azul y acijado»).

El problema es que el arabismo *acije* o *aceche* era el nombre genérico para designar distintos sulfatos de minerales diferentes: de hierro, cobre, cinc y cobalto, y cada uno de estos compuestos tiene un color particular: el sulfato de hierro es verde, el de cobre, azul, el de cinc, blanco, el de cobalto, rojo. Por tanto, en sentido estricto, el sustantivo *acije* necesita un adjetivo para especificar su color.

Desde la Edad Media, *acije* es sinónimo de *caparrosa*[13] y *vitriolo*. De la misma forma que su sinónimo de origen árabe, la caparrosa o el vitriolo podían ser de color verde, azul, blanco o rojo, según el tipo de sulfato de que se tratase.

En el español actual, *acijado* equivale exclusivamente a una tonalidad de verde, que se corresponde con el color del sulfato de hierro o caparrosa verde (Maíllo Salgado 1998: 62). Así se refleja de manera inequívoca en diccionarios especializados como el de Sanz/Gallego (2001: s. v. *acije*): 'coloración estándar constituida por la específica *verde caparrosa*'. En la Edad Media, el verde era uno de los colores posibles del *aceche* o *acije*, aunque, como se ha dicho, no era el único:

> es la piedra quel dizen calcadiz, et esta es una de las maneras de azeche que ya nombramos: *verde es de color* […] es otrossi la quarta manera de azech que auemos dicho: *blanco*

13 El *DECH* considera una posible, aunque no segura, etimología árabe *qubrusî* 'vitriolo de Chipre' (de cobre) para *caparrosa*. No obstante, el *OED* considera más probable que las formas latinas medievales *cuprosa* y *cuperosa*, que estarían en el origen de los derivados en las diferentes lenguas europeas, provengan de un hipotético *(AQUA) CUPROSA 'de cobre' (*OED*: s. v. *copperas*). Para Corominas, la forma AQUA CUPROSA no es la etimológica, sino un latinismo tardío explicable por etimología popular.

> *es de color* […] es la piedra del azech que dizen calcatar […] et a las uezes fallan en esta pedaços del *azech uermeio* que ante diximos, que quandol tuellen una tela que tiene de suso paresce de dentro amariello […] es meior aquello que a *color de laton* […] ay otro [azech] en que fallan uenas como color doro […] quando la quebrantan, luze […] & por essol llaman *[azech] estrellenno* (*Lapidario* 1250?–1279?, Kasten/Nitti 2002: s. v. *aceche*).

En consecuencia, tanto el *acije* como la *caparrosa* podían aludir al color verde, azul, blanco o incluso rojo. Para complicar más la situación, el acije se combinaba con otros ingredientes para fabricar tinta y otros colorantes de color negro[14]. Es muy conocida la composición de la llamada *tinta ferrogálica*, que se hacía mezclando agallas de roble, acije o caparrosa verde (sulfato de hierro), agua y goma arábiga. El acije se extraía evaporando el agua de tierras ferrosas, como las que se encuentran en el entorno de Riotinto en Huelva. Este río, para los árabes, recibía el nombre de *Acije*, como advierte Covarrubias:

> Azige. Cierta tierra con que se tiñen las lanas y los paños de negro. Antonio Nebrisensse: para tinta. Atramentum, atramenti. Yo presumo ser lo mesmo que llamamos en Castilla tierra de Seuilla. Es también *Azige* nombre de vn rio en el Andalucia entre Palos y Huelva dicho ansí por la muchedumbre de esta tierra que en aquellos lugares se saca (1611 Covarrubias *Suplemento*, *NTLLE*: s. v. *azige*).

La vinculación del acije o aceche con el color negro la recoge *Autoridades*[15], que sigue en lo esencial a Covarrubias:

> Azeche: Cierto genero de tierra negra que se halla en las bocas de los minerales de cobre, a manera de sal: la qual ordinariamente sirve para hacer tinta, y por otro nombre se llama tierra de Sevilla […]. Puede venir esta voz del río Azéche, que corre bien cerca de Sevilla: vocablo Arábigo, que vale tanto como rio negro o rio tinto (1726 *Dicc. Autoridades*, *NTLLE*: s. v. *azeche*).

Consecuentemente, *acijado* se traduce al latín como «nigrescens, fuscus» (*DRAE* 1770, *NTLLE*: s. v. *acijado*), es decir, 'negruzco, oscuro, moreno, de tez morena' (*OLD*: s. v. *fuscus*). Con todo lo anterior, las posibilidades cromáticas del acije aumentan: verde, azul, blanco, rojo o negro[16]. Sin embargo, el color acijado del

---

14 «E quando quisieren fazer a los blancos que les nazcan pelos prietos, tomen del azech e del çumo del estierco de los asnos e del seuo de las cabras tanto de lo uno como de lo al e cueganlo todo en uno» (*Libro de Moamyn* 1250?–1300?, Kasten/Nitti 2002: s. v. *aceche*).

15 «Azige. Lo mismo que azéche. Viene del arábigo Zig, que significa cosa negra» (1726 *Dicc. Autoridades*, *NTLLE*: s. v. *azige*).

16 Pocos investigadores se pronuncian sobre el color acijado. Para Matarín Guil/Ruescas Granados (2000: 169), el *acijado* de los inventarios almerienses debe de ser verde; Martín Morales (2016: 14) defiende, basándose en *Autoridades*, el color negro de *acijado*.

hilo de los inventarios debía ser uno bien definido e inequívoco, pues en ningún caso aparecen aclaraciones desambiguadoras.

Los ejemplos que figuran en los repertorios lexicográficos tampoco son muy reveladores, salvo dos del siglo XVII, el mismo siglo de nuestros inventarios. El primero es un ejemplo de Fr. Agustín de Vetancurt, de 1698: «Los chicozapotes son pardos en el color de la cáscara y *azijada* la carne con unas pepitas negras; es sana y muy suave» (*DHLE*: s. v. *acijado*). El otro ejemplo es un verso de Fr. Damián Cornejo (1629–1707): «Fieros caballos de color de *acije*» (Pagés 1902, *NTLLE*: s. v. *acije*).

Este segundo ejemplo solo sirve para que pierda fuerza la hipótesis del acije verde. El primero es mucho más interesante, puesto que podemos saber de qué color es la carne del chicozapote, denominación popular del fruto del árbol tropical *Manilkara zapota*, de cuya corteza se extrae una sustancia gomosa con la que los mayas fabricaban el chicle. Los frutos de este árbol son unas bayas de unos diez centímetros de diámetro, cuya cáscara es marrón y la pulpa de color amarillo rojizo, *acijado* en palabras de Vetancurt[17].

Aunque los escribanos de los inventarios del *CorLexIn* y *CORDEREGRA* no sintieron la necesidad de aclarar el significado del *hilo acijado*, sí podemos obtener datos interesantes de las concordancias de otras designaciones de color que se extraen del programa *AntConc*. Hemos visto que cuarenta y nueve ejemplos de *acijado* (de un total de cincuenta y cinco) se localizan en el siglo XVII en las ocho provincias andaluzas. Cabe preguntarse cómo se llamaba a ese hilo acijado fuera de esta región, pues parece evidente que más allá del territorio andaluz tendría que haber almohadas y cojines bordados con hilo del mismo color, aunque denominado de otra forma. Después de revisar las combinaciones de *hilo* con adjetivos de color en el *CorLexIn* nos llamó la atención la de «hilo almacigado», por varias razones.

En primer lugar, lo más llamativo es que el color almacigado jamás aparece en inventarios andaluces (ver mapa 2)[18]. Pero además, *almacigado*, muy significativamente, suele combinarse con el sustantivo *hilo* (en veintidós de los treinta y tres ejemplos), exactamente igual que *acijado*; además, como este, solo se aplica al color de telas bordadas con las que se hacían almohadas, paños, toallas,

17 Recuérdense también los ejemplos medievales de *aceche bermejo* y *aceche estrelleño*, que apuntan asimismo a tonos dorados y rojizos.

18 Figura en Ávila, Badajoz, Cáceres, Cuenca, Guadalajara, La Rioja, Madrid, Puerto Rico y Toledo.

camisas y cojines. Por último, se combina frecuentemente con el color azul («azul y almacigado»)[19].

De todo lo anterior, podría establecerse una especie de distribución complementaria, en la que el mismo color se denominara *acijado* en Andalucía y *almacigado* en otras regiones de la Península. Si se acepta esta sinonimia, bastaría con determinar el valor de *almacigado* para poder visualizar cómo eran las almohadas de hilo acijado del corpus.

Afortunadamente, el sentido de *almacigado* en el siglo XVII no es tan problemático como el de *acijado*. *Almacigado* ha significado siempre 'de color amarillo o de almáciga' (*DHLE*, s. v.). La almáciga es una resina cuyo color oscila entre el amarillo anaranjado profundo y el cobrizo profundo (Sanz/Gallego 2001: s. v. *almáciga*), esto es, un color muy parecido al de la «carne acijada» del chicozapote. Además, este color rojizo podría aplicarse sin problemas a los «fieros caballos de color de acije», del ejemplo citado por Pagés.

Por otra parte, la sinonimia de *acije* y *caparrosa* nos da otras pistas interesantes. En el *CorLexIn* encontramos veintinueve ejemplos de *caparrosa*, procedentes de Ávila, Vizcaya, Burgos, Cantabria, León, Palencia, Soria, Guipúzcoa, Valladolid y Zamora. Este color se aplica exclusivamente al hilo con que se labran lienzos para elaborar almohadas, paños, toallas y sábanas. Cinco ejemplos son especialmente significativos:

> Una almohada labrada con *caparrosa* açul (Candeleda, Ávila, 1646).
> Una almuada de lienço, nueba, labrada de *caparrosa* y açul (Revilla del Campo, Burgos, 1639).
> Una almuada de lienzo delgado labrada de azul y *caparrossa* (Villalobos, Zamora, 1654).
> Vna madeja de yllo rojo que llaman *caparroso* (Tolosa, Guipúzcoa, 1633).
> Quatro madejas de yllo rojo *caparroso* (Tolosa, Guipúzcoa, 1633).

En los veinticuatro casos restantes, *caparrosa* aparece sin adjetivar, como en «dos paños de manos de lienço, vno labrado con *caparrosa*» (Solanilla, León, 1662), por lo que se plantean las mismas dudas que suscitaba el color *acije*. No obstante, estos cinco ejemplos pueden ayudarnos a entender el auténtico significado de *caparrosa* en los inventarios. Los tres primeros parecen descartar que la caparrosa no adjetivada sea azul: si así fuera, es decir, si el color por defecto de la caparrosa fuera el azul, no tendría sentido repetir el adjetivo, como ocurre en el ejemplo de Ávila, ni tampoco sería razonable que se combinara con este color («caparrosa y azul»), en los ejemplos de Burgos y Zamora.

19 Conviene recordar la combinación de «azul y acijado», citada anteriormente.

Mapa 2: *Acijado*, *almacigado* y *caparrosa* en el s. XVII

Los dos ejemplos guipuzcoanos son todavía más interesantes: el tasador que dicta el inventario se ve en la obligación de precisar qué se entiende por «hilo caparroso»: «yllo rojo que llaman caparroso». Probablemente este tasador no estaba familiarizado con la denominación *caparrosa* y cree conveniente glosarla, gracias a lo cual desvela el sentido de los casos de *caparrosa* no adjetivados, que se corresponden, otra vez, con un tono rojizo (un rojo especial «caparroso»), precisamente el mismo del hilo acijado y del almacigado, según la interpretación que venimos defendiendo.

Los inventarios del siglo XVII, analizados en su conjunto, parecen indicar que el hilo acijado andaluz se llamaba *almacigado* o *caparrosa* en otras regiones. Los tres tienen en común que se aplican exclusivamente a lienzos (y al hilo con que se bordan estos lienzos), con los que se hacían almohadas, paños, toallas, camisas, etc. Son, por otra parte, las tres únicas designaciones de color que tienen esta especificidad de aludir únicamente al color de un hilo y no al de otras cosas. Además, se distribuyen complementariamente: *acijado* en Andalucía, *almacigado* en una extensa región del centro-oeste de la Península y *caparrosa* en el norte (ver mapa 2). *Almacigado* apunta inequívocamente a tonos rojizos; *caparrosa*, también, a juzgar por la aclaración de un tasador de Tolosa. De ahí, no parece aventurado afirmar que el hilo acijado de los inventarios andaluces (el «hilo acijado de Sevilla») tuviera un color rojizo o cobrizo, similar al de las aguas del Río Tinto, el Acije de los árabes, y al de la carne acijada del chicozapote. Habrá que seguir leyendo inventarios con la esperanza de que algún

tasador constate, como su colega guipuzcoano, la existencia de una madeja de hilo rojo llamado «acijado».

2.4 *Orón*

De orón 'espuerta de esparto grande y redonda' destacamos los siguientes ejemplos:

> Un *hurón* pequeño de nueçes. Hanega y media de habas en un *hurón* pequeño […] dos hanegas de linaça en un *hurón* […] Una trox llena de trigo. Un *hurón* pequeño de trigo (1561 Baza, Granada, Espinar Moreno/Espinar Jiménez 2015: 78–79); Dos *orones* de esparto (1641 Pitres, Granada, *CORDEREGRA*); Vn *orón* donde abía seis fanegas de zebada (1648 Alcalá la Real, Jaén, *CorLexIn*); Vn *orón* pequeño, hasta ocho o nuebe fanegas de trigo en él (1649 Cuevas de Almanzora, Almería, *CorLexIn*); Un *oronzillo* pequeño (1655 Alcalá la Real, Jaén, *CorLexIn*); Un *orón* con diez fanegas de habas (1661 Montefrío, Granada, *CorLexIn*); Vn *orón*, y en él, como seis fanegas de linaza (1748 Baza, Granada, *CORDEREGRA*); Quatro jerpiles, siete espuertas y un *orón* de bedriado (1753 Mecina Bombarón, Granada, *CORDEREGRA*).

Como se observa en el mapa 3, el sustantivo *orón* en el siglo XVII se empleaba preferentemente en Andalucía oriental (seis casos en Granada, ocho en Jaén y tres en Almería)[20] y Murcia (tres ejemplos). La Academia lo recoge por primera vez en la edición de 1956, con tres acepciones: una general, 'serón grande y redondo', y dos murcianas: 'sitio en que se guarda el trigo en las casas de la huerta' y 'especie de tubo de grandes dimensiones, hecho de pleita, para contener grano'[21]. La edición de 2014 elimina las dos acepciones murcianas y mantiene solo la general de 'serón'. Oudin (1607) y Sobrino (1705) registran *orón* como una especie de dique para contener el desbordamiento de los ríos. Terreros en 1787 indica que esta palabra ya no estaba en uso en su época. Antes encontramos testimonios en Rosal (1611, *NTLLE*, s. v. *orón*): «serón lleno de tierra», y en Nebrija (1495, *NTLLE*, s. v. *orón*): «orón lleno de tierra».

Dialectalmente, se localiza *orón* (*horón* o *(h)urón*) en Granada, Almería, Málaga[22], Murcia, Jaén (zona de Segura) y Cuenca (Idáñez de Aguilar 2015: 375).

---

20 El punto más occidental donde se localiza *orón* en el *CorLexIn* es Teba, en la provincia de Málaga.

21 Señala Corominas que es frecuente que la misma palabra signifique 'granero' y 'cesta, canasto'. Entre los ejemplos cita este de *horón* y otros como *troj(a)* (*DECH*: s. v. *cenacho*).

22 En el *ALEA* III, lámina 754 ('recipiente de esparto para guardar granos, panes, etc.'), figura en las localidades granadinas de Gor, Guadix, Charches, Lugros y Caniles (*urón*); en las malagueñas de Benahavís y Estepona, y en las almerienses de Vélez Rubio, Oria, Alcóntar y Tahal (*urón*).

Mapa 3: *Orón* en el s. XVII

Como murcianismo aparece en García Soriano (1932: s. v. *horón*) y en el *Vocabulario del noroeste murciano* de Gómez Ortín (1991: 308). Como andalucismo se registra en Alcalá Venceslada (1951/1980, s. v. *horón*), en el *Vocabulario de la Alta Alpujarra* de García de Cabañas (1967) y en el léxico almeriense de Torres Montes (2004: 263; véase también *THLA*).

3 *Molle(d)o* y *lagarto* en los partes médicos del *CORDEREGRA*

Este capítulo servirá para presentar otra posibilidad de estudio léxico a partir de los datos del corpus del reino de Granada. En este caso, se centrará la atención en el otro gran bloque textual que compone el corpus, las declaraciones de testigo, y más concretamente, las declaraciones de los cirujanos, barberos o flebotomianos. Estos facultativos dictaban ante el juez partes médicos o «declaraciones de esencia», en los que describían técnicamente el alcance y gravedad de las heridas que se denunciaban en los juicios penales celebrados en la Real Chancillería de Granada. Hasta la fecha, hemos podido transcribir treinta y siete declaraciones de cirujanos y sangradores granadinos del periodo comprendido entre 1700 y 1795. En todas ellas los médicos trataban de describir con precisión qué partes del cuerpo habían sido dañadas, qué instrumento había provocado las heridas y cuál era el pronóstico y gravedad de las mismas. Desde el punto de vista lingüístico, los informes «chirúrgico-legales» son muy interesantes porque recogen

tanto las voces populares como los tecnicismos médicos que se empleaban para nombrar las partes del cuerpo y las dolencias de las víctimas.

En el siglo XVIII se asistió a una pugna decisiva entre las designaciones romancistas y latinistas en el ámbito de la medicina (*casco* frente a *cráneo*, *derramamiento* o *flujo de sangre* frente a *hemorragia*, *quijada* frente a *mandíbula*, etc.). Vamos a analizar en concreto el debate que ocurrió con los nombres de la parte carnosa del brazo, es decir, de la «carne del brazo que, cuando se flexiona el antebrazo haciendo fuerza adquiere forma de bola» (*ALEA* V, 1628).

En el setecientos, el español contaba con tres voces tradicionales para esta designación: *morcillo*, *molledo* y *lagarto*, y con dos cultismos de reciente incorporación: *bíceps* y *músculo*. Aunque la batalla la han ganado en español los latinismos *bíceps* y *músculo*, dos de las designaciones populares dieciochescas, *molleo* (más concretamente, la variante *mollero*) y *lagarto* se han conservado dialectalmente, tanto en Andalucía como en Canarias y partes de América.

El cultismo *músculo* había empezado a utilizarse en español a finales del siglo XV, pero todavía se veía en el XVIII como latinismo puro:

> *Músculo* es palabra latina y diminutivo de la voz *mus*, que significa 'ratón', porque lo que en nuestro cuerpo llamamos *músculo* lo parece, assí en lo veloz del movimiento como en tener cabeza, vientre y cola, por lo qual también nuestros vulgares le llaman *lagarto*, *murecillo* o *morcillo* (Martínez 1728: 40).

Efectivamente, como indica el médico novator Martín Martínez, MUSCULUS (MUS+-CULUS, *OLD*: s. v. *mus*, *muris* y *musculus*), literalmente 'ratoncito', se empleaba ya en latín tanto para designar al roedor como para referirse metafóricamente en anatomía al «instrumento del movimiento» (1734 *Dicc. Autoridades*, *NTLLE*: s. v. *músculo*)[23]. Se localiza en diccionarios de español desde Oudin (1607), que todavía remite para su definición al sinónimo *murezillo*, más habitual en la época.

Murecillo y las variantes *morecillo* y la más frecuente *morcillo* comparten etimología con *músculo*. Se trata obviamente de un derivado de MUS, MURIS 'ratón', al que se le añade el diminutivo medieval *-eci(e)llo*. Como indica Martínez (1728), estas variantes se sentían como coloquiales en el siglo XVIII, en contraste con el tecnicismo anatómico *músculo*.

---

23 Por vía popular, MUSCULUS dio en español *muslo*, que se documenta desde el siglo XIII (*DECH*: s. v. *mur*); dialectalmente ha pervivido la designación *muslo del brazo* para el bíceps (*ALEA* V, 1268; *ALEANR* VII, 981).

La otra designación vulgar del músculo era el sustantivo *lagarto*, que encontramos en español al menos desde 1542, en referencia tanto al muslo como al bíceps braquial:

> Tirole un golpe a Marcelino que le derribó la falda toda y el escudo y Marcelino le dio otro que le acertó en el *lagarto* del muslo, que le hizo hincar la rodilla en tierra (Anónimo, *Baldo*, 1542, *CORDE*).
> Hirieron al maestre de campo en un brazo muy mal, porque fue en el *lagarto*, y segund pareció tenía ponzoña la flecha (Anónimo, *Tercera relación anónima de la jornada que hizo Nuño de Guzmán*, 1544, *CORDE*).
> Músculo o *lagarto*: es una parte orgánica, compuesta de carne, tendones, fibras membranáceas, nervios, arterias y venas, y es instrumento de los movimientos voluntarios (Torres Villarroel, Diego de, *Anatomía de todo lo visible e invisible*, 1738–1752, *CORDE*).

En latín LACERTUS (y la variante dialectal *LACARTUS, *DECH*: s. v. *lagarto*) designaba tanto al reptil como a la parte del brazo comprendida desde el hombro hasta el codo (*OLD*: s. v. *lacertus*). Los tratados médicos medievales, escritos en latín, recogen el sintagma *lacerti brachiorum* (Barcia Goyanes 1993: 269); incluso en el lenguaje anatómico castellano del siglo XV se empleó el cultismo *lazertos* (Herrera 1996: s. v. *lagarto*). Pero la forma patrimonial *lagarto* fue la que tuvo más éxito en español. Es la variante que usan los barberos del siglo XVIII en el reino de Granada: «más abajo, en el *lagarto* de el brazo […] le ha curado otra contusión» (Pinos Puente, Granada, 1722, *CORDEREGRA*) y la que se conserva dialectalmente en Andalucía (Villarrodrigo, Ja200, *ALEA* V, 1268) y sobre todo en Canarias[24]. Actualmente, *lagarto* pervive en español como designación metafórica de un corte de carne del cerdo ibérico, de la parte próxima al lomo.

Por último, los cirujanos granadinos del setecientos designan al bíceps del brazo con la voz popular *molle(d)o*:

> Vio tenía en el braço y costado derecho una herida de forma que tenía pasado el brazo por el *molledo* (El Salar, Granada, 1700, *CORDEREGRA*).
> A cuyo tiempo le dijo el Salvador Miranda a el declarante no fuese desvergonzado, y sacando un cuchillo le tiró a el declarante un golpe con él y lo hirió en el *molledo*, que fue el de el brazo disquierdo (Cúllar Vega, Granada, 1753, *CORDEREGRA*).
> Le halló una herida en el brazo disquierdo en la parte alta de el *molleo* hecha con instrumento cortante y punzante como navaja o cuchillo (Cúllar Vega, Granada, 1753, *CORDEREGRA*).
> Cortó cuero, gordura y membrana carnosa, y asimismo una contusión en el brazo hizquierdo en el *molleo* (Atarfe, Granada, 1737, *CORDEREGRA*).

---

24 Casi siempre en la forma completa *lagarto del brazo* (lagartọ l bra:ṣo): la variante *legarto* se localiza en las localidades Go 2, Tf 3, 4, 30, 31 y GC 1, 4, 11 (*ALEICan* II, 500).

Esta forma *molleo*, en su variante *mollero*, sobrevive en Andalucía y Canarias (*DHECan*: s. v. *mollero*). En el *ALEA* (V, 1628) *mollero* (y *los molleros*) es la designación más frecuente para el bíceps braquial, extendida ampliamente por todas las provincias. En Canarias se localiza en una veintena de localidades (*ALEICan* II, 500)[25].

Además, *mollero* 'bíceps' se conserva en Guatemala, El Salvador[26], República Dominicana, Perú (*Diccionario de Americanismos —DA—*: s. v. *mollero*), Puerto Rico[27] (Álvarez Nazario 1972: 129), Cuba y zona central del norte de México (Lope Blanch 1971: 49–51). En Puerto Rico[28], la variante *molleto* se aplica a la persona de raza negra, grande y musculosa (*DA*: s. v. *molleto*).

4 Conclusiones

Los «corpus de pequeño dominio» como el *CORDEREGRA* o el *CorLexIn* se inscriben en el paradigma que Campos Souto (2016: 56) denomina *High Quality Data*, especialmente útil para el estudio historicolingüístico del léxico dialectal. Gracias a la comparación de los inventarios del corpus granadino con los no andaluces del *CorLexIn* se ha podido determinar con procedimientos estadísticos (test de *log-likelihood*) la presencia de los diez andalucismos más frecuentes en el corpus de Granada y determinar con precisión el significado y la distribución geográfica en el siglo XVII de cuatro de ellos: *tiradizo*, *medianillo*, *acijado* y *orón*.

Por otra parte, las declaraciones de los cirujanos del *CORDEREGRA* nos han brindado la oportunidad de estudiar la vitalidad dialectal de dos romancismos anatómicos del siglo XVIII: *molle(d)o* y *lagarto*, con los que se designaba al bíceps del brazo. Actualmente, *molle(d)o*, en sus variantes *mollero* y *molleto*, sobrevive en Andalucía, Canarias y algunas partes de América. *Lagarto* se conserva en las designaciones andaluzas y canarias del bíceps, según se atestigua, respectivamente, en el *ALEA* y *ALEICan*.

---

25 LP 10; Go 40; Tf 41, 50; GC 1, 2, 3, 11, 30, 40; Fv 1, 2, 20, 31; Gs 1 y Lz 1, 3, 4, 20, 30. En Tf 6 se registra *molleho*.

26 «Me ha salido un *mollero* en el brazo por jalar tantas hojas del calendario» (*CE*).

27 En Puerto Rico se emplea frecuentemente en sentido figurado, tanto en el lenguaje deportivo («no podemos ir a un juego pensando que podemos imponer *mollero* cuando hasta nuestros jugadores más atléticos carecen del físico para hacerlo») como en el político («con ese *mollero* político poco se podrá hacer») (*CE*).

28 «Mira a ese maldito *molleto* con una muchacha blanca […]. Él es solamente otro *molleto* que se casó con una blanca, haciéndola tan negra como él» (Piri Thomas, *Por estas calles bravas*, 1999, Google libros).

Referencias bibliográficas

Alcalá Venceslada, Antonio (1951/1980): *Vocabulario andaluz*. Madrid: Gredos.

ALEA = Alvar, Manuel (1960–1973): *Atlas lingüístico y etnográfico de Andalucía*. Granada: Universidad de Granada-CSIC.

ALEANR = Alvar, Manuel (1979–1980): *Atlas lingüístico y etnográfico de Aragón, Navarra y Rioja*, 12 vols. Zaragoza/Madrid: Diputación Provincial de Zaragoza/CSIC.

ALEICan = Alvar, Manuel (1975–1978): *Atlas lingüístico y etnográfico de las Islas Canarias*, 3 vols. Las Palmas de Gran Canaria: Publicaciones del Excmo. Cabildo Insular.

Alvar Ezquerra, Manuel (1999): «Pervivencia de los andalucismos del *DRAE*», en Eduardo Forastieri Braschi, Julia Cardona y Humberto López Morales (eds.), *Estudios de lingüística hispánica: homenaje a María Vaquero*. Puerto Rico: Universidad de Puerto Rico, 56–72.

Álvarez Nazario, Manuel (1972): *La herencia lingüística de Canarias en Puerto Rico: estudio histórico-dialectal*. San Juan: Instituto de Cultura Puertorriqueña.

Barcia Goyanes, Juan José (1978–1993): *Onomatología anatómica nova. Historia del lenguaje anatómico*, 9 vols. Valencia: Universidad.

Calderón Campos, Miguel (2011): «La documentación archivística del reino de Granada como fuente lexicográfica», en Pilar Carrasco Cantos y Francisco Torres Montes (eds.), *Lengua, historia y sociedad en Andalucía. Teoría y textos*. Madrid/Fráncfort: Iberoamericana/Vervuert, 123–154.

Calderón Campos, Miguel (2015): *El español del reino de Granada en sus documentos (1492–1833)*. Berna: Peter Lang.

Campos Souto, Mar (2016): «Lexicografía del futuro para la lengua del pasado», en Rosalía Cotelo García (coord.), *Entre dos coordenadas. La perspectiva diacrónica y diatópica en los estudios léxicos del español*. San Millán de la Cogolla: Cilengua, 33–71.

CE = Davies, Mark: *Corpus del español*. <http://www.corpusdelespanol.org> [último acceso: 10/10/2017].

Company Company, Concepción (2012): «El español del siglo XVIII. Un parteaguas lingüístico entre México y España», en M.ª Teresa García-Godoy (ed.), *El español del siglo* XVIII. *Cambios diacrónicos en el primer español moderno*. Berna: Peter Lang, 255–291.

CORDE = Real Academia Española: Banco de datos (*CORDE*) [en línea]. *Corpus diacrónico del español*. <http://www.rae.es> [último acceso: 10/10/2017].

CORDEREGRA = Calderón Campos, Miguel/M.ª Teresa García Godoy (dirs.): *Corpus diacrónico del español del Reino de Granada*. 1492–1833. <http://www.corderegra.es> [último acceso: 10/10/2017].

CorLexIn = Morala Rodríguez, José R. (dir.): *Corpus Léxico de Inventarios (CorLexIn)*. <http://web.frl.es/CORLEXIN.html> [último acceso: 29/04/2017].

Corriente, Federico (1999): *Diccionario de arabismos y voces afines en iberorromance*. Madrid: Gredos.

DA = Asociación de Academias de la Lengua Española (2010): *Diccionario de americanismos*. Lima: Santillana.

DECH = Corominas, Joan/José Antonio Pascual (1980–1991): *Diccionario crítico etimológico castellano e hispánico*, 6 vols. Madrid: Gredos.

DHECan = Corrales, Cristóbal/Dolores Corbella (2013): *Diccionario Histórico del Español de Canarias*, 2 vols. La Laguna: Instituto de Estudios Canarios.

DHLE = Real Academia Española (1960–1996): *Diccionario histórico de la lengua española*. Madrid: RAE.

Espinar Moreno, Manuel/María Espinar Jiménez (2015): «Bernardino Benalbara y su testamento: notas sobre alimentación en Baza», *Estudios sobre Patrimonio, Cultura y Ciencias Medievales* 17, 55–82.

García de Cabañas, Mª. Jesús (1967): *Vocabulario de la Alta Alpujarra*. Madrid: CSIC, Anejos del BRAE XIV.

García Soriano, Justo (1932): *Vocabulario del dialecto murciano: con un estudio preliminar y un apéndice de documentos regionales*. Madrid: C. Bermejo, impresor.

Gómez Ortín, Francisco (1991): *Vocabulario del noroeste murciano. Contribución lexicográfica al español de Murcia*. Murcia: Editora regional de Murcia.

Herrera, Mª. Teresa (1996): *Diccionario español de textos médicos antiguos*. Madrid: Arco/Libros.

Idáñez de Aguilar, Alejandro F. (2015): *Léxico de la región prebética. Límites del lenguaje andaluz y del murciano*. Murcia: Editum.

Kasten, Lloyd A./John J. Nitti (2002): *Diccionario de la prosa castellana del Rey Alfonso X*, 3 vols. Nueva York: The Hispanic Seminary of Medieval Studies.

Lope Blanch, Juan M. (1971): «El léxico de la zona maya en el marco de la dialectología mexicana», *Nueva Revista de Filología Hispánica* 20, 1, 1–63.

Maíllo Salgado, Felipe (1998): *Los arabismos del castellano en la baja Edad Media: consideraciones históricas y filológicas*. Salamanca: Ediciones Universidad de Salamanca.

Martín Morales, Francisco Manuel (2016): *Glosario del ajuar doméstico en la Sevilla de Velázquez*. <http://www.academia.edu> [último acceso: 29/09/2017].

Martínez, Martín (1728): *Anatomía completa del hombre*. Madrid: Imprenta de Bernardo Peralta.

Martínez de Mazas, José (1794): *Retrato al natural de la ciudad y término de Jaén*. Jaén: Imprenta de D. Pedro de Doblas.

Matarín Guil, Manuel F./Antonio Ruescas Granados (2000): «La vida cotidiana en los inicios del siglo XVII. El espacio privado. El caso de la taha de Boloduy», en Valeriano Sánchez Ramos (ed.), *El reino de Granada en el siglo XVII*. Almería: Instituto de Estudios Almerienses, 159–175.

Miñano y Bedoya, Sebastián (1826–29): *Diccionario geográfico-estadístico de España y Portugal*, tomo VII. Madrid: Pierrat-Peralta.

Morala Rodríguez, José Ramón (2014): «El *CorLexIn*, un corpus para el estudio del léxico histórico y dialectal del Siglo de Oro», *Scriptum Digital* 3, 5–28.

Morala Rodríguez, José Ramón (2015): «Los inventarios de bienes y el léxico del siglo XVII en el Archivo Histórico Provincial de Cádiz», en Teresa Bastarín y M.ª del Mar Barrientos (eds.), *Lengua y cultura en el Archivo Histórico Provincial de Cádiz*. Cádiz: Universidad de Cádiz, 137–174.

Morala Rodríguez, José Ramón (2017): «Voces derivadas en documentación notarial del siglo XVII», *Cuadernos del Instituto Historia de la Lengua* 10, 135–163.

Moya Torres y Velasco, Francisco Máximo de (1730): *Manifiesto universal de los males envejecidos que España padece*. Madrid: Librería de Francisco Laso.

Navarro Gavilán, Blanca (2014): *La sociedad media e inferior en Córdoba durante el siglo XV. Familia y vida cotidiana*. Córdoba: Universidad de Córdoba. Tesis doctoral.

Nebrija, Elio Antonio de (1495?/1989): *Vocabulario español-latino*. Madrid: Arco/Libros.

NTLLE = Real Academia Española (2001): *Nuevo tesoro lexicográfico de la lengua española*. <http://ntlle.rae.es/ntlle/SrvltGUISalirNtlle> [último acceso: 10/10/2017].

OED = *Oxford English Dictionary Online*. Oxford: Oxford University Press. <www.oed.com> [último acceso: 10/10/2017].

OLD = Glare, P. G. W. (1982): *Oxford Latin Dictionary*. Oxford: Oxford University Press.

Oudin, César (1607): *Tesoro de las dos lenguas francesa y española. Thresor des deux langues françoise et espagnolle*. París: Marc Orry.

Pagés, Aniceto de (1902): *Gran diccionario de la lengua castellana, autorizado con ejemplos de buenos escritores antiguos y modernos*. Madrid: Sucesores de Rivadeneyra.

Rosal, Francisco del (1611): *Origen y etymología de todos los vocablos originales de la Lengua Castellana. Obra inédita […] copiada y puesta en claro […] por el P. Fr. Miguel Zorita de Jesús María*. Ms. 6929 (s. XVIII) de la BNE.

Sánchez-Prieto Borja, Pedro (2011): *La edición de textos españoles medievales y clásicos. Criterios de presentación gráfica*. San Millán de la Cogolla: Cilengua.

Sanz, Juan Carlos/Rosa Gallego (2001): *Diccionario del color*. Madrid: Akal.

Sobrino, Francisco (1705): *Diccionario nuevo de las lenguas española y francesa*. Bruselas: Francisco Foppens.

Terreros y Pando, Esteban de (1787): *Diccionario castellano con las voces de ciencias y artes y sus correspondientes en las tres lenguas francesa, latina e italiana*. Madrid: Viuda de Ibarra.

TLHA = Alvar Ezquerra, Manuel (2000): *Tesoro léxico de las hablas andaluzas*. Madrid: Arco/Libros.

Torres Martínez, Marta (2013): «De la vida doméstica en la ciudad de Jaén: léxico documentado en la carta de dote de Ana María de Morales (1791)», en Manuel Cabrera Espinosa y Juan Antonio López Cordero (eds.), *V Congreso virtual sobre historia de las mujeres*. Jaén: Archivo Histórico Diocesano de Jaén.

Torres Montes, Francisco (2004): *Nombres y usos tradicionales de las plantas silvestres en Almería (estudio lingüístico y etnográfico)*. Almería: Diputación de Almería e Instituto de Estudios Almerienses.

Zerolo, Elías (1895): *Diccionario enciclopédico de la lengua castellana*. París: Garnier hermanos.

Pedro Sánchez-Prieto Borja y Delfina Vázquez Balonga

El léxico en los documentos de la Comunidad de Madrid (ss. XVI–XIX)[1]

Resumen: El propósito de este trabajo es examinar la documentación conservada en los archivos madrileños entre los siglos XVI y XIX para obtener el léxico de la vida cotidiana más significativo. De este modo es posible la comparación interna del vocabulario en la Comunidad de Madrid y también con las zonas geográficas próximas a través de documentos de esta época. Por otro lado, analizamos la pervivencia de este léxico por la comparación con los atlas lingüísticos sincrónicos de Madrid y Castilla-La Mancha. Esta labor se lleva a cabo en el proyecto *ALDICAM-CM*, cuya finalidad principal es elaborar un atlas lingüístico diacrónico de la Comunidad de Madrid.

Palabras clave: Léxico, Dialectología diacrónica, Documento archivístico, Comunidad de Madrid

Abstract: This paper's main aim is to examine archival documents from Madrid region from the 16th to the 19th century in order to obtain the most widely known vocabulary of everyday life. This way it will be possible to compare vocabulary from Madrid region and its next areas through contemporary documents. We will also analyse continuity of this vocabulary by comparing the synchronic linguistic atlas of Madrid region and Castilla-La Mancha. This work is carried out within the project *ALDICAM-CM*, whose main goal is to make a diachronic linguistic atlas of the Madrid region.

Keywords: Vocabulary, Diachronic Dialectology, Archival Document, Madrid Region

1 Estudios sobre el habla de la CM

Los usos lingüísticos madrileños han sido objeto de numerosas referencias tanto en el ámbito popular como entre los autores literarios. Un antecedente del reflejo de la lengua castiza madrileña se encuentra en el teatro breve de Don Ramón de la Cruz; así, en *Los bandos de Lavapiés* (h. 1800), el autor introduce algunos usos vulgares, como la vacilación de vocales átonas (*nenguno* 'ninguno', *empruperios* 'improperios'), pérdida de [-d-] intervocálica (*toítos*), la inserción de palatal

1 Este trabajo ha sido realizado dentro del proyecto «Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid» (*ALDICAM-CM*) (S2015/HUM-3443), financiado por la Comunidad de Madrid y el Fondo Social Europeo.

antihiática (*friyendo* 'friendo'), palatalización de -DY- (*presillos* 'presidios'), además de formas seguramente arraigadas entonces (*Madril* 'Madrid'). Más tarde, ha sido seguramente Galdós el mejor observador de ciertos rasgos que, sin ser exclusivos de Madrid, podrían haber estado presentes en las clases populares del siglo XIX. Conocida es la caracterización lingüística de Fortunata, que resulta esencial para fijar al personaje en contraste con su oponente, Jacinta (Russell 1971), y que puede sintetizarse en la expresión «pa chasco»[2]. En la misma novela, hay personajes como Mauricia la Dura o Izquierdo «Platón», que también reflejan el habla madrileña de los estratos más humildes frente a los representantes de la burguesía (Andreu 1983). Pero quizá el texto más expresivo sea el discurso de «Pujitos» en los *Episodios Nacionales*[3], seguramente madrileño, en el que se incluyen algunos rasgos tan sorprendentes como una supuesta aspiración de [s-] inicial (*jeñores*), junto a otros más esperables: *güenos*, *golvimos*, *menistro*, *dimonios*, *vendío*, *puebro*, *juera* 'fuera'[4]. En el teatro cómico del siglo XX volvemos a encontrar caracterizaciones lingüísticamente marcadas (Seco 1970). Por ejemplo, en *El chico de las Peñuelas o No hay mal como la envidia,* de Carlos Arniches (1915), el personaje de Sole elimina las consonantes implosivas (*salú* 'salud'), omite sílabas (*pa* 'para', *tié* 'tiene'), y reduce el diptongo de algunas palabras (*pos* 'pues'), además de usar opciones léxicas coloquiales (*cacho de novio*) (*Teatro completo*, I, 1177).

Pero ha sido solo en los últimos años cuando ha habido un interés científico por el estudio del habla de Madrid, y esto en dos ámbitos paralelos; por un lado, se han desarrollado estudios sociolingüísticos que han puesto el foco en la complejidad poblacional, debida a la inmigración desde otras regiones españolas, mientras que últimamente se ha encuestado a la población extranjera[5] (Cestero, Molina y Paredes 2015). Desde este punto de vista, ha tenido un importante

2 Tampoco se puede pasar por alto el significativo fragmento en el que Fortunata es corregida en su manera de hablar: «No se dice *diferiencia*, sino *diferencia*. No se dice *Jacometrenzo*, ni *Espiritui Santo*, ni *indilugencias*. Además *escamón* y *escamarse* son palabras muy feas, y llamar *tiologías* a todo lo que no se entiende es una barbaridad. Repetir a cada instante *pa chasco* es costumbre ordinaria». También se cita que «las eses se le convertían en jotas y se comía muchas sílabas» (Andreu 1983).

3 Biblioteca Castro I, 421.

4 El madrileñismo del personaje parece avalado por su definición como «un majo decente… personaje sublimado por el oficio de obra prima, el de carpintero y el de platero» (*ib*. 419).

5 Puede citarse el proyecto INMIGRA, coordinado por Florentino Paredes en la Universidad de Alcalá. Sobre los estudios recientes de sociolingüística de la población inmigrante, mencionaremos Sancho Pascual (2015), entre otros.

papel el trabajo de campo del proyecto *PRESEEA* y sus consiguientes investigaciones (Moreno Fernández 2015). Se han publicado trabajos sobre la atenuación (Molina Martos 2015 y Cestero Mancera 2015), las funciones del diminutivo y el pronombre personal (Paredes García 2015a y 2015b), las locuciones verbales (Penadés Martínez 2016) y la intensificación (Molina Martos 2010), por mencionar algunos ejemplos notables.

Otro ámbito de estudio ha sido el del Madrid rural, es decir, las poblaciones que todavía en el siglo XX han conservado modos de vida tradicionales (agricultura, ganadería, silvicultura, artesanía). En esta línea, aunque no exclusivamente, se sitúa el *Diccionario de madrileñismos* de Alvar Ezquerra (2011a), basado en trabajos previos que recogen voces de pueblos y ciudades de la Comunidad de Madrid, como el estudio sobre el habla del valle de Lozoya (Sacristán 1990)[6]. Con una metodología rigurosa se han llevado a cabo encuestas que han conducido al *Atlas Dialectal de Madrid* (*ADiM*), de 2015, en el que se cartografían los datos obtenidos en las localidades que se sitúan en la periferia de la Comunidad, por considerar que en ellas la influencia del habla de la urbe es menor. Se han publicado también 16 «etnotextos» de diferentes pueblos, junto con el estudio de algunos mapas representativos que muestran la diversidad lingüística, sobre todo léxica, de la provincia (García Mouton y Molina Martos 2017).

Para épocas pasadas se ha llevado a cabo algún trabajo, pero casi nunca a partir de textos anclados geográficamente, es decir, de documentos con lugar de emisión explícito, con excepción de los estudios sobre el *Fuero de Madrid* (h. 1200; Millares Carlo *et al.* 1932); más recientes son los del *Fuero de Alcalá*, anterior a 1247 (Torrens Álvarez 2002). Diversos volúmenes de la serie *Textos para la Historia del Español* recogen documentación madrileña: III, sobre el Archivo Municipal de Alcalá de Henares, con fondos de 1252 a 1684 (Paredes García 2005)[7]; IV, con fondos de la Inclusa de Madrid conservados en el Archivo Regional de la Comunidad de Madrid, de entre 1590 y 1662 (Sánchez-Prieto Borja y Flores Ramírez 2006); V, sobre documentos del Archivo Municipal de Daganzo de Arriba, entre 1584 y 1649 (Paredes García 2010), y VIII, de Arganda del Rey, de 1579 a 1691 (Vázquez Balonga 2014). También citamos el material disponible en línea del *Corpus Léxico de Inventarios* (*CorLexIn*, <http://web.frl.es/CORLEXIN.html>), con documentos de archivo del siglo XVII emitidos en la ciudad de Madrid, pero también en localidades como Pinto y Torrelaguna. En cuanto a trabajos de investigación concreta con documentos de archivo de

6 Sobre las voces propias de Madrid, v. Alvar Ezquerra 2011b.

7 Aunque los más antiguos no serían emitidos sino por el arzobispo de Toledo.

la Comunidad de Madrid, mencionamos a Paredes García (2003), Gómez Seibane y Vázquez Balonga (2013), Vázquez Balonga (2014, 2015) Morala (2015), Vázquez Balonga y Sánchez-Prieto Borja (2015, 2017), Sánchez-Prieto Borja y Vázquez Balonga (2018).

2 El proyecto *ALDICAM* y los archivos municipales

A pesar de los trabajos previos, faltaba para Madrid y su tierra una labor de recopilación exhaustiva de sus fondos que atendiera tanto a la escasa documentación medieval como a la mucho más abundante de los siglos XVI y XVII, y más aún del XVIII y XIX, para poder trazar así toda la trayectoria histórica de la variedad madrileña, y enlazar de este modo con los estudios sobre las hablas actuales. Ello se justifica muy en particular por el peso que suponemos que tuvo la ciudad de Madrid en la conformación del español moderno. La ocasión surgió gracias a un proyecto financiado por la Comunidad de Madrid y llevado a cabo por investigadores de varios centros madrileños[8]. El objetivo es seleccionar, transcribir, editar digitalmente y poner a disposición de investigadores e interesados en general un amplio corpus documental que permita consultas dinámicas de formas y palabras respecto de diferentes parámetros y, sobre todo, que proyecte a mapa de manera inmediata cualquier elemento buscado, y ello mediante una tecnología que ya está disponible desde 2015 en el corpus *CODEA*[9]. Dentro de este proyecto, se ha seleccionado hasta ahora documentos de los siglos XIII al XIX de Alcalá de Henares, Alcorcón, Aranjuez, Arganda, Arroyomolinos, Cadalso de los Vidrios, Camarma de Esteruelas, Chinchón, Colmenar de Oreja, Colmenar Viejo, Coslada, Daganzo de Arriba, El Escorial, Getafe, Guadarrama, Hoyo de Manzanares, Madrid, Moralzarzal, Navalcarnero, Parla, Titulcia, Valdemoro, Rascafría, Robledo de Chavela, San Lorenzo de El Escorial y Torrejón de Ardoz, entre otros.

3 Los documentos de municipios madrileños en *ALDICAM*: tipología y materias

Para nuestro propósito actual, el estudio léxico de localidades de la Comunidad de Madrid, nos serviremos de piezas transcritas y revisadas de varios municipios madrileños, sobre todo de Arganda del Rey, Camarma de Esteruelas, Daganzo

8 Forman parte del proyecto *ALDICAM-CM* (S2015/HUM 3443) las universidades de Alcalá (grupo coordinador) y Complutense de Madrid, y el Consejo Superior de Investigaciones Científicas (ILLA-CSIC) <http://aldicam.blogspot.com.es/>.

9 <http://www.corpuscodea.es>.

de Arriba, Hoyo de Manzanares y El Escorial y Valdemoro. Limitaremos el examen a documentos datados entre los siglos XVI y XVIII, ambos inclusive. Dentro de la gran variedad de tipos y asuntos tratados en las piezas conservadas en los archivos madrileños, especialmente los de la capital, nos hemos circunscrito a aquellos que podían ofrecer un vocabulario más rico y variado, así como una mayor diferencia de registro (v.i. 4.). La documentación municipal de Arganda del Rey abarca desde los siglos XVI a finales del siglo XVIII. En la parte primera de los siglos XVI y XVII, hay principalmente documentos de tipo notarial: ventas, trueques, inventarios *post-mortem* y dotes, testamentos y otras escrituras. Podemos destacar el memorial de los vecinos de Arganda acerca del uso de la ribera del Jarama (1584), el testamento de Juan Blanco (1607), bienes de los hijos de Custodio Sanz (1677), dote de Ana Mejorada (1687), tutela de José Milano (1690) y la escritura para dorar un retablo (1691) (Vázquez Balonga 2014). Para el léxico, sobre todo tienen un papel importante los inventarios de bienes y los testamentos. En el siglo XVIII, en cambio, se trata en la mayor parte de los casos de expedientes judiciales elaborados por la justicia municipal de la propia localidad (Mañas García 2017): por cazar con galgos (1775), por el robo de una bayoneta (1779), por la venta conjunta de desperdicios y carnes (1785), por insultos (1791), por el robo de una potra castaña (1792), por deudas por comprar ropa (1796).

Otro tanto cabe decir de las piezas procedentes de Camarma de Esteruelas (Gallardo López 2017), entre 1754 y 1765; se trata en gran parte de los casos de declaraciones de testigos, que tienen interés como reflejo, siquiera parcial, de la lengua de los declarantes. Por ejemplo, por haber encontrado una vaca en un plantío de árboles (1754), por intento de violación a una mujer (1754), por lesiones a un carretillero que había entrado en un campo de trigo (1754), por acusaciones de fraude contra una vendedora, por robo de trigo (1759) o paja (1760), o de una pollina (1765), o por dejar al ganado comer en un campo de centeno (1761).

Los documentos de Daganzo de Arriba provienen de la edición y estudio publicados por el equipo GITHE (Paredes García 2010). Son 47 piezas escritas entre los siglos XVI y XVII. Otra parte está datada en los siglos XVII (un documento) y XVIII (dos documentos), inserta en el corpus *ALDICAM*. En este fondo predominan documentos notariales relacionados con traspasos de bienes. Tenemos, por ejemplo, diferentes piezas del proceso de las curadurías de Juan Gordo (1587), Andrés Martín (1588) y Alonso Bartolomé de Soria (1633), la petición de soldadas de Francisco Gasco (1588), un proceso de acreedores (1589–1592) y diversas almonedas, testamentos e inventarios de bienes (1591–1594), además de una carta de reclamación de deuda (1598); los últimos expedientes son un nombramiento de tutores (1637) y una escritura de poder (1649).

Tiene gran interés la documentación del Archivo Municipal de El Escorial, población más antigua que la de San Lorenzo, situada junto al monasterio jerónimo fundado por Felipe II. Conserva abundantes piezas desde el siglo XVI al XVIII, y son de particular interés la de justicia civil y, sobre todo, criminal. Muchos procesos tienen que ver con la caza y pesca furtiva, así como con el aprovechamiento ilegal de pastos y montes reales. No serían pocos quienes aprovechaban los recursos del bosque, y así, por ejemplo, se abrió un proceso contra unos hombres acusados de ir a cazar chivos (1763), y otros que tenían sebo de gamo y ciervo (1770). Además, hay procesos sobre los asuntos más variopintos: mandato que prohíbe apedrear perros (1534), devolución de un rocín con lesiones (1587), denuncia y averiguación por vender aceite en mal estado (1587), pleito por un sangrado mal hecho a un caballo (1585), denuncia a Monedero, pastelero, por deudas (1588), incumplimiento de promesa de matrimonio (1699), protesta de un vecino por una cacera (1700), denuncia a un albañil por hacer mal un horno de obra, proceso por bestialismo (1723), denuncia por heridas a una vaca (1769–1770), proceso judicial por la aparición de dos pasquines injuriosos (1771), atropello de un animal de carga por un carruaje, entre otros muchos.

El Archivo del Monasterio de El Escorial conserva un importante fondo (Mediavilla Martín 2010), del que han sido transcritas 28 piezas relativas a las obras de construcción, ampliación y reparación, así como a administración económica del monasterio (Agujetas Ortiz 2017): memoria de los precios de herramientas (1562 y 1567), requisitos para hacer las paredes del plantel de la Fresneda (1563), encargo de ladrillos (1569 y 1631), gastos de fábrica del monasterio (1596)[10], defectos en el chapado y solado de la escalera del panteón (1640), carta privada sobre los daños de un incendio (1673), reedificación tras incendio (1677), carta privada sobre caída de un rayo en el cimborrio (1679), carta del prior sobre un incendio por rayo, carta privada sobre estragos causado por varios lobos (1771).

Del Archivo Municipal de Hoyo de Manzanares hemos seleccionado, entre otros, un acuerdo sobre la leña (1673), sobre la subida del trigo (1673), varios testamentos (1679 1700, 1706) e inventarios de bienes (1704, 1707), una carta de dote (1680), una postura sobre la extracción de corteza para carbón (1682), las condiciones de los *fabriqueros* de carbón (1682), un acuerdo sobre corta de chaparros y enebros del monte (1704), acuerdos e informes sobre uso de caminos (1752) y petición para hacer unas cercas (1752)[11].

10 Aunque suele señalarse 1584 como fecha de conclusión de las obras, lo cierto es que los remates, arreglos y ampliaciones fueron constantes.

11 Fuera del límite temporal marcado, encontramos un acuerdo sobre la extracción de la *chavasca* ‘leña menuda‘ en la ladera del Ejido (1857).

En el Archivo Municipal de Valdemoro destacamos diferentes acuerdos sobre venta de pan, viñas y vendimias (1656), inventario de majuelos (*id.*), y otro acuerdo sobre el agostadero (1657), petición para hacer la fiesta de toros (1674), denuncia de la fuga de un muchacho, criado de un sargento (1712), instrucciones sobre cómo construir la barca del Jarama (1793), un apeo de las lindes (1777) y una lista de materiales y maderas de una obra (1787).

Se comprenderá que ante tal variedad de asuntos tratados, el vocabulario ofrecerá una notable riqueza léxica en diversos ámbitos, sobre todo en los de referente material[12].

4 La lengua de los documentos de municipios de la CM. Caracterización general

El estudio histórico de la lengua de las localidades de Madrid en el pasado está, lógicamente, condicionado por los materiales textuales disponibles y por el estado de la edición de los fondos. En cuanto al primer aspecto, hay que señalar que muchos pueblos de la tierra de Madrid[13] conservan solo documentos tardíos, como Torrejón de Ardoz o Parla, mientras que El Escorial cuenta con amplios fondos desde el siglo XVI. Tampoco la variedad registral de los documentos y la diversidad sociolingüística que transparentan es tan amplia en estas localidades como en la ciudad de Madrid (Sánchez-Prieto/Vázquez Balonga 2017). En cuanto al segundo aspecto, se han transcrito hasta ahora documentos anteriores al siglo XIX, sobre todo, de Alcalá de Henares[14]. Arganda del Rey, Camarma de Esteruelas, Hoyo de Manzanares, Daganzo de Arriba, El Escorial y Valdemoro.

Señalaremos una serie de rasgos que, sin ser exclusivos de esta zona, permiten establecer, en algunos casos, características lingüísticas de la Comunidad de Madrid. En el vocalismo notamos la diptongación en el sustantivo postverbal

12 En lo que sigue, los documentos son citados por el lugar y año de emisión. Puede verse la tabla completa de documentos transcritos para el proyecto *ALDICAM* en <http://aldicam.blogspot.com.es/>. Aparte de estos, hemos citado algunas piezas en proceso de transcripción y edición.

13 Se ha de notar que la unidad territorial que se atribuye modernamente a la Comunidad de Madrid está justificada históricamente por el concepto de «tierra de Madrid», que viene a corresponder de manera bastante aproximada a la provincia (Jiménez Rayado 2010).

14 Por su situación especial como ciudad universitaria y centro político, religioso y cultural, no hemos considerado para este estudio la documentación de Alcalá de Henares, ya que se aleja del carácter rural de las otras localidades. Puede verse en la obra ya citada de Paredes (2005).

entriega 'entrega' (Corpa, 1765). La vacilación de átonas /e/-/i/ es constante, tanto en los ss. XVI y XVII como a lo largo de todo el XVIII: *difinir* (Daganzo de Arriba, 1589), *pidido* (Daganzo de Arriba, 1594), *lezencia, defunta, metad* (Hoyo de Manzanares, 521, 1704), *bessitar* (El Escorial, 1771), *dilinquido* (Arganda del Rey, 1798). También hay numerosos ejemplos de /o/-/u/: *Úrsola, sepoltura* (Arganda del Rey, 1607), *groñía* (El Escorial, 1771), *rigoroso* (Arganda del Rey, 1784). Aunque no tan común, tampoco falta la oscilación entre /a/-/e/ átonas (*celedores,* Valdemoro, 494, 1756; *tabarnero,* El Escorial, 1771).

En el consonantismo, todavía en el siglo XVIII aparece *f-* en *fecho* (Arganda del Rey, 1775 y 1789); por contra, *hunda* 'funda de almohada' se encuentra en Daganzo de Arriba en 1588. Se confunde la labial y la velar en *frígola* 'frívola' (Camarma de Esteruelas, 1754). El grupo *-pt-* debía de pronunciarse reducido (*setiembre*, 1797). La labial intervocálica se conserva en *toballas* 'toallas' (Daganzo de Arriba, 1594), variante muy extendida en el español de los siglos de Oro[15]. Ya antes parece faltar una conciencia de la composición de los grupos consonáticos, y en 1587 vemos *agciones* 'acciones' (Daganzo de Arriba, *CODEA* 1779). En Daganzo de Arriba (1568) se sigue dando la forma *amos* 'ambos' («amos a dos») en 1568 (*CODEA* 1778). La palabra procedente de SALICEM 'sauce' aparece en su forma reducida *saz* (Daganzo de Arriba, 1591). Entre los resultados de VENTULU y VENTILARE, retenemos *benlo* 'bieldo', y el verbo *albelar* (Daganzo de Arriba 1588 y 1594, respectivamente). De POPULUS procede *pobo* 'chopo' (Daganzo de Arriba, 1591; Paredes García 2010: 36). En las sibilantes, no se encuentran fenómenos de indistinción /s/y/θ/, pero sí algunos posibles casos de asimilación, como *zencillo/a* en Hoyo de Manzanares a principios del siglo XVIII («de pared *zencilla*», 521, 1704). Igualmente, se registra la anticipación, seguramente común en la lengua hablada («plaça *plública*», Valdemoro, 491, 1674). En un documento de nivel excepcionalmente bajo se observa confusión de líquidas, sobre todo cuando hay otra próxima: *artar,* 'altar', *er* 'el' (El Escorial, 1771)[16].

Hay alternancia entre el verbo con y sin prefijo en *apastar* (Camarma de Esteruelas 1754, «havía *apastando* dos reses vacunas»), y *desapartar* 'apartar' (Camarma de Esteruelas 1760, «y el confesante los *desapartó* y puso en paz»).

En cuanto al diminutivo, alterna la forma *-illo* con *-ito*. En los siglos XVI y XVII parece dominar más *-illo*, en consonancia con el uso general de la época

15 En el citado *CorLexIn*, *toballa* se documenta en el siglo XVII en provincias tan distantes como Huesca, Córdoba, Murcia, Ciudad Real, Segovia, Navarra y Guipúzcoa.

16 Este fenómeno es muy frecuente en las notas de abandono de expósitos escritas con mano inhábil para la Hermandad del Refugio, incluso avanzado el siglo XIX, por lo que puede considerarse un rasgo del habla popular de la región de Madrid.

(Náñez Fernández 1973): *arquilla, sartenilla* (Daganzo de Arriba, 1594); *mesilla* (Arganda del Rey, 1691). Tampoco falta en el siglo siguiente: *juzguillo, perrilla* (Arganda del Rey, 1775). No obstante, hay ejemplos de *-ito* en los siglos XVI a XVIII: *arquita, sillita, taleguita* (Daganzo de Arriba, 1594), *pradito*, *Gamita* (Hoyo de Manzanares, 1704), *barreñita* (Hoyo de Manzanares, 1702). La forma con *-ico*, si bien se considera arcaica y aparece más en documentos del siglo XVI (*açadoncico* en Daganzo de Arriba, 1588), también debió pervivir más tiempo en la lengua coloquial, a juzgar por el uso de *dobloncico* en un documento escrito por mano inhábil y en tono informal en El Escorial en 1771. No hay que olvidar, por otro lado, la presencia de numerosos diminutivos lexicalizados a partir de los siglos XVI y XVII en localidades como Arganda y Daganzo de Arriba (*ropilla, calcetas, rastrillo*) (Miguel Franco 2010: 42–43 y Vázquez Balonga 2014: 20).

Sin salir del nivel morfológico, hay que destacar la existencia en los siglos XVI y XVII del neutro de materia, localizado en Daganzo de Arriba (*lana blanco, lana prieto*) (Miguel Franco 2010: 47) y Arganda del Rey (*plata falso, plata barniçado*) (Vázquez Balonga 2014: 21). Este fenómeno de concordancia aparentemente masculina para sustantivos femeninos incontables ha tenido un largo seguimiento en la dialectología sincrónica (Fernández-Ordóñez 2006 y 2007) y se localiza, sobre todo, en zonas más septentrionales de la Península. Sin embargo, la presencia en lugares ubicados en el centro-este de la Comunidad indica una isoglosa diferente, al menos hasta el siglo XVIII (Gómez Seibane/Vázquez Balonga 2013).

Como forma de tratamiento, la estructura más compleja se aplica al rey: «Valga para el reinado de su magestad el señor don Carlos» (Camarma de Esteruelas, 1760). Encontramos «*señor*» ante nombre para una autoridad (Camarma de Esteruelas, 1754, «ante el señor Andrés de Lucas, alcalde ordinario de esta villa»), «ante dicho señor juez» (Camarma de Esteruelas, 1756); *su merced* parece referirse al alcalde ordinario (Camarma de Esteruelas, 1754, 1760). Para el uso más general parece que se prefería la forma *usted*, como cuando se trata a Ana Isidro, vendedora (Camarma de Esteruelas, 1756), o *ustedes*, propio de un nivel coloquial para las autoridades locales de El Escorial en un pasquín de protesta (El Escorial, 1771). En este mismo documento se puede ver una muestra de *vosotros,* para «los pobres de El Sitio».

En lo que concierne al pronombre átono objeto, el leísmo con referente animado es general desde el siglo XVI: «cuando algún perro parece que anda por el dicho sitio los peones e algunos oficiales le tiran muchas piedras y le dan mucha grita hasta que *le* matan» (El Escorial, 1534); «introduciéndose en el cuarto y dormitorio, en el que *le* hallé [a Lorenzo Garrido]» (Camarma de Esteruelas 1754). También se emplea *le* con referente cosa: «sin *le* revocar [el testamento]»

(Arganda del Rey, 1677), «rezivió juramento por Dios nuestro señor y una señal de cruz en forma de derecho, el que *le* hizo como se requiere [el juramento]» (Camarma de Esteruelas, 1756). En cambio, con referente no contable se adopta *lo:* «el azeite que Mena bende (…) se sufre *benderlo*, (…) *calentarlo* en una sartén (…) *lo* viesen *calentallo*» (El Escorial, 1587). Podemos concluir, por tanto, que el sistema referencial estaba extendido en Madrid desde el siglo XVI (Sánchez-Prieto Borja y Vázquez Balonga, 2018)[17].

En cuanto al laísmo, se documenta temprano: «Requirió a Alonso Monedero de Fraça (…) con la dicha María Díez sobre los treinta e cuatro reales que le pide, el cual dixo que solamente *la* debe veinte e un reales y cuatro libras» (471, El Escorial, 1588). Sin embargo, no parece que se extendiera al unísono en la Comunidad de Madrid, y así en Arganda *la* todavía compite en desventaja con *le* en la segunda mitad del siglo XVII, pero en el XVIII parece ya generalizado[18] (Sánchez-Prieto Borja y Vázquez Balonga, 2018). En Camarma y Hoyo de Manzanares lo documentamos en el siglo XVIII: «la metad que *la* á tocado a la dicha Ana Martín» (Hoyo de Manzanares, 521, 1704); «que haviendo pasado la que declara el día siete del presente mes y año de la fecha a la casa de Ana Isidro Baldavero, vezina de esta billa, con motivo de que *la* vendiese un cuartillo de leche» (Camarma de Esteruelas, 1756).

La actualización del sustantivo ha alcanzado ya índices comparables a los de hoy, e incluso el artículo aparece en secuencias que luego se fijaron sin él, como *con motivo:* «con *el motivo* de estar el testigo en su era guardando sus granos» (Camarma de Esteruelas, 1759).

La forma del adverbio es *ansí* en 183 ocasiones en la Comunidad de Madrid, mientras que *as*(*s*)*í* aparece 215, lo que indica que la primera no estuvo tan lejos de superar a la general. En la documentación de Daganzo de Arriba incluida en *CODEA*, encontramos *asín* 52 veces entre 1586 y 1666, mientras que *as*(*s*)*í* aparece solo 10, entre 1589 y 1782. Resulta curiosa la distribución de las variantes *ahora* y *aora* frente al tradicional *agora*, pues las innovaciones parecen concentrarse en la ciudad de Madrid, Alcalá de Henares y Arganda del Rey; en esta última población, durante la segunda mitad del siglo XVII, *agora* solo aparece 1 vez, por 4 de *a*(*h*)*ora*.

17 No obstante, tenemos ejemplos de referentes incontables con *le:* «que el pasto ni el restrojo no nos *le* hemos de llevar» (Camarma de Esteruelas, 1754).

18 Con todo, hay varios casos de laísmo singular en los documentos de Arganda en este período: «la traiga luto [a María Herranz]» (1677), «dándola el alimento» (1688); en contraste, podemos citar ejemplos como «le quedó un hijo [a María Sanz]» (1690) (Vázquez Balonga 2014: 20).

El lugar en donde puede indicarse con *a* («una tierra de fanega y media poco más o menos, *a* la raya de Torrebermeja», «una viña *al* camino de Pajares», Arganda del Rey, 1677; Camarma de Esteruelas 1754, «hallándose […] *a* la puerta de dicho señor alcalde»)[19]. Se emplea «a presencia» donde hoy diríamos «en presencia» (Camarma de Esteruelas 1754, «y *a presenzia* de algunas personas que havía en un corral inmediato a dicho paxar»).

La expresión «donde/do dicen» por 'llaman' es frecuente: «una tierra en término d'esta villa, *do dicen* los Cornicabros» (Daganzo de Arriba, 1594); «la cuarta parte de doce fanegas de tierra *donde dicen* Quiebracarros y la Machana» (Arganda del Rey, 1677); «*donde dizen* las Cañas» (Camarma de Esteruelas 1754). También aparece la forma «que llaman» en Hoyo de Manzanares («la zerca *que llaman* de la Cavilda», 1704) y en Arganda del Rey («el parage *que llaman* la Cabeza Gorda», 1775).

5 El léxico local de municipios madrileños. Entre la lengua administrativa y la lengua viva

Una aproximación al léxico de la documentación madrileña incluida en *ALDICAM* ha de empezar por intentar diferenciar registros lingüísticos dentro de esos mismos documentos, pues ello puede darnos una pista acerca de qué palabras se adscriben al estilo formal propio de los textos administrativos y judiciales y cuáles corresponden al uso general, o incluso al nivel sociolingüístico bajo. En las piezas de los expedientes judiciales resulta relativamente fácil delimitar las secciones formularias de las declaraciones de testigos. Y así oponemos segmentos como (1) y (2):

(1) ante vuestra merced en la vía y forma que más haya lugar en derecho y premisas sus solemnidades, me querello grave y criminalmente de Felipe Díaz, otro vezino de esta dicha villa, y expresando el caso que motiva mi justa querella con verdadero relato de él (Camarma de Esteruelas, 1754).

(2) Hombres, echen ustedes fuera con mil demonios esos bueyes, que no estamos en año de que ninguno se coma las haziendas de otros (Camarma de Esteruelas, 1754).

No ha de pensarse, sin embargo, que la distancia entre la lengua administrativa (y jurídica) y la general sea insalvable; en este sentido, no compartimos la adscripción que suele hacerse de esta a los lenguajes de especialidad, ni hoy ni en el pasado (Castellón 2001); al contrario, destacamos la gran permeabilidad de la lengua de uso respecto de la jurídica. P. ej., la palabra latina *interim* debió

19 El empleo de la preposición *a* para valor 'en donde' fue frecuente en toda la historia del español; encontramos ejemplos todavía en el siglo XVIII (Sánchez-Prieto Borja 2000).

extenderse en la lengua coloquial y popular: «y vuestra señoría me perdone el que le canse. *Ínterin* yo quedo rogando a Dios, guarde a vuestra señoría felices años» (*CODEA* 2305, Madrid 1754). Incluso se puede ver en los niveles más bajos, como en una nota de abandono de una expósita: «sus padres la esponen a la piadás [‘piedad’] de ustedes, *ínterin* Dios les mejora de fortuna» (Madrid, 1741)[20].

La contrapartida es que no todos los segmentos que se ponen en boca de testigos son reproducción literal de lo dicho por estos, pues las declaraciones se modifican para adaptarlas al decoro exigible en un auto judicial, o bien por la tradición de emplear palabras propias de la lengua jurídica o administrativa. Así, en la declaración de «tía Galinda», se pone en boca de esta, «*¿qué á de ser*?, que me están *quitando el crédito* diciendo que echo harina en la leche» (Camarma de Esteruelas, 1756), donde parece del todo realista «qué ha de ser», mientras que «quitar el crédito» es probablemente responsabilidad del escribano.

Otras veces el escribano da una interpretación muy ajustada de las palabras de los testigos, lo que añade verosimilitud de fidelidad: «y agarrándole sin resistencia alguna le dejó caer en el suelo y espresándole con ironía y cólera engañosa: Aora te boy a contar un *cuento*», principió a derramar patadas crueles en todo su cuerpo» (Arganda del Rey, 1784), lo que permite entender el uso de *cuento* en el contexto de la violencia que estaba a punto de aplicar al muchacho. También cabe adscribir al estilo conversacional *llevar* para ‘cobrar’: «Y solo por esto y la cebada que gastó con dos caballerías en aquella noche *le llebó* cincuenta reales», «Por solo medio día que asistió en dicha posada y hizo una comida *le llebó* dichó Lorán doscientos reales» (El Escorial, 1763)[21].

La relación de las vocablos y expresiones con un registro bajo es garantía de su autenticidad en el testimonio recogido, y así leemos: «se principió a leer el mandamiento de execución que encabeza Mateo Sancho y oído por el dicho Juan prorrumpió la expresión de que *se cagaba* en él» (Arganda del Rey, 1796)[22]. Precisamente en este segmento se puede apreciar el marcado contraste registral entre la lengua del notario (*prorrumpió la expresión*) y la de los testigos.

---

20 También la literatura del siglo XIX: «ansina se ha de hacer, que *ínterin* quel otro se guarda el dinero de la nación el puebro no come» (Episodios Nacionales, I, 421).

21 Este uso aparece en *Autoridades* (1734).

22 En el mismo corpus de *ALDICAM* hay otro ejemplo en El Escorial: «respondieron que se limpiaban el culo con ella [la vara] y que se cagaban en ella» (El Escorial, 1768). Documentamos «cagarse en alguien» ya en 1653: «Siempre que venga el trigo y la cebada, se recibirá, y la comerán los hombres y las bestias, y después la cagarán, como yo me cago en Vm. y en sus órdenes» (Juan Francisco Ustarroz, *Carta al maestro Gil González Dávila*, ed. A. Paz y Melià, Madrid, 1890).

Los insultos parecen recogidos con fidelidad por los escribanos, pues a veces son causa de querella. Así, encontramos *deslenguada* y *desvergonzada* (Camarma de Esteruelas, 1760). De uso bastante extendido en los siglos XVII y XVIII es *pícaro*: «a que volvió a dezir la testigo: *Pícaro*, ¿a qué vienes aquí?» (El Escorial, 1708), «la insultó con las expresiones de que era una *pícara* advenediza forastera» (Arganda del Rey, 1791). El *DLE* define esta palabra como «Listo, espabilado» o bien «tramposo y desvergonzado»; a tenor de los empleos encontrados parece de polaridad mucho más negativa[23]. Más dificultad causa saber a qué nivel pertenece *divertido* 'ocupado, entretenido': «pues aunque todos estavan allí inmediatos algunos de dichos gallegos estaban *divertidos* segando» (Camarma de Esteruelas, 1754).

6 Léxico de los documentos y clasificación temática

La clasificación del léxico histórico puede adoptar diferentes criterios que, lógicamente, habrán de adaptarse a la tipología textual examinada, aparte de los usos propios de la naturaleza administrativa y jurídica. Para facilitar el estudio, nos aproximamos al léxico con la clasificación del *Atlas Dialectal de Madrid* (2015), aunque el que se documenta y el vocabulario obtenido por encuesta no son totalmente equiparables. No obstante, para el entorno del hogar, complejo en algunos documentos, como los inventarios de bienes, tenemos en cuenta la clasificación pensada para este tipo de textos propuesta por Carriazo (2012).

6.1 El cuerpo humano y la persona

Este grupo referencial abarca no solo las partes del cuerpo, sino también sustantivos de agresiones corporales y nombres y adjetivos relativos a la persona. Por ello, donde más aparecen estos elementos es en las causas criminales por agresiones, muchas veces por intentos de abuso sexual o por peleas, llamadas a menudo *quimeras* (Camarma de Esteruelas, 1740, entre otros). Así, tenemos *mano* (Arganda del Rey, 1784), *espaldas* (Camarma de Esteruelas, 1754), *pescuezo, pie* (Camarma de Esteruelas, 1740), *sangre* (Arganda del Rey, 1791). Para describir las agresiones, aparecen *estocadas, cuchilladas* (El Escorial, 1668), *cachetes, manotada, puñaladas*[24] (El Escorial, 1708), *rempujón* (Camarma de Esteruelas,

23 En Galdós lo vemos aplicado a Godoy, al que se le llama también *monstruo* en los *Episodios Nacionales,* I.

24 Se documenta, en concreto, la forma fijada *coser a puñaladas:* «la havía de *coser a puñaladas*» (El Escorial, 1708). Igualmente se encuentra la expresión *moler a palos* (El Escorial, 1723).

1754), *golpes, patadas* (Arganda del Rey, 1784), *azotes* (Arganda del Rey, 1791), *empellones* (Camarma de Esteruelas, 1740; El Escorial, 1768), *palos* (El Escorial, 1723), así como lesiones: *cardenal* (El Escorial, 1708). Con referencia a las personas, documentamos *mozo* (El Escorial, 1723; Arganda del Rey, 1784), *chicuelo* (Arganda del Rey, 1784) y el tipo madrileño *majo* (*id*., 1796).

En cuanto a los calificativos de carácter, están los de tipo ofensivo: *bellaco* (El Escorial, 1587), *inocente* (El Escorial, 1723), *deslenguada, desvergonzados* (Camarma de Esteruelas, 1740), *traidor, tonto* y *alcagüetes* (El Escorial, 1771), *pícara, advenediza, forastera, ladrona, escandalosa* (Arganda del Rey, 1791), hasta llegar a otros más duros, registrados por necesidad de fidelidad en el proceso sobre un papel difamatario (*cabrón consentido* y *puta,* El Escorial, 1771). En este mismo documento, llama la atención el uso de *alcahueta* y *mujer de vodegón* para insultar a una mujer[25]. Hay que destacar también cómo se llega a apuntar el insulto en expresiones fijadas (*un ato de ladrones,* El Escorial, 1768). Por su parte, las referencias al físico se indican en algunas ocasiones: *mal encarado* (El Escorial, 1723), *moreno, descolorido* (Corpa, 1765), *desmelenada, perturbada* 'alterada' (Daganzo de Arriba, 1782).

6.2 Indumentaria

En el ámbito de la indumentaria, la fuente más completa es la de los inventarios de bienes con o sin tasación, aunque también otro tipo de documentación, como las denuncias, que incluyen descripciones de individuos. Por ejemplo, en dos casos de búsqueda de mozos huidos: *montera valenciana* y *gujón* (Corpa, 1765); *chupa, calzones, vestido militar* (Colmenar de Oreja, 1712). El término *gujón* será, seguramente, una mala interpretación de *jubón*, forma habitual en inventarios de bienes de los siglos XVII y XVIII (Arganda del Rey, 1688; Hoyo de Manzanares, 1706; Camarma de Esteruelas, 1765). Otras prendas que aparecen en documentos son *balona, bestido* (de hombre) (Arganda del Rey, 1688), *cabezones* (El Escorial, 1723), *casaca, capa, corbata, buelos, escusalí* 'delantal pequeño', *devantal, guardapiés, manteleta, capotillo, paletina* 'adorno que cubre el pecho', *puño, vasquiña, zapatos* (Camarma de Esteruelas, 1765). Los tejidos suelen ser los más comunes: *paño, sempiterna, lino, lienço, cáñamo* (Arganda del Rey, 1688), *tafetán, rasoliso, terciopelo, damasco, prinsipela* y *griseta* (Camarma de Esteruelas, 1765)*, damasquillo, seda, piñuela* (Hoyo de Manzanares, 1706),

25 Según *Autoridades* (1726), «parece que se ha criado en un bodegón» era una expresión general para designar «mala crianza», por lo que podría ser equivalente a 'mujer maleducada, baja' y, por extensión, 'liviana, promiscua'.

entre otros. Como complementos, podemos destacar *faldriquera* (Arganda del Rey, 1798) y *bolsillo* (El Escorial, 1771).

6.3 Vida doméstica

En este amplio grupo tenemos que hacer distinciones. En el ajuar de cocina, registramos un vocabulario abundante; en casi todos los inventarios y testamentos de los fondos estudiados aparecen *sartén, artesa, cazo, olla, asador, almirez, caldero* o *caldera.* Encontramos también *varillas* para cerner y *parrillas* (Arganda del Rey, 1691). Como recipientes, destaca *botijón* (Camarma de Esteruelas, 1754). Igualmente hay vajilla, llamada normalmente *vedriado.* No faltan alimentos («Productos perecederos»), como *queso* (Arganda del Rey, 1691), *tozino* y *manteca* (Hoyo de Manzanares, 1706), además de otros citados en los procesos judiciales: *harina, leche* (Camarma de Esteruelas, 1756), *pan, bacallado*[26] (El Escorial, 1763). En la parte de mobiliario y accesorios encontramos numerosas referencias a muebles comunes (*mesa, silla, cama, arca*). La madera para hacerlos suele ser de *pino* (Daganzo de Arriba, 1594; Hoyo de Manzanares, 1706), aunque también otros como *pobo* 'álamo blanco, chopo blanco', menos común (Daganzo de Arriba, 1591).

Para el grupo que hemos denominado «Decoración», podemos destacar la presencia de algunos elementos como el *guadamacil* y el *paramento* (Daganzo de Arriba, 1594) y, sobre todo, obras de pintura o dibujos: *pintura* (Arganda del Rey, 1691), *cuadro* y *cuadrito* (Hoyo de Manzanares, 1706) y *estampas* (Camarma de Esteruelas, 1754). El vestir las camas de la familia tenía gran importancia: *almohadas, colchón, jergón, sábanas, mantas.* Las sábanas se hacen de diversos tejidos, como *cáñamo*, y variados lienzos como el de Aroca (Arganda del Rey, 1688). Encontramos una manta de *sayal berbín* y una *colcha manchega* (Hoyo de Manzanares, 1706). También vemos alguna mención a la *frazada* 'manta peluda' (Arganda del Rey, 1691). Relacionado con los tejidos, pero para las labores domésticas, tenemos *lana* (Daganzo de Arriba, 1591), instrumentos para espadar el lino (*espadilla,* Hoyo de Manzanares, 1706) y para tejer (*urdidor* y *casillar,* Arganda del Rey, 1691).

En cuanto a las armas, estas aparecen con frecuencia en los inventarios de bienes y en las causas judiciales: *arcabuz* (Daganzo de Arriba, 1594; Arganda del Rey, 1690), *escopeta, espadín* (El Escorial, 1668); *pistolas de munizión* (Camarma

26 La forma *bacallao* es normal en la época (*Autoridades* 1726), y también se registra a menudo con la forma ultracorrecta con *–d–* (1822, *CORDE*).

de Esteruelas, 1765), *cachiporra* (El Escorial, 1768), *bayoneta* (Arganda del Rey, 1779).

Las partes de la vivienda son descritas a menudo: *pajar* y *corral(es)* (El Escorial, h. 1700; Camarma de Esteruelas, 1754, entre otros), *establo, barraca* (El Escorial, 1723), *dormitorio* (Camarma de Esteruelas, 1754). Como corral para los cerdos, también registramos *corte* (Camarma de Esteruelas, 1765). En Hoyo de Manzanares se citan, además, la *tinada* (1706). No se puede olvidar el caso de *cacera* 'cauce para agua', que en este ejemplo concreto se refiere a un desagüe hecho en el corral (El Escorial, h. 1700). En la vivienda se puede citar la presencia de *aldava* y *zerrojo* (El Escorial, 1708).

6.4 Familia, fiestas religiosas y creencias

Se registran en el grupo de la familia los elementos *madre, padre, hijos, hermanos, marido* en localidades como Camarma, El Escorial, Arganda del Rey, Hoyo y Daganzo de Arriba. Encontramos también referencias a las fiestas religiosas, en concreto, las *Nabidades* y las *Máscaras* 'Carnaval' (El Escorial, 1771), además de *Carnestolendas* (Arganda del Rey, 1798). No tan importante para la liturgia y la celebración, pero sí como fecha de referencia para las labores agrícolas y el inicio del otoño es el día *de San Miguel* (El Escorial, 1668). En las expresiones orales alusivas a la religión, tenemos *Por vida de Dios* (El Escorial, 1723).

6.5 Juegos y ocio

Las referencias a los juegos y al ocio son escasas, aunque algunos nombres sí aparecen en diversos documentos judiciales: *juego del villar,* llamado también *de trucos* (El Escorial, 1723).

6.6 Tiempo atmosférico

Hay pocos testimonios, pero sí se puede destacar *culebrina* 'relámpago' (San Lorenzo de El Escorial, 1732), *estar la noche clara* y *hacer luna* (Arganda del Rey, 1798).

6.7 Accidentes de terreno, huerta y árboles

En cuanto a los accidentes de terreno, se puede destacar la aparición de *cerro* (Arganda del Rey, 1775), *peña* (*id*., 1798), *cerviajo* y *cuesta* (Daganzo de Arriba, 1782). Para los caminos, *bereda* (Arganda del Rey, 1677). En la toponimia menor se refleja el uso de *cabeza* como 'monte pequeño' en el paraje *Cabeza Gorda* (Arganda del Rey, 1775). En ocasiones se mencionan las *huertas* o *huertos*

(Arganda del Rey, 1605; *id*. 1677). Entre los árboles, contamos *encina mayor, roble* y *alcornoque* (Hoyo de Manzanares, 1682), *chaparro* (*id*., 1706), *álamos* (Arganda del Rey, 1754), y entre las plantas, *mielgas* (*id*., 1775).

6.8 Labores del campo

Como es de esperar en este contexto rural, las labores del campo ocupan una buena parte del léxico. Tenemos utensilios tan empleados como el *arado* o el *yubo/yugo/ubio,* partes del arado como la *belorta, timón, telera, rexa, cama* así como *carros, carretas* (Camarma de Esteruelas, 1769) con sus *bujes* (Hoyo de Manzanares, 1706). Asimismo, encontramos menciones a instrumentos de trabajo agrícola: *azadón jarero* (Hoyo de Manzanares, 1706), *azadón retamero, arnerillo* (Camarma de Esteruelas, 1765), *trillo, bielo, palas, orcas, carreta* (*id*., 1769). Otros son la *piqueta, hoz de podar, rastrillos* (Arganda del Rey, 1691). En estas labores también se emplean cestos y otro tipo de recipientes: *espuerta, esportillo, canastillo, escusabarajica* (Daganzo de Arriba, 1594); *escriño, esportón* (Arganda del Rey, 1691). No sería tan común el *hierro para sacar criadillas de tierra* (Hoyo de Manzanares, 1706). Algunos de los frutos que se recolectan son *cañamón* (Arganda del Rey, 1688), *cebada* (*id*., 1691)*, guisantes, almortas y garbanzos* (Camarma de Esteruelas, 1769).

Para las tierras, hay *prado de pasto siego, zerca* 'prado cercado' (Hoyo de Manzanares, 1706), *tierra de labor, majuelo, suerte* (Arganda del Rey, 1677), *tierra cañamar* (*id*., 1688). Tiene importancia la viticultura en algunos municipios: *cepas, viña* y *viña aragonés* (Arganda del Rey, 1677); también se apunta la existencia de *olibo* (Daganzo de Arriba, 1594; Arganda del Rey, 1688). La tierra puede ser un *centeno* (Arganda del Rey, 1677; Camarma de Esteruelas, 1775), un *trigo* (El Escorial, 1768), un *trigal* (Camarma de Esteruelas, 1754), o llamarse *tierra ricial* (Camarma de Esteruelas, 1761). Otros campos se denominan *barbecho* (Daganzo de Arriba, 1591), *rastrojera* (Valdemoro, 1671), *herrén* (Hoyo de Manzanares, 1706), *espiga, agostadero* 'lugar donde agosta el ganado' (El Escorial, 1768)[27]. El montón de haces es la *hacina* (Camarma de Esteruelas, 1754). La unidad de medida más empleada para superfice es *fanega* (Daganzo de Arriba, 1594, entre otros), y para la capacidad de los granos, *celemines* y *arrobas* (Arganda del Rey, 1688).

27 «por cuya causa no se á abierto ni dado permiso para que entren en ella a pastar la espiga y agostadero ganados algunos». Es posible que *espiga*, más allá del significado habitual, signifique 'terreno con espigas donde pasta el ganado'.

6.9 El vino y el aceite

En algunos lugares, sobre todo los más meridionales, tiene un papel destacado la elaboración del vino, como es el caso de Arganda del Rey, aunque en muchas localidades se menciona esta bebida y elementos relacionados con ella debido a que su consumo estaba generalizado: *vino* (Daganzo de Arriba, 1588; Hoyo de Manzanares, 1706), y *vino moscatel* (El Escorial, 1771). Este se guarda en *tinajas* (Arganda del Rey, 1691; Daganzo de Arriba, 1591), en un *cuero* (Daganzo de Arriba, 1588), en un *jarro toledano* (Hoyo de Manzanares, 1706) y en un *pellejo* (El Escorial, 1771). Para el zumo de uva agraz se emplea la *agrazera* (Camarma de Esteruelas, 1765). También aparecen *mosto* (Daganzo de Arriba, 1591) y la *cueva*, que servía de bodega (Arganda del Rey, 1677). En cuanto al aceite, podemos encontrar la mención al término (Daganzo de Arriba, 1594; Arganda del Rey, 1691; Camarma de Esteruelas, 1754), y recipientes para este: *cántaro* (Daganzo de Arriba, 1594) y, más específica, la *azeitera* (Hoyo de Manzanares, 1706).

6.10 Animales domésticos

En el grupo de los animales domésticos, son numerosas las menciones a las caballerías, tanto caballos como mulas que servían para el transporte y las labores agrícolas. Para el híbrido entre asno y caballo, encontramos *mula* (Daganzo de Arriba, 1591 y 1782; Camarma de Esteruelas, 1754 y 1769, El Escorial, 1774) y *macho* (Daganzo de Arriba, 1591; Arganda del Rey, 1677; Colmenar de Oreja, 1712). De las formas de denominar al asno joven, hay que mencionar *pollina* (Arganda del Rey, 1677 y 1690; Camarma de Esteruelas, 1769; Corpa, 1765; Campo Real, 1798) y *pollino* (El Escorial, 1774; Daganzo de Arriba, 1782), *borrica* (Daganzo de Arriba, 1591; El Escorial, 1723), *borriquilla* (El Escorial, 1723)[28]. Como cría del asno se emplea también *zerril* (Corpa, 1765). Entre los caballos, *rocín* (El Escorial, 1570), *potra* (Arganda del Rey, 1792) y *yegua* (Arganda del Rey, 1688).

A las caballerías y asnos también se las describe, sobre todo cuando se hacen tasaciones o en caso de extravío: [una potra] *estrellada* (Arganda del Rey, 1792) y [una pollina] *patoja* (Corpa, 1765). En un documento de finales del siglo XVI que demanda los daños a un albéitar, se cita que el rocín está *manco* (El Escorial, 1585).

28 La *borriquilla* del documento de El Escorial, es joven, ya que dice que es «como de un año».

De especial valor consideramos una denuncia por pérdida de una pollina en Campo Real en 1798, ya que contamos varios adjetivos sobre el animal: *caveza acarnerada, vaja de hombros, anquirredonda, bragada, ojalada roja.* También se mencionan partes de su cuerpo, entre las que destacamos los *vasos* 'pezuñas' y los *corbejones*. Los aparejos de las caballerías también son citados: *albardones, estribos, fierro* (Arganda del Rey, 1688), *albarda* (Corpa, 1765). Hay términos para su pelaje, como una borrica *ruzia* y otra *parda* (Daganzo de Arriba, 1591), un caballo *castaño* (Colmenar de Oreja, 1712; Arganda del Rey, 1798) o una pollina *negra* (Arganda del Rey, 1677). En Camarma de Esteruelas (1769) tenemos una pollina *vociblanca.*

En zonas con tradición de ganado vacuno se encuentran animales de este grupo: *vaca, ternera* y *torete* (Camarma de Esteruelas, 1769), *buey* y *novillo* (Hoyo de Manzanares, 1706). En Daganzo de Arriba encontramos *becerro* (1591). La vaca puede ser *cerril* o *domada* (*id.*), y también *picarona* (El Escorial, 1769; v.i. 7.9). Para el pelaje, hay adjetivos para vacas y bueyes como *roxa* y *castaño,* usados como nombre propio pero alusivo a su color, mientras que también se registra un curioso *buey conejo* (Hoyo de Manzanares, 1706). Otro animal de gran importancia en el mundo rural madrileño ha sido el cerdo. Al adulto se lo puede denominar *puerco* (Daganzo de Arriba, 1591), pero esta voz solo aparece en el siglo XVI, para después documentar *macho de zerda simental* (Camarma de Esteruelas, 1765), los genéricos *ganado moreno* (El Escorial, 1723) y *ganado de cerda* (El Escorial, 1768), *cerdo*, *cerda* (Camarma de Esteruelas, 1769) y *gorrino* (El Escorial, 1771); a la cría, *cochinillo* (Camarma de Esteruelas, 1760) y *lechón* (Daganzo de Arriba, 1591 y 1594).

Dentro de los animales domésticos, del perro se puede especificar su clase: *alanos, dogos, lebreles, galgos, sabuesos* (El Escorial, 1697), *perro galgo, podenco, juzguillo,* o, sencillamente, con un diminutivo (*perrilla*, Arganda del Rey, 1775). Si es mestizo se llama *atravesado* o *mestizo* (El Escorial, 1697). Otros animales domésticos citados, aunque con menos frecuencia, son las aves: *gallinas, gallo, pollos, gansos* (Daganzo de Arriba, 1591).

Además de las referencias a los animales, en un documento relativo a la venta de los despojos de carnicería tenemos *cabezas, carrillada, chofes, asadura,* y los nombres genéricos *inmundicia, despoxos, reliquias* y *asquerosidades* (Arganda del Rey, 1784). En un proceso contra unos cazadores, hay referencias al *sebo* y las *costillas* de los ciervos (El Escorial, 1770). A los animales se los guarda en *establos* (v.i.) y necesitan *paja* (Camarma de Esteruelas), también se menciona el estiércol, llamado *freza* (Camarma de Esteruelas, 1754). Para la apicultura, conservamos el testimonio de los *basos de colmena* (Arganda del Rey, 1691).

6.11 Caza

En algunos lugares la caza, sobre todo la furtiva, tuvo una presencia continua en el desarrollo de los procesos penales. Esto se ve de una manera especial en El Escorial, una zona rica en bosque y caza limitada por ser posesión real. En sus documentos encontramos alusiones a la fauna cazada: *res gamuna* y *res zerbuna* (El Escorial, 1770), *chivos* (*id*., 1763), *perdiz, liebre* (Arganda del Rey, 1775). También vemos léxico relativo a los instrumentos para esta actividad: *escopeta, valas, pólvora, perdigones, martillo de escopeta, papel de pólvora, plomo* (El Escorial, 1770), *percha* (Arganda del Rey, 1775).

6.12 Oficios

Dentro de los oficios, hay menciones constantes a las diferentes profesiones de las personas que aparecen en la documentación. Los cargos municipales son *alcalde mayor, secretario* (El Escorial, 1771), *postillón* (Arganda del Rey, 1779), *guarda* (El Escorial, 1763). No faltan otras ocupaciones comunes en todos los pueblos de la época: *cura, panadero, tabernero* (El Escorial, 1771), *herrero, jornalero* (Camarma de Esteruelas, 1740), *pastor* (Campo Real, 1798 y Camarma de Esteruelas, 1754) y *carretillero* (Camarma de Esteruelas, 1754). En El Escorial se documenta *albéitar* (h. 1585), *pastelero* (1588), *ermitaño* (1616), *albañil* (1624), *mercader* (1723), *operario, fiel, pagador* (1763); *cuadrilleros* (1768), *criado, veedor* y *conserje* (1771), *platero* (1778); para la carnicería, *tablajero* (Arganda del Rey, 1784). Encontramos algún oficio femenino, como *labandera* (Arganda del Rey, 1791); de una mujer se dice que tiene *como oficio remendar* (El Escorial, 1723). No faltan los grados y empleos militares: *brigadier, granaderos, sarjento* (Colmenar de Oreja, 1712). Más infrecuente es el caso de *fabriqueros* (Hoyo de Manzanares, 1682), *padre obrero* (El Escorial, 1771) y *mozo de espuelas* (*id*., 1778).

7 Origen y documentación histórica de algunas voces

Las voces seleccionadas para su estudio se adscriben en su mayor parte al léxico material de la Comunidad de Madrid, mientras que otras corresponden a diversas parcelas referenciales, muy dispares entre sí (términos para las diferentes edades de la persona, insultos, etc.). Especialmente difíciles de clasificar son los verbos, pues los que hemos destacado no pertenecen al léxico material (*llevar* por 'cobrar', El Escorial, 1624, y *visitar el bolsillo*, el Escorial, 1771, por 'gastar'). Los sustantivos examinados parecen, en principio, más reveladores de la distribución geográfica del léxico. Por otra parte, es posible conjeturar la continuidad

histórica de este vocabulario[29]. Podemos, pues, sospechar que ciertas palabras documentadas en nuestro corpus madrileño pudieron tener gran antigüedad en los mismos enclaves geográficos, de acuerdo con el principio señalado de la escasa movilidad a lo largo del tiempo del léxico material para técnicas de cultivo, herramientas y aperos agrícolas[30], mientras que la terminología para prendas de vestir o tejidos conoció más mudanzas, por ser más permeable al neologismo, sobre todo por préstamo (Vázquez Balonga 2015: 134; Varela Merino 2009: 105). Este léxico material puede revelar fronteras lingüísticas muy antiguas (v.i. *ubio* y *yugo*). El problema metodológico que se plantea en un espacio como este es la falta de documentación temprana, por lo que la etapa medieval queda completamente oscura, y casi solo los fueros de Madrid (ca. 1200) y de Alcalá (ante 1247) son muestra de este vocabulario medieval. Examinamos aquí la historia de algunas voces que pueden esclarecer la historia y la geografía interna del léxico de la Comunidad de Madrid y que, en algunos casos, tal vez quepa catalogarlas de madrileñismos.

7.1 El cuerpo humano y la persona

Las edades del hombre son objeto de diversas denominaciones. Nos fijaremos en *chicuelo*, que interpretamos como 'adolescente', pues, en efecto, el documento señala que sería de doce años «poco más o menos»: «acobardado el *chicuelo* de tan impensado aprieto y peligro» (Arganda del Rey, 1784). El sufijo *-uelo* es inusual en esta zona, por lo que atribuimos expresividad al derivado (Náñez 1973). El agresor de este muchacho, de mayor edad, es llamado *mozo*: «contra el relacionado Juan Antonio Aguilar, *mozo* soltero de este expresado vecindario». Este puede desempeñar diversos oficios, como «mozo de espuelas» (San Lorenzo de El Escorial, 1777–78) (v.i. 7.2).

No se distinguirían, en cambio, por su trabajo, los llamados *majos*: «que era una escandalosa que los *majos* la davan pañuelos de seda» (Arganda del Rey, 1796). Con el sentido moderno, el sust. *majo*, según el *DECH* procedente de *majo* 'mazo', debió circular pasada la mitad del siglo XVIII en la ciudad de Madrid. En

29 Ejemplo extremo es la voz *bacelar* 'viña joven', documentada por primera vez en la *Nodicia de quesos* de 974 y recogida en el sur de León, Zamora y Salamanca por la dialectología moderna con el mismo sentido (Álvarez Tejedor 2007: 205).

30 Por ejemplo, *belorta*, *telera* se remontan a la Edad Media, pero tenemos testimonios en el siglo XVI, como el derivado *belortón* en Getafe en 1579 (Vázquez Balonga 2015: 260).

Autoridades (1734) ya se recoge de esta manera: «El hombre que afecta valentía y guapeza en sus acciones. Comúnmente se llama assí a los que viven en los arrabales d'esta corte». Entre las primeras documentaciones literarias destaca José Cadalso: «un *majo* del Barquillo no hablaría con más bajo estilo» (1768), y Nicolás Fernández de Moratín en *El arte de putear* (ca. 1771). Desde la ciudad debió de extenderse a los pueblos colindantes ya a finales del siglo XVIII. Posiblemente, la idea de *majo* como 'habitante de los arrabales' llevó a un sentido de 'hombre de baja calaña, pendenciero', lo que podría adecuarse a la situación descrita en Arganda del Rey.

Los majos serían, al menos algunos de ellos, *pícaros*. Sin embargo esta voz no parece usarse en nuestro corpus con el sentido habitual de «personaje de baja condición, astuto, ingenioso y de mal vivir, protagonista de un género literario surgido en España en el siglo XVI» (*DLE*, s. v.), sino más bien en el de 'inmoral', 'sinvergüenza' e incluso 'corrupto', como se aprecia en Galdós[31]. Así, en una denuncia por intento de violación se narra lo siguiente: «a que volvió a dezir la testigo: *Pícaro*, ¿a qué vienes aquí?» (El Escorial, 1708); y en una causa por injurias, «la insultó con las expresiones de que era una *pícara* advenediza forastera» (Arganda del Rey, 1796). El mismo sentido de 'corruptela' parece tener *picardía*: «El señor alcalde mayor y el padre obrero y el señor secretario, ustedes aquí no son llamados sino para ser alcagüetes de las *pircardías* de el señor don Ánjel» (El Escorial, 1771). Observamos, pues, el proceso de peyorización de *pícaro* y *picardía*, de manera que esta es definida en *Autoridades* como «acción baxa, vileza, ruindad, engaño y maldad», donde se señala la equivalencia con lat. NEQUITIA.

Las relaciones personales no siempre transcurrían por vías pacíficas. Antes al contrario, las peleas y disputas eran frecuentes. Tanto para riña verbal como física se empleó la palabra *quimera*, recogida con este valor en el *DLE*. En nuestro corpus aparece en Camarma de Esteruelas (1760): «el dicho Diego Martínez, dijo que habiendo armado la *quimera* en la cassa de el confesante, le dijo su muger que tales *quimeras* las tubiesen en la calle». La voz tuvo en el castellano de los siglos de Oro el sentido, más proximo al etimológico, de «aquello que se propone a la imaginación como posible o verdadero, no siéndolo» (*DLE*, s. v.), mientras que el valor que tiene en nuestro texto solo lo encontramos en el siglo XVIII: «y la España toda está tan quieta y dócil, que ha años que no se oye ni una *quimera* de garrotazos» (Diego de Torres Villarroel, *Visitas y visiones con*

31 «[…] el príncipe de la paz, ese monstruo […] para que la nación se vea libre de pícaros», (*Episodios Nacionales*, I, 508).

don Francisco de Quevedo por a corte, 1727–28, *apud CORDE*). Así, en Oudin (1607) es 'reverie', mientras que en Terreros (1788) vale tanto 'riña' como 'fantasía'.

En las disputas había golpes, como *cachete*: «y la havía dado muchos *cachetes*» (El Escorial, 1708). En *CORDE* solo lo documentamos desde 1758 (Padre Isla, *Historia del famoso predicador Fray Gerundio de Campazas alias Zotes*); sin embargo, parece leerse «a cachetes» en un documento emitido en Daganzo de Arriba en 1666: «Porque me tiene enfadado. Y diziendo y aciendo <…> a *cachetes* contra el dicho Domingo Martínez» (*CODEA* 2384), empleo que concuerda con Covarrubias (1611): «el golpe que se da con el puño cerrado»[32]. Esta acepción es más acorde con el contexto que la moderna de «golpe ligero que se da en la cara u otra parte del cuerpo con la palma de la mano», que recoge el *DLE*.

Empellón era poco frecuente en el siglo XVIII, y solo parece sobrevivir en algunos autores como el padre Isla y Meléndez Valdés: «y los echaba y echó a *empellones* de su casa» (Camarma de Esteruelas, 1760). La forma más habitual, *empujón*, conoció una variante popular, *rempujón*, documentada en Arganda del Rey: «el de menor estatura le dio un *rempujón* en el pecho» (Camarma de Esteruelas, 1754). La voz aparece en Juan de Timoneda (1666), así como en Lope de Vega, pero en el siglo XVIII parece confinada al habla popular (en *Autoridades*, *empujón* se define como «el golpe o *empellón* que se da a otro»).

Los daños no eran recibidos solo por las personas, sino también por las cosas: «que la havía de dar azotes dando *porrazos* y pedradas a la ventana de la casa» (Camarma de Esteruelas, 1760). Con la acepción de dar un *porrazo* a una persona, lo encontramos desde 1607 (San Juan Bautista de la Concepción, *Memoria de los orígenes de la descalcez trinitaria*, *apud CORDE*). En Franciosini (1620) es «mazzata, bastonata». En el contexto todavía es «Golpe que se da con la porra o con otro instrumento».

Portazo no parece que fuera una palabra muy difundida, al menos durante el siglo XVIII: «y dando un grande *portazo* con este desaire hubo de retirarse» (Arganda del Rey, 1796). Solo la documentamos desde el siglo XVIII (Padre Isla, 1758, *apud CORDE*) y, entre los diccionarios, la recoge por primera vez *Autoridades* (1737).

7.2 Oficios

En nuestro corpus aparece *mozo de espuelas* (San Lorenzo de El Escorial, 1777–1778), voz frecuente en los siglos XVI–XVII, pero presente solo un par de veces

32 La misma explicación se lee en *Autoridades* (1729) para *cachete*.

en el XVIII y XIX en *CORDE*, por lo que ya en la época de nuestro documento debía de ser sintagma poco habitual. Siguiendo con oficios y ocupaciones, señalamos «padre obrero» (El Escorial, 1771), que era la persona que debía de estar al cargo de los trabajadores del Monasterio de San Lorenzo[33]. Encontramos también *destaxero* en esta misma localidad en 1587 (no aparece en *CORDE*, pero sí en las fuentes lexicográficas desde Franciosini, 1620)[34]. *Jornalero* se encuentra en Camarma de Esteruelas en 1760, pero la voz estaba ya documentada al menos desde 1529 (en *ADiM* es la respuesta a «obrero eventual» en Cubas de la Sagra, al sur de la Comunidad). *Operario* está documentado en El Escorial en 1763: «Alfonso de Llamas, residente en dicho Real Sitio, pagador de los *operarios* en el camino», aquí ya en el sentido moderno, que parece neologismo del XVIII, mientras que los empleos encontrados desde principios del siglo XVII parecen referirse a la labor evangélica de los sacerdotes[35]. En el *ADiM* vemos *obrero*, *peón* y *trabajador*, pero no *operario*. De estas, *peón* está en cartas del Monasterio de San Lorenzo: «son menester de ofiçiales y *peones*» (1563), «haviéndose descolgado un lienço de cornisa de cantería, que cae a la parte del seminario con grave daño de dos *peones*» (1673).

Mayor interés tiene para nuestro propósito la voz *fabriquero*, que localizamos solo en la documentación de Hoyo de Manzanares (1682); se aplica a los que se dedicaban a elaborar carbón y a cortar corteza. Encontramos la voz en *Documentos sobre música de la catedral de Sigüenza* (1714–1750, en *CORDE*). En *Autoridades* (1732) se recoge la acepción de encargado de la fábrica de una catedral, además de la más general de 'fabricante'. En la edición del *Diccionario* académico de 1914 es «operario que en los montes trabaja en el carboneo». Parece, pues, que este significado es una innovación tardía en la lexicografía, pero ya atestiguada a finales del siglo XVII en la localidad de Hoyo de Manzanares.

7.3 Indumentaria

Destacamos una voz que no encontramos en la mayoría de las fuentes lexicográficas: *paletina* («Item, una *paletina* de gasa con motas encarnadas», Camarma

33 En otra documentación aparece «padre de güérfanos» (Madrid, 1600), referido seguramente al responsable de la Inclusa de Madrid, y también tenemos el testimonio de «padre de los locos», para el encargado de cuidar a los enfermos mentales (h. 1793) (Sánchez-Prieto Borja y Vázquez Balonga 2017).

34 En *Autoridades* es «el que hace la obra a destajo o por un tanto».

35 «[…] el padre Lorenço Luzero, el más antiguo *operario* de los que tiene oy aquella labor» (Manuel Rodríguez, *El Marañón y Amazonas. Historia de los descubrimientos*, 1684, en *CORDE*).

de Esteruelas, 1765)[36]. En Terreros (1788) es «adorno de la garganta que baja al pecho» y proviene del francés *palatine* y este del italiano *pallatina.*

Entre los complementos de la indumentaria está *faldriquera*, muy frecuente en el siglo XVII, pero que declina en el XVIII y solo en autores tradicionalistas como Pereda se encuentra en el XIX. En nuestro corpus aparece en Arganda del Rey (1798): «abrí la puerta con la llabe que tenía en la *faldriquera*». En La Aceveda el *Diccionario de madrileñismos* da *faldiquera*, seguramente por cruce con *falda*. Que el término no debía ser muy corriente, al menos en la lengua escrita, parece corroborarlo el que en este mismo documento se emplee *bolsillo* para el mismo concepto, el de bolsa de tela interior, no cosida a la ropa: «y la llabe tenía en mi *bolsillo*». Con este mismo sentido parece usarse en El Escorial (1771): «¿Qué diremos de el cura, que sale de la tarberna de bessitar el *bolsillo* a las onze de la noche?».

7.4 Vida doméstica

Entre las voces sobre la construcción y equipamiento de la casa, incluimos aquí *cacera*, que el *DLE* define como «Zanja o canal por donde se conduce el agua para regar», pero que puede ser también 'desagüe', como se ve en «y contra derecho a guiar las aguas que de las llubias caen de su tejado por una *cacera* que ha hecho nueba al corral de su cassa» (El Escorial, h. 1700); aparece también en Valdemoro en 1889: «y havido un Puentecillo para pasar la *cacera*»[37]. En el *ADiM* (García Mouton y Molina 2015), solo aparece en Santa María de la Alameda y Cubas de la Sagra (en Suroeste). El *Diccionario de madrileñismos* la muestra en Bustarviejo, Chinchón, Colmenar Viejo y Titulcia. Obviamente, es un derivado de *caz* 'cauce', lo que apunta a que esta voz debió de estar difundida en Madrid.

La palabra *corte* para 'corral de cerdos, pocilga' tiene tradición medieval (ambas son equivalentes, y así el lugar donde se celebran los juicios puede ser el *corral*, p. ej., en el *Fuero de Alcalá*). En nuestro corpus la encontramos en Camarma de Esteruelas en 1765, en el sentido de divisiones para albergar a los cerdos: «un pajar algo undido con sus *cortes* para cerdos y un huerto». Aunque es un término de largo recorrido, hoy en día aparece documentado en el *ADiM*, precisamente en la zona este de Madrid, donde se ubica Camarma de Esteruelas: Meco y Carabaña, muy cercanos a esta localidad, y Alalpardo y Colmenar de

36 Según el *NTLLE*, la última vez que se recogió *paletina* en el diccionario académico fue en la edición de 1989.

37 Aparece también en la obra *Colectivismo agrario* de Joaquín Costa, que vivió en Graus, provincia de Huesca (*CORDE*, 1893 y 1898).

Oreja, más alejadas, pero igualmente en la franja oriental[38]. En el *ALECMAN*, resulta llamativo que *corte* sea el término elegido por la práctica totalidad de los encuestados de la provincia de Guadalajara para este concepto, frente a la preferencia en Toledo por *pocilga*. Por lo tanto, estamos ante un regionalismo documentado en el siglo XVIII y todavía vivo en Madrid, sobre todo al este.

Es interesante el término *casillar*; en un inventario de bienes de Arganda del Rey en 1691 (*CODEA* 1842), «un urdidor y un *casillar*». Aunque no aparece en el diccionario académico ninguna acepción que sea satisfactoria para el contexto, ha sido señalado en el *Tesoro Léxico de las Hablas Andaluzas* (*TLHA*) de Alvar Ezquerra como un instrumento para el tejido, en concreto, 'caja de madera utilizada para guardar los accesorios de la urdidera', pero la aparición en el *CorLexIn* en la provincia de Guadalajara en 1640 desmiente que sea un andalucismo (Torres Martínez 2014: 155), pues su presencia en un documento de la Comunidad de Madrid confirma la idea de que la voz, al menos en esa época, estuvo difundida en el centro peninsular.

Añadimos aquí, en el grupo de armas, *cachiporra*, que aparece en un registro de El Escorial (1768). Es una voz antigua, datada según el *DECH*, en su variante *cachiporro* y *cachiporro*, en la primera mitad del siglo XVI, originada a partir de una partícula *cach-/caz-* y el sustantivo *porra*. Lo cierto es que se encuentra en el siglo XVI en *CORDE*, pero parece tratarse de un insulto. Con el sentido de 'palo que termina en una bola' del *DLE* no se ve en *CORDE* hasta Don Ramón de la Cruz en 1766 («de payo, con una *cachiporra* y un pañuelo atado»).

7.5 Familia y fiestas

En otros ámbitos referenciales, como el de las fiestas, incluimos *máscaras*, que en el contexto interpretamos referida al Carnaval («en regalando un dobloncico de a ocho para las máscaras y un pellejo de bino moscatel», El Escorial, 1771). Para el mismo concepto, resulta curiosa la pervivencia del término *Carnestolendas* («y digo que haviéndome combidado a cenar en la noche de martes de *Carnestolendas*», Arganda del Rey, 1798), raro en el siglo XVIII. En el *Diccionario de madrileñismos* se documenta *carrastroliendas* en Alameda, con el sentido de «broma carnavalesca». En el *ADiM* (2015), en cambio, no se ha registrado el uso del término *Máscaras* ni *Carnestolendas* para denominar a la fiesta de Carnaval, que ha quedado simplemente con esta denominación en singular o en plural en todo el territorio de la Comunidad de Madrid.

38 En otros municipios de la zona oeste, como Valdemorillo y Cenicientos, se emplea la voz *pocilga* (*ADiM* 2015).

7.6 Juegos y ocio

El juego de billar era ya conocido en el siglo XVIII, pero en El Escorial (1723) se emplea un sinónimo, *trucos*: «su tienda, que está zerca de la de Monseur Dupón, en que este tiene juego de villar o *trucos*». Debía de ser similar al que describe el *DLE*: «m. pl. Juego de destreza y habilidad que se ejecuta en una mesa dispuesta a este fin con tablillas, troneras, barras y bolillo». En el *Tesoro de las tres lenguas* de Vittorio Girolamo Vittori, de 1609, se lee «Trucos, el juego de los trucos o truecos, *une sorte de ieu de billard*, *una sorte di gioco di pirla*».

7.7 El tiempo atmosférico

Encontramos *culebrinas* en San Lorenzo de El Escorial (1732), que según el *DLE* es «meteoro eléctrico y luminoso con apariencia de línea ondulada», pero falta en *Autoridades*. Sorprendentemente, en el *ADiM* la única respuesta en todas las localidades encuestadas es *relámpago*. En el *ALECMAN* es general *culebrina* en Guadalajara, Cuenca y Albacete, mientras que en Ciudad Real la forma más frecuente es *culebrilla*.

7.8 Labores del campo

El término general es *campo*, pero cuando este se encuentra sembrado puede nombrarse por el cereal que en él se cultiva, de manera que *trigo* vale 'tierra sembrada de trigo': «habiendo salido el declarante a ver *un trigo* que tiene en el término de ella» (Camarma, 1775). Con el mismo sentido se emplea *un centeno*: «Se han comido un *centeno* que tenía mío propio» (Camarma de Esteruelas, 1761).

Un término poco documentado es *ricial*: «hasta llegar a una tierra *ricial* de zevada» (Camarma de Esteruelas 1761), «y mediodía un *rizial* y erial» (Camarma de Esteruelas, 1776). En el *DLE*, *ricia* es «Campo que se siembra aprovechando las espigas que quedan en él», aunque interpretamos que significa lo que nace del grano que quedó en el terreno tras la siega, o bien el prado que retoña tras cortarlo, aunque claramente en el contexto tiene el primer significado. En el *CORDE* solo aparece en 1944 en un tratado de ganadería[39]. En Alvar Ezquerra (2011a) se recoge *ricia* en Torres de la Alameda.

Un sintagma que solo hemos documentado en Hoyo de Manzanares es «pasto (y) siego»: «Más un pradito de *pasto siego* y monte de chaparro que está

39 «[…] para pasto principalmente después de un corte, el ricial o retoño, y los pastan toda clase de ganado» (E. González Vázquez, *Alimentación de la ganadería y los pastizales españoles*, Madrid-Barcelona 1944, *CORDE*).

en término d'esta villa y llaman el pradito de los Arroyuelos» (1704). Suponemos que la forma masculina en lugar del postverbal *siega* se debe a la atracción de *pasto*.

Tampoco hemos encontrado fuera de la provincia de Madrid la voz *cerviajo*: «se arrojó violentamente del pollino dicha mi muger, sin reparar el peligro a que se expuso por defender su honor, pues cayó de un *cerviajo* y se halló sin lesión» (Daganzo de Arriba, 1782, *CODEA* 1848). La palabra aparece un par de veces más en la provincia de Madrid: «Otra [tierra] al de Valdezarza con su *cerviajo* de álamos negros» (*Diario de Madrid*, 3 de abril de 1803), «linda: Norte, Pascual de la Peña; Este, *cerviajo* por medio, Valentín Martínez y Paulina Sánchez» (Boletín Oficial de la Provincia de Madrid, 8 mayo 1941). Debió de tener el sentido de 'borde alto que delimita una finca o separa esta del camino'.

Cabe destacar también *pobo* 'álamo blanco, chopo', que se encuentra en Daganzo de Arriba (1591). La primera documentación en *CORDE* es de 1582 (Luis Gálvez de Montalvo, *El pastor de Fílida*)[40]. En el *DLE*, y en toda la tradición académica desde sus inicios, desde *pobo* se remite a *álamo blanco* y la base etimológica es la misma que *chopo*, el latín POPUS. Según indica el *DECH*, está extendido «en la provincia de Madrid y en otras partes». Su presencia en la toponimia (*Povedilla*, Albacete, *Poveda de la Sierra*, Guadalajara, *Poveda,* Ávila, entre otros de zona central) sugiere, según los autores del *DECH*, la existencia de una forma arcaica *POPUS, que dio lugar a *pobo,* mientras que *chopo* vendría de la forma diminutiva POPULUS[41]. Como en otros casos, aquí encontramos una preferencia en la actualidad por las formas estandarizadas, ya que en el *ADiM* (2015) se registran únicamente para 'álamo' las voces *álamo* (*blanco/negro*), incluso con la alternancia de *chopo,* y no se encuentra ya *pobo*[42].

Entre los cultivos destacamos *almortas*: «Más cuatro fanegas de almortas» (Camarma de Esteruelas, 1769). La coherencia histórica de la presencia de esta variante en el este de la Comunidad de Madrid queda confirmada por la coincidencia con toda la provincia de Guadalajara, como se observa en el *ALECMAN*, mientras que esta misma fuente recoge *prínsoles* en Toledo, *pitos* en Ciudad Real y *guijas* en Albacete.

---

40 Hay que notar que la edición es de Madrid y que el autor nació probablemente en Guadalajara y estudió en Alcalá de Henares (Martínez San Juan 1999: 17).

41 En la comarca de Alcalá, que incluye Daganzo de Arriba, se encuentra el topónimo menor *La Poveda* de Arganda del Rey.

42 Lo mismo sucede con el *ALECMAN*, donde tanto el álamo blanco como el álamo negro son llamados de este modo, y en ocasiones se prefiere *chopo.*

La palabra procedente de IUGUM conoció diversas variantes, de entre las que documentamos en nuestro corpus *yugo* («Item, un *yugo* para carro, echo pedazos el ombligo, con sus costillas», Camarma de Esteruelas, 1765), *yubo* («Una carreta para bueyes también bastantemente maltratada con un *yubo* para mulas», *id*., 1769) y *ubio* («y con su *ubio* viejo», «un *ubio* carretero de bueyes biejo», «más un *ubio* de arar bueyes bueno», Hoyo de Manzanares, 1704). La variante *yubo* parece dominar en Madrid en los siglos XVI y XVII, y así Vázquez Balonga (2015) la documenta en Alcorcón, Getafe, Arganda y Carabaña, mientras que *yugo* la encuentra en Olías del Rey, Mocejón, Ajofrín y Orgaz, en la provincia de Toledo. La forma *yubo* ha perdido extensión en la zona central, ya que en el *ALECMAN yugo* es general en Guadalajara, Cuenca y norte y centro de Toledo, mientras que en Ciudad Real, donde se añade *toza*, y Albacete, domina *ubio*. Asimismo, en el *ADiM* se documenta solamente la forma estandarizada *yugo.*

El *bieldo* es utilizado para aventar la paja y separarla del grano. Encontramos las variantes (*b*)*ielo* y *viela*: «tres palas, un *bielo* grande, cuatro o cinco orcas»; «Item tres *vielas*, y unas horcas viejas, tres rodillos y dos *vielos* pequeños» (Camarma de Esteruelas, 1769). El *DLE* solo recoge *bieldo* y *bielda*. En Vázquez Balonga (2015), en Alcorcón y Arganda del Rey en los siglos XVI y XVII hay *bielgo*, además de *bielga* en la primera localidad, y en Carabaña, *bielo*; esta última forma se documenta también en Mocejón (Toledo). Todo indica que la variante *bielo* estuvo arraigada en la Comunidad de Madrid, y aún en la actualidad es documentada por Alvar Ezquerra (2011a) en Paracuellos del Jarama[43]. Por otra parte, el *ALECMAN* muestra *bielo* en el NE de Toledo.

7.9 Animales domésticos

Numerosas variantes presentan los nombres del cerdo. En Camarma de Esteruelas se emplea *cerdo* («un cerdo de un año», «una cerda», 1769), mientras que en El Escorial aparece *gorrino* (1771). Hoy, en las localidades cercanas a Camarma encuestadas en el *ADiM* (Alalpardo y Meco), se atestigua *guarro*, *cochino*, *cerdo* en la primera y *cochino*, *gorrino* en la segunda, y en Santa María de la Alameda y Robledo de Chavela se emplean tanto *gorrino* como *cerdo*, según este mismo atlas dialectal. El verraco es «cerdo semental» y «macho de zerda simental» (Camarma de Esteruelas, ca. 1765). Para el lechón se emplea *cochinillo*: «por lo cual les dio un *cochinillo* para la feria» (Camarma de Esteruelas, 1760). Esta forma está extendida en la actualidad en Guadalajara y algunos puntos minoritarios del sur de Castilla-La Mancha (*ALECMAN*).

43 Sin embargo, en el *ADiM* se prefiere *bieldo, bielda* para el concepto 'instrumentos para aventar' y no queda rastro de las otras variantes.

El ganado vacuno es presentado de manera genérica como *res vacuna*: «halló las pisadas y freza de dichas *reses vacunas*» (Camarma de Esteruelas, 1754), pero el animal hembra es denominado *vaca* («Una *baca* llamada la Hermosa»), mientras que su cría es la *ternera* («Una baca llamada la Pajarita con una *ternera* de siete meses»); en el *ADiM, ternero* se documenta para el animal recién nacido en el norte y este de la Comunidad, y es raro cuando tiene entre dos y seis meses, pues pasa a denominarse *becerro*. El macho de cierta edad es *torete* («un *torete* de cuatro años», Camarma de Esteruelas, 1769). La res adulta para el trabajo se denomina *buey* («un par de *bueyes* de lavor, uno de ocho años y otro de nuebe»). Este recibe diversas denominaciones por su pelaje o capa. Entre estas llama la atención *conejo*, seguramente por sus tonos pardos como los del conejo de monte: «eszepto que el *buey conejo* á fenezido y en lugar de aquel la mando uno que al presente tengo que se llama Naranjo» (Hoyo de Manzanares, 1702). Otras documentaciones de este término se dan en Ávila («un *buey conejo*», *CODEA* 2389, Palacios Rubios, 1749), Segovia («un novillo conejo», *CorLexIn*, Revenga, 1655) y Zamora («dos bueyes mansos, uno negro y otro conejo», *CorLexIn*, Mahíde, 1664). Puede suponerse, pues, que es una voz del castellano norteño y central. No hemos encontrado más documentaciones de «vaca *picarona*» que la de El Escorial de 1769; no sabemos si guardaba alguna relación con el adj. *pícaro* (v.i.) o tal vez con *picón*[44], aunque lo más verosímil tal vez sea lo primero, por tratarse de una vaca que tiene tendencia a cornear, lo que podría cuadrar con el contexto («una vaca picarona[45] […] la pegué un rejonazo bien puesto»).

Notable variedad alcanzan los nombres del asno o burro. A pesar de ser estos los términos generales en español, en nuestro corpus el más frecuente es *pollino*, que, sorprendentemente falta en el *ADiM*. Según el *DLE pollino* vale tanto 'asno' como 'asno joven'. En nuestros testimonios, puede referirse al animal adulto: «Una *pollina* pelo negro voziblanca de edad de treze años tasada por dicho maestro en doszientos reales», «otra *pollina* de un año del mismo pelo voziblanca» (Camarma de Esteruelas, 1776), «una *pollina* que tenía mía propia, pelo castaño oscuro, de edad de siete años» (544, Corpa, 1798). El término se encuentra también en Alcalá de Henares (*CODEA* 255, 1601): «y le traían de un lugar a otro en un *pollino* por no poder andar»; en Daganzo de Arriba, al este de la provincia: «y se dirigió sola al labadero de Torote, conduciéndose con las ropas que había de labar en un *pollino* nuebo que tengo» (1782). En El Escorial,

44 «Dicho de una caballería: de dientes incisivos superiores que sobresalen de los inferiores, por lo cual no pueden cortar bien la hierba» (*DLE*, s. v.).

45 En Terreros y Pando (1788), *picarón* vale 'tramposo'.

en cambio, aparece *borriquilla* (1723), aunque también *pollino* (1774). Parece, pues, claro, que esta voz se concentra en el este de la Comunidad de Madrid, aunque su uso se ha visto reducido en favor de las palabras generales del español aprendidas en la escuela, por lo que falta en *ADiM*. Tampoco el *ALECMAN* da *pollino* para 'asno joven' más que en unos pocos puntos del centro de Guadalajara y sur de Albacete. En el este de Guadalajara y en Toledo y Ciudad Real es general *buche*. Sin embargo, el *ADiM* solo trae *buche* para 'pollino' en Lozoya, quizá un resto de una forma más extendida. Esta voz es de origen desconocido, tal vez de procedencia vasca (*DECH*). No se encuentra en el *COSER*.

La cría del asno es llamada *cerril*: «y la pollina que se ha llevado es de bastante altura, pelo negro, algo patoja, y al presente está criando un *zerril*» (Corpa, 1765). El término se aplica como adjetivo a otros animales («Más una baca *zerril* que se llama Gamita», Hoyo de Manzanares, 1706), como se recoge en el *DLE*, pero la peculiaridad del empleo en la localidad de Corpa consiste en la sustantivación tal vez especializada para la cría del asno. En la descripción del burro se utilizan, entre otros adjetivos, *patojo* («y la pollina que se ha llevado es de bastante altura, pelo negro, algo *patoja*», Corpa, 1798), término que documentamos en el siglo XX en *CORDE* (8 casos), pero referido a persona. En el mismo documento se emplean los términos *acarnerada* («la cabeza *acarnerada*»), *anquirredonda*[46], *bragada*, *ojalada*[47]. No hemos encontrado *vociblanco* («Una pollina pelo negro voziblanca», Camarma de Esteruelas, 1769), aunque sí *bocinegro* («un perro llamado Bezerrillo, bermejo, *bocinegro* y mediano», *CORDE*, López de Gómara, 1554), que era el que tenía cerco de otro color alrededor del hocico.

En cuanto al nombre para el mulo, destacamos que se prefiere *macho* en documentos de la zona este de la Comunidad de Madrid (Arganda del Rey y Colmenar de Oreja) en los siglos XVIII y XVIII, y esa voz coincide con los lugares donde se registra en *ADiM* (2015) que, salvo unos casos en la zona oeste (Navacerrada y Cenicientos), se ubica en el área este, incluso en la misma localidad (Carabaña y Colmenar de Oreja). En el *ALECMAN*, *macho* aparece con frecuencia en la provincia de Guadalajara, limítrofe con la zona este de Madrid.

Encontramos *freza* 'estiércol' en Camarma de Esteruelas (1754): «halló las pisadas y *freza* de dichas reses vacunas». En el *CORDE* aparece con este sentido en Joaquín Costa, nacido en Monzón (Huesca). En las encuestas del *ADiM* no se registra esta voz para 'estiércol', por lo que debe ser un término en claro desuso.

46 «Dicho de una caballería: Que tiene las ancas muy carnosas y convexas» (*DLE*, s. v.).

47 «Dicho de una res vacuna: Que tiene alrededor de los ojos, formando líneas circulares, el pelo más oscuro que el resto de la cabeza».

Para los canes, el término general es *perro*; el actual *galgo* aparece como sustantivo («cazando con dos *galgos*», Arganda del Rey, 1775), pero también en aposición («unos *perros galgos*»). El diminutivo parece indicar no que sea cachorro (*cf. cochinillo*) sino el tamaño pequeño o mediano: «llebaban una *perrilla*». Se habla en el mismo documento de «un podenco o *juzguillo*»; esta voz, que no hemos documentado en otros lugares, podría emparentarse con *gozque*, a través de un diminutivo *guzguillo*, cambiado tal vez en *juzguillo*. El perro mestizo o cruzado es *atravesado* («unos perros galgos y otro *atrabesado*», *id.*); según el *DLE* se aplica a cualquier especie animal.

8 Posibles regionalismos en los documentos de la CM

Llegados a este punto, y más en la perspectiva de la continuidad de este trabajo que en la de alcanzar conclusiones definitivas, pueden apuntarse algunas reflexiones que afectan, sobre todo, a la metodología de la investigación. El estudio del léxico de la documentación municipal examinada para el proyecto *ALDICAM-CM* revela la existencia sólida de una economía y de una cultura genuinamente rural (agrícola y ganadera) en la Comunidad de Madrid. A dicha cultura se ha ido sobreponiendo la influencia urbana, gracias al contacto incesante con la ciudad de Madrid, que es progresivo desde el siglo XVII, como acredita la documentación.

El léxico rural no se manifiesta tal vez en toda su fuerza, debido a la mediación de los escribanos que elaboran los documentos a partir de los usos aportados de las declaraciones de los vecinos de los pueblos en los que ejercen su labor, pero tenemos numerosos indicios de la autenticidad registral de esos profesionales de la escritura. Las voces rurales serían recogidas de manera fiel, generalmente, a partir del testimonio de denunciantes y testigos.

Esto favorece la profundidad histórica de este vocabulario, al menos todavía para el recogido en el siglo XVIII (*corte* como 'pocilga', *yugo, belorta*). De las voces tradicionales a veces quedan solo reminiscencias parciales, pues las formas de mayor rusticidad pudieron ir renovándose en los siglos XVIII y XIX en favor de las estandarizadas. Por último, hay que señalar que el léxico documentado tiene continuidad a día de hoy, como el caso de *cacera,* aún presente en el *ADiM* (2015). Otras veces, observamos más bien la sustitución léxica de las formas tradicionales por las nuevas (*relámpago* por *culebrina*), pero este fenómeno puede ser muy reciente[48]. La comparación entre Madrid y Castilla-La Mancha gracias a

48 Es lo que parece apuntarse por las diferencias entre *ALECMAN* y *ADiM*. La realización de las encuestas del *ALECMAN* en los años 80 y 90 del siglo pasado favorecería la pervivencia de léxico más diferencial, dialectal y rural en las provincias colindantes de Toledo y Guadalajara que en Madrid.

los atlas dialectales arroja formas de más antigüedad en esta última, quizá también por el mayor ruralismo de La Mancha (*buche*, general en *ALECMAN*, falta en *ADiM*). Este y otros casos nos llevan a plantear la viabilidad de un *continuum* areal entre las provincias de Guadalajara y Toledo, que abarcaría a Madrid. Ha de suponerse que cuando el oeste de Guadalajara y el NE de Toledo coinciden, esas formas coincidentes se podrían proyectar también a Madrid en etapas pasadas. Sin embargo, no puede descartarse que la ciudad de Madrid provocara una ruptura en la continuidad geográfica.

Varias son las situaciones posibles en la disposición del léxico por áreas. En primer lugar, Madrid, puede representar un espacio relativamente diferenciado, en el que sobresalen voces que apenas han sido registradas fuera de la Comunidad: *cerviajo*, *pobo*, *pasto siego*, *tierra ricial*. En otros casos, el léxico de Madrid se adscribe al espacio centro-norte. Así parece verse en algunas formas como *buey conejo* y *corte* 'pocilga'. Por último, apreciamos coincidencias con el sur, como *casillar* 'caja para la urdidera', documentado en Andalucía y en Arganda del Rey, y la variante *bielo* de *bieldo,* también localizada en el norte de Toledo en la actualidad.

Referencias bibliográficas

ADiM = García Mouton, Pilar/Isabel Molina Martos (2015): *Atlas Dialectal de Madrid*. <http://adim.cchs.csic.es/> [último acceso: 20/10/2017].

Agujetas Ortiz, María (2017): *Edición y estudio lingüístico de documentos del Archivo del Monasterio de El Escorial (ss. XVI–XVIII)*. Trabajo de Fin de Grado, Alcalá de Henares: Universidad de Alcalá.

ALDICAM-CM = Atlas lingüístico diacrónico e interactivo de la Comunidad de Madrid. <http://aldicam.blogspot.com.es/> [último acceso: 15/10/2017].

ALECMAN = Moreno Fernández, Francisco/Pilar García Mouton (2003): *Atlas Lingüístico (y Etnográfico) de Castilla-La Mancha*. <http://www.linguas.net/alecman/> [último acceso: 20/10/2017].

Alvar Ezquerra, Manuel (2011a): *Diccionario de madrileñismos*. Madrid: La Librería.

Alvar Ezquerra, Manuel (2011b): «Voces usadas en la Comunidad de Madrid con otras marcas diatópicas en el *DRAE*», *AEF* XXXIV, 5–21.

Álvarez Tejedor, Antonio (2007): «El dialecto leonés. El sur del dominio: Zamora», en José Ramón Morala Rodríguez (coord.), *Ramón Menéndez Pidal y el dialecto leonés (1906–2006)*. Burgos: Fundación Instituto Castellano-Leonés de la Lengua, 177–206.

Andreu, Alicia G. (1986): «Diálogo de voces en *Fortunata y Jacinta*», en David Kossof *et al*. (eds.), *Actas del octavo Congreso de la Asociación Internacional de Hispanistas*, I. Madrid: Ediciones Istmo, 153–158.

Autoridades = Real Academia Española (1726–1739): *Diccionario de la lengua castellana*. <http://web.frl.es/DA.html> [último acceso: 05/10/2017].

Carriazo Ruiz, José Ramón (2012): «Categorización, clasificación, repertorización onomasiológica y estudio lexicológico del vocabulario doméstico de los inventarios de bienes de San Millán: etnografía lingüística y dialectología de los Siglos de Oro», *Cuadernos del Instituto de Historia de la Lengua* V, 7, 125–141.

Castellón, Heraclia (2001): *El lenguaje administrativo: formas y uso. Información General*. Granada: La Vela.

Cestero Mancera, Ana María/Isabel Molina Martos/Florentino Paredes García (eds.) (2015): *Patrones sociolingüísticos de Madrid*. Berna: Peter Lang.

Cestero Mancera, Ana María (2015): «La atenuación lingüística en el habla de Madrid: un fenómeno sociopragmático variable», en Ana María Cestero Mancera, Isabel Molina Martos y Florentino Paredes García (eds.), *Patrones sociolingüísticos de Madrid*. Berna: Peter Lang, 365–412.

CODEA = *Corpus de Documentos Españoles Anteriores a 1800*. <http://corpuscodea.es/>. [último acceso: 11/10/2017].

CORDE = Real Academia Española: *Corpus Diacrónico del Español*. <http://www.rae.es/recursos/banco-de-datos/corde> [último acceso: 20/09/2017].

CorLexIn = Morala Rodríguez, José Ramón (dir.): *Corpus Léxico de Inventarios*. <http://web.frl.es/CORLEXIN.html> [último acceso: 11/10/2017].

DECH = Corominas, Joan/José Antonio Pascual (1980–1991): *Diccionario Crítico Etimológico Castellano e Hispánico*. Madrid: Gredos.

DLE = Real Academia Española: *Diccionario de la Lengua Española*. <http://dle.rae.es/?w=diccionario> [último acceso: 11/10/2017].

Fernández-Ordóñez Martín, Inés (2006): «Del Cantábrico a Toledo: el 'neutro de materia' hispánico en un contexto románico y tipológico (primera parte)», *Revista de Historia de la Lengua Española* I, 67–118.

Fernández-Ordóñez Martín, Inés (2007): «Del Cantábrico a Toledo: el 'neutro de materia' hispánico en un contexto románico y tipológico (segunda parte)», *Revista de Historia de la Lengua Española* II, 29–81.

Gallardo López, Álvaro (2017): *Documentación municipal madrileña del siglo XVIII: Camarma de Esteruelas*. Trabajo de Fin de Grado, Alcalá de Henares: Universidad de Alcalá.

Gómez Seibane, Sara/Delfina Vázquez Balonga (2013): «¿Más huellas del neutro de materia en los Siglos de Oro? Algunos testimonios de la región de Madrid», *Revista de Filoloxía Asturiana* 13, 53–70.

Jiménez Rayado, Eduardo (2010): *La villa y la tierra de Madrid en los albores de la capitalidad (siglos XIV–XVI)*. Madrid: Asociación Cultural Al-Mudayna.

Mañas García, Sandra (2017): *Edición y estudio de documentos municipales de Arganda del Rey*. Trabajo de Fin de Grado, Alcalá de Henares: Universidad de Alcalá.

Martínez San Juan, Miguel Ángel (1999): *Estudio y edición del* Pastor de Fílida *por Luis Gálvez de Montalvo*. Tesis doctoral, Universidad Complutense de Madrid. <http://biblioteca.ucm.es/tesis/19972000/H/3/H3065501.pdf> [último acceso: 21/10/2017].

Mediavilla Martín, Benito (2010): *Inventario de documentos. Real Biblioteca de El Escorial*. San Lorenzo de El Escorial: Ediciones Escurialeneses.

Miguel Franco, Ruth (2010): «Comentario morfosintáctico», en Florentino Paredes García (coord.), *Textos para la Historia del Español V. Archivo Municipal de Daganzo de Arriba*. Alcalá de Henares: Servicio de Publicaciones de la Universidad de Alcalá, 41–52.

Millares Carlo, Agustín (1932): *Glosario*, en Rafael Lapesa (transcrip.), *Introducción de Pedro Rico López. El Fuero de Madrid y los derechos locales castellanos, por Galo Sánchez y Sánchez. Texto del Fuero*. Madrid: Artes Gráficas Municipales.

Molina Martos, Isabel (2010): «Difusión social de una innovación lingüística. La intensificación en el habla de las jóvenes madrileñas», *Oralia: Análisis del discurso oral* 13, 197–214.

Molina Martos, Isabel (2015): «Estrategias de atenuación en el barrio de Salamanca de Madrid», en Ana María Cestero Mancera, Isabel Molina Martos y Florentino Paredes García (eds.), *Patrones sociolingüísticos de Madrid*. Berna: Peter Lang, 349–364.

Molina Martos, Isabel/Pilar García Mouton (2017): *Las hablas rurales de Madrid. Etnotextos*. Berna: Peter Lang.

Morala Rodríguez, José Ramón (2015): «Datos para la historia del 'neutro de materia' en castellano», *Revista de Filología Española* 95, 2, 307–337.

Moreno Fernández, Francisco (2015): «Hablar madrileño», en Ana María Cestero Mancera, Isabel Molina Martos y Florentino Paredes García (eds.), *Patrones sociolingüísticos de Madrid*. Berna: Peter Lang, 91–116.

Náñez Fernández, Emilio (1973): *El diminutivo: historia y funciones en el español clásico y moderno*. Madrid: Gredos.

NTLLE = Real Academia Española: *Nuevo Tesoro Lexicográfico de la Lengua Española*. <http://www.rae.es/recursos/diccionarios/diccionarios-anteriores-1726-1992/nuevo-tesoro-lexicografico> [último acceso: 30/10/2017].

Oudin, César (1607): *Tesoro de las dos lenguas francesa y española*. París: Marc Orry.

Paredes García, Florentino (2003): «Un cuadernillo del censo de vecinos de Alcalá de 1618», *Anales Complutenses* 15, 99–216.

Paredes García, Florentino (2005): *Textos para la Historia del Español III. Archivo Municipal de Alcalá de Henares*. Alcalá de Henares: Servicio de Publicaciones de la Universidad de Alcalá.

Paredes García, Florentino (coord.) (2010): *Textos para la Historia del Español V. Archivo Municipal de Daganzo de Arriba*. Alcalá de Henares: Servicio de Publicaciones de la Universidad de Alcalá.

Paredes García, Florentino (2015a): «Funciones subjetivadoras del diminutivo en el habla de Madrid», en Ana María Cestero Mancera, Isabel Molina Martos y Florentino Paredes García (eds.), *Patrones sociolingüísticos de Madrid*. Berna: Peter Lang, 117–153.

Paredes García, Florentino (2015b): «Nuevos datos sobre el uso y las funciones de los pronombres átonos de tercera persona en Madrid», en Ana María Cestero Mancera, Isabel Molina Martos y Florentino Paredes García (eds.), *Patrones sociolingüísticos de Madrid*. Berna: Peter Lang, 177–125.

Penadés Martínez, Inmaculada (2016): «Las locuciones verbales en el habla de Madrid (distrito de Salamanca)», en Ana María Cestero Mancera, Isabel Molina Martos y Florentino Paredes García (eds.), *Patrones sociolingüísticos de Madrid*. Berna: Peter Lang, 251–286.

Russell, Robert H. (1982): «De Fortunata y su habla», en Eugenio de Bustos (coord.), *Actas del Cuarto Congreso Internacional de Hispanistas*, II. Salamanca: Universidad de Salamanca, 543–551.

Sacristán Jerez, Julián (1990): *El habla del valle de Lozoya*. Tesis doctoral. Madrid: Universidad Complutense de Madrid.

Sánchez-Prieto Borja, Pedro (2000): «La preposición *a* con valor 'lugar en donde' en castellano antiguo», en Annick Englebert *et al.* (eds.), *Actes du XXIIe Congrès International de Linguistique et de Philologie Romanes, Bruxelles, 23–29 juillet 1998*, II. Tubinga: Max Niemeyer Verlag, 393–406.

Sánchez-Prieto Borja, Pedro/Delfina Vázquez Balonga (2017): «Hacia un corpus de beneficencia en Madrid (siglos XVI–XIX)», *Scriptum Digital* 6, 83–103. <http://www.scriptumdigital.org/documents/06_SD06_03_SanchezPrieto_VazquezBalonga.pdf> [último acceso: 25/10/2017].

Sánchez-Prieto Borja, Pedro/Delfina Vázquez Balonga (2018): «Toledo frente a Madrid en la conformación del español moderno: el sistema pronominal átono», *Revista de Filología Española*, 98, 1, 185–215.

Sánchez-Prieto Borja, Pedro/Ana Flores Ramírez (2006): *Textos para la Historia del Español IV. Archivo Regional de la Comunidad de Madrid*. Alcalá de Henares: Servicio de Publicaciones de la Universidad de Alcalá.

Sancho Pascual, María (2015): «Integración sociolingüística de los inmigrantes ecuatorianos en Madrid. Datos sobre el diminutivo según los corpus Ispie-Madrid y *PRESEEA-Madrid*», en Ana María Cestero Mancera, Isabel Molina Martos y Florentino Paredes García (eds.), *Patrones sociolingüísticos de Madrid*. Berna: Peter Lang, 155–176.

Seco, Manuel (1970): *El habla de Madrid en el teatro de Arniches*. Madrid: Alfaguara.

Terreros y Pando, Esteban de (1786–1788): *Diccionario castellano con las voces de ciencias y artes correspondientes en las tres lenguas francesa, latina e italiana*. Madrid: Viuda de Ibarra.

TLHA = Alvar Ezquerra, Manuel (2000): *Tesoro Léxico de las Hablas Andaluzas*. Madrid: Arco Libros.

Torrens Álvarez, María Jesús (2002): *Edición y estudio lingüístico del Fuero de Alcalá (Fuero Viejo)*. Alcalá de Henares: Fundación Colegio del Rey.

Torres Martínez, Marta (2014): «Inventarios de bienes en establecimientos benéficos jienenses (siglo XIX)», en María Águeda Moreno Moreno y Marta Torres Martínez (eds.), *Estudios de léxico histórico español*. Salamanca: Luso Española Ediciones, 129–265.

Varela Merino, Elena (2009): *Los galicismos en el español de los siglos XVI y XVII*. Madrid: Consejo Superior de Investigaciones Científicas.

Vázquez Balonga, Delfina (2014): *Textos para la Historia del Español VIII. Archivo Municipal de Arganda del Rey*. Alcalá de Henares: Servicio de Publicaciones de la Universidad de Alcalá.

Vázquez Balonga, Delfina (2015): *Léxico en la documentación de Toledo y Madrid en los siglos XVI–XVII*. Tesis doctoral. Alcalá de Henares: Universidad de Alcalá.

Vázquez Balonga, Delfina/Pedro Sánchez-Prieto Borja (2015): «¿Seseo en el centro peninsular?», *Revista de Historia de la Lengua Española*, 10, 201–217.

Vittori, Girolamo (1609): *Tesoro de las tres lenguas francesa, italiana y española*. Ginebra: Philippe Albert & Alexandre Pernet.

María Jesús Torrens Álvarez

El proyecto *ALDICAM-CM* y el ejemplo de los fueros de Alcalá para el estudio de la historia del léxico[1]

Resumen: Presentamos el *Altas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid*, un proyecto por completo innovador que permitirá la visualización inmediata sobre un mapa de los resultados de cualquier tipo de búsqueda, simple o compleja, sobre los documentos del corpus. Dicho corpus estará formado por un amplio número de documentos archivísticos madrileños de los siglos XIII al XIX. El *ALDICAM-CM* permitirá conocer mejor la lengua hablada y escrita en Madrid y comprobar la hipótesis de que esta variedad contribuyó de manera decisiva a la formación del español moderno. Asimismo, ejemplificamos el interés de este proyecto para la historia del léxico con un análisis de los fueros de Alcalá, de los siglos XIII y XVI, respectivamente.

Palabras clave: *ALDICAM-CM*, Madrid, Dialectología diacrónica, Fueros de Alcalá

Abstract: The *Interactive Diachronic Linguistic Corpus of the Comunidad de Madrid* is a completely innovative project, because it will allow to visualise the results of any simple or complex search in the documents of the corpus as a map. This corpus will consist of a large number of archival documents of Madrid written from the 13th to 19th centuries. *ALDICAM-CM* will allow more knowledge about the language spoken and written in Madrid, and will possibly support the hypothesis that Madrid's variety was decisive in the formation of modern Spanish. Likewise, we exemplify the interest of this corpus in the history of the lexicon with an analysis of the «fueros» of Alcalá, of the thirteenth and sixteenth centuries, respectively.

Keywords: *ALDICAM-CM*, Madrid, Diachronic dialectology, «Fueros» of Alcalá

1 Introducción y antecedentes

Hoy en día resulta poco menos que impensable trabajar en historia de la lengua sin el fundamento empírico que proporcionan los corpus textuales informatizados.

1 Este trabajo se enmarca en el proyecto de investigación financiado por la Comunidad de Madrid *Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid*, ref. S2015/HUM-3443-ALDICAM-CM.

La edición digital y la creación de corpus diacrónicos han dado lugar al desarrollo tanto de las Humanidades digitales, como de la Lingüística de corpus, al hacer posible el análisis cuantitativo de un volumen de datos lingüísticos inimaginable hace pocos años.

El problema, como señalan Rodríguez Molina/Octavio de Toledo y Huerta (2017) para el caso del *CORDE*, pero que puede extenderse a otros corpus y bases de datos más limitados que el académico, es que este incremento cuantitativo «no ha venido acompañado de una reflexión pareja que pondere la dimensión cualitativa». Estos autores se ocupan de un asunto fundamental: la indistinción entre texto y testimonio, y entre autor y copista, que lleva a atribuir a copias tardías la fecha y autoría del original no conservado, lo que compromete la fiabilidad filológica de muchos de los textos medievales recogidos en el *CORDE* como fuente de datos lingüísticos. Esto ocurre principalmente con los textos literarios medievales, en su mayoría transmitidos por testimonios muy posteriores al original y con una historia textual a veces imposible de reconstruir.

No menos importante, por cuanto atañe directamente a esa misma fiabilidad filológica de los textos, es la diversidad de criterios de presentación gráfica y el grado de intervención de los responsables de cada edición, desde la absoluta fidelidad paleográfica al manuscrito, a la modernización gráfica e incluso morfológica, pasando por numerosas soluciones intermedias.

Esta disparidad de criterios característica de los grandes corpus y bases de datos hechos a partir de ediciones previas, sumada a la falta de lematización, no solo se traduce en las dificultades prácticas para la búsqueda de palabras, de las que hay que imaginar todas las formas posibles que pudo adoptar a lo largo de la historia o que han podido darle sus editores, sino que hará inviables las investigaciones sobre los niveles gráfico y fonético-fonológico. Con esto en absoluto queremos decir que las ediciones normalizadas (que no modernizadas) no sean válidas para los estudios de historia de la lengua: los análisis sintácticos o léxicos, objeto primordial para los que participamos en este libro, se verán beneficiados de unas ediciones críticas que palíen en cierta medida la ausencia de lematización y de posibilidades de marcación gramatical de los textos, al diferenciar, por ejemplo, entre *del*, *d'él* o *de·l* 'dele' lo que en el manuscrito es siempre *del*. Habrán de ser conscientes, eso sí, de que muchas de las soluciones presentadas serán el resultado de la interpretación del editor, obligado a posicionarse ante los innumerables problemas filológicos que el texto plantea.

Estos dos grandes problemas de índole filológica tienen una implicación lingüística fácil de entender, pero recurrentemente infravalorada o, más bien, ignorada. Durante décadas la historia de la lengua española, con la salvedad de la época de los orígenes, se ha escrito a partir de los textos literarios, a pesar de los

graves inconvenientes señalados, no por bien conocidos, tenidos en consideración. En cuanto a los criterios de edición, parece claro que una única presentación no puede satisfacer las diferentes necesidades de los usuarios, pues la forma del texto determina en gran medida las posibilidades de su explotación mediante los sistemas de búsqueda informática.

En este marco de necesidades insatisfechas vio la luz la red *CHARTA* (www.charta.es), que aglutina en estos momentos una veintena de grupos de investigación con un objetivo principal: la creación de un corpus electrónico de textos archivísticos de los siglos X al XIX, en su inmensa mayoría originales con data explícita, y con aspiraciones de que esté representado todo el ámbito geográfico hispánico (www.corpuscharta.es). Los documentos son editados por los miembros de *CHARTA* con unos criterios comunes, y se ofrecen en una triple presentación: paleográfica, que reproduce con fidelidad la escritura del documento; crítica, que normaliza los usos gráficos, la puntuación… para facilitar la intelección del texto; y la imagen digital, siempre que se tengan los permisos pertinentes (Sánchez González de Herrero *et al*. 2013).

Los presupuestos metodológicos de *CHARTA* y su corpus son los que están en la base del *Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid* (*ALDICAM-CM*, http://aldicam.blogspot.com.es/), proyecto financiado por la Comunidad de Madrid y coordinado por Pedro Sánchez-Prieto Borja[2]. A partir de un amplio corpus de documentos de entre los siglos XIII (no los hay anteriores) al XIX escritos por personas de los municipios que hoy constituyen esta comunidad autónoma, se pretende ofrecer unos materiales que sirvan de base para estudios lingüísticos diacrónicos a fin de llegar a conocer, con el debido apoyo en los datos, tanto las posibles variedades y variaciones internas, como el papel del habla de Madrid en la configuración del español moderno.

2 El proyecto cuenta con el consorcio de tres grupos de investigación: el Grupo de Investigación de Textos para la Historia del Español (GITHE), de la Universidad de Alcalá y coordinado por Pedro Sánchez-Prieto Borja; el grupo Cambio Lingüístico e Historia del Español (CaLiHE), del Consejo Superior de Investigaciones Científicas y coordinado por María Jesús Torrens Álvarez; y el grupo Archivos y Patrimonio Documental (ArPaDoc), de la Universidad Complutense de Madrid y coordinado por Concepción Mendo Carmona. Participan, asimismo, otros cinco grupos asociados: CiTeHi (Alicia Sánchez Díez, Universidad Complutense de Madrid), PADICE (María del Val González de la Peña, Universidad de Alcalá), GRANATVM (Javier Rodríguez Molina, Universidad de Granada), LexUNED (José Ramón Carriazo, UNED) y ReLiR (Bautista Horcajada, Universidad Complutense de Madrid).

La sincronía actual del habla madrileña ha recibido la atención de sociolingüistas y dialectólogos, como García Mouton/Molina (2009, 2015, 2017), Fernández-Ordóñez (2005) o Cestero/Molina/Paredes (2015). Pilar García Mouton e Isabel Molina (2015) son las responsables del *Atlas Dialectal de Madrid* (*ADiM*), para el que han encuestado dieciséis localidades rurales de toda la periferia de la Comunidad de Madrid, con 1188 mapas léxicos en su primera etapa que se irán publicando en la web progresivamente, así como una selección de archivos sonoros y etnotextos (estos últimos editados en formato libro en 2017). También está representado el espacio rural madrileño, con cuatro enclaves encuestados, en el *Corpus Oral y Sonoro del Español Rural* (*COSER*), dirigido por Inés Fernández-Ordóñez y pensado para el análisis de la variación dialectal de diversos aspectos gramaticales. Ambos trabajos siguen los presupuestos metodológicos de la dialectología tradicional, esto es, la entrevista a informantes rurales, de edad avanzada y escasa escolarización, que en lo posible hayan vivido siempre en su lugar de nacimiento, para garantizar la escasa influencia en su habla de otras variedades lingüísticas.

En cuanto a estudios diacrónicos se refiere, han sido cuatro los volúmenes de la colección *Textos para la Historia del Español*, dirigida por Pedro Sánchez-Prieto, dedicados a la documentación histórica madrileña: el III, sobre fondos del Archivo Municipal de Alcalá de Henares; el IV, con documentos del Archivo Regional de la Comunidad de Madrid; el V, dedicado a Daganzo; y el VIII, a Arganda del Rey, publicaciones de las que se han derivado otras aportaciones que profundizan en aspectos lingüísticos concretos.

No es en absoluto desconocida la importancia de la llamada «norma madrileña» en la configuración del español moderno. El traslado de la Corte del reino a Madrid en 1561 y el prestigio asociado a los comportamientos lingüísticos de los cortesanos favoreció la difusión de dichos rasgos, que se extenderían en mayor o menor medida en la península ibérica y en América (Bustos/Santiago 2002), siendo posiblemente también centro de innovación de fenómenos como el laísmo (Sánchez-Prieto/Vázquez 2018). Los rasgos lingüísticos de esa norma madrileña de los Siglos de Oro, empero, no han sido todavía debidamente analizados y descritos, ni tampoco los posibles cambios «de abajo arriba» y los procesos de nivelación lingüística motivados por la fuerte e incesante llegada de inmigrantes del área rural y de la más variada procedencia geográfica. Para ello resulta imprescindible contar con un nutrido corpus de textos antiguos escritos por personas naturales y vecinas de las poblaciones que hoy forman la Comunidad de Madrid, textos sobre aspectos sociales, económicos, culturales… propios del lugar o de su entorno, y a ser posible representativos de una diversidad sociolingüística amplia en la que tengan la mayor cabida los escribientes «inhábiles» o de baja formación, que pueden reflejar de manera más directa el habla popular (muestras de gran valor son

las notas de abandono de niños en la Inclusa de Madrid, Sánchez-Prieto/Flores 2005). Pero, como es lógico, estos testimonios son siempre escasos y se puede decir que inexistentes antes del siglo XVII, pues no solo el nivel de alfabetización de la sociedad disminuye a medida que retrocedemos en la historia, sino que los escritos que han tenido más posibilidades de conservarse son necesariamente los que resultaban de importancia e interés para los grupos sociales dominantes.

2 Objetivos y características

Las características fundamentales del proyecto se explicitan en su propio título: *Atlas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid*. Se persigue la creación de un atlas lingüístico, pues los datos lingüísticos buscados por el usuario tendrán una salida directa a un mapa de la Comunidad de Madrid, lo que permitirá visualizar de manera inmediata la distribución geográfica de variantes de todo tipo: gráficofonéticas (*escribir-escrevir…, fazer-hazer-hacer, ierba-hierba-yerba*…), morfológicas (*avía-avié, el-la eglesia*, sufijos, superlativos…), sintácticas (cuestiones de concordancia de género y número, usos de los tiempos y modos verbales…) o léxicas (formas para referirse, por ejemplo, a un animal como el *cerdo-puerco-cochino-marrano*, onomástica y toponimia…). Al menos en un principio, el corpus no estará lematizado ni tendrá marcas morfológicas o sintácticas que den el trabajo hecho al investigador, sino que habrá de ser este quien introduzca las palabras o secuencias que desee o espere encontrar. No obstante, el hecho de que, aparte de los mapas, se dé al usuario acceso a la lectura secuencial del texto íntegro, ofrecido en la triple presentación ya señalada (imagen digital, transcripción paleográfica y presentación crítica), y de que los resultados de sus consultas también se ofrezcan en forma de concordancias, con el contexto anterior y posterior a la secuencia buscada, hace que la tarea de extracción e interpretación de los fenómenos morfosintácticos objeto de estudio se vea muy facilitada.

Naturalmente, para que las consultas obtengan resultados es necesario que previamente se haya construido un corpus textual amplio de documentos archivísticos, cuya cronología para la Comunidad de Madrid va del siglo XIII, fecha de los documentos romances más antiguos conservados, al siglo XIX. Hay que decir que la representación de los distintos siglos es necesariamente muy desigual, pues es muy escasa la documentación medieval, mientras que del XVI en adelante los materiales son muy abundantes[3]. Aspiramos a ofrecer 1500

3 Debido al despegue urbanístico no solo de la capital, sino de las llamadas «ciudades dormitorio», no son pocos los municipios madrileños actuales que solo disponen de fondos de los siglos XVIII y XIX, o incluso solo del XIX.

documentos seleccionados y transcritos por los miembros del proyecto, si bien una de las principales ventajas de los corpus electrónicos es que siempre podrán seguir creciendo.

Este atlas lingüístico diacrónico, a diferencia de los pocos existentes en otros ámbitos románicos diseñados para un periodo histórico concreto, como son los del francés o el inglés medieval, y a diferencia también de los atlas lingüísticos de las lenguas modernas basados en encuestas dirigidas, es interactivo, lo que significa que en lugar de ofrecer mapas fijos dibujados a partir de la información lingüística seleccionada y procesada por los investigadores responsables del atlas, vierte sobre un mapa digital los resultados de cualquier búsqueda que el usuario realice. Se podrán generar, así, tantos mapas como consultas sobre el corpus se conciban, lo que a su vez lo convierte en un instrumento válido y atractivo tanto para especialistas como para el público en general.

Para ello el usuario dispondrá de una serie de parámetros que podrá combinar según sus intereses: cronología, localidad, materia o tema, tipología documental, canon paleográfico… Esta interacción entre los distintos factores, principalmente de naturaleza temática, diacrónica, diatópica y sociolingüística, ofrece unas posibilidades de investigación que van mucho más allá de la dialectología histórica y la historia de la lengua, pues es igualmente útil para historiadores de los más variados aspectos de la sociedad y la vida madrileña en los siglos pasados. De hecho, el equipo lo formamos filólogos, paleógrafos e historiadores, que trabajamos en nuestras respectivas parcelas, pero también conjuntamente, pues la lengua es en gran medida manifestación de la realidad social de sus hablantes y de un contexto histórico dado.

Asimismo, aparte de la explotación que del *ALDICAM-CM* puedan hacer los especialistas, creemos que sus características hacen de él un material con grandes posibilidades de aplicación en el ámbito de la enseñanza en niveles preuniversitarios, por lo que está entre nuestros objetivos el diseño de recursos didácticos. También para el público general, que podrá satisfacer las curiosidades más diversas sobre la historia de los municipios madrileños, convencidos como estamos de que la divulgación ha de ser objetivo esencial de nuestro proyecto.

Por las razones señaladas, el *ALDICAM-CM* es un proyecto por completo innovador, que aúna las Humanidades digitales en su desarrollo más puntero con el trabajo histórico y filológico riguroso, con el objetivo de ofrecer a todo tipo de usuarios el acceso gratuito a un corpus de documentos archivísticos fiable en su descripción y transcripción, base de los más diversos estudios y consultas. Para nuestros intereses en particular, el corpus electrónico y las herramientas de explotación harán de él una fuente imprescindible para el análisis lingüístico de la variedad o variedades madrileñas a lo largo de la historia, al permitir

establecer con datos seguros y cuantificables los fenómenos y su distribución geográfica a lo largo del tiempo.

3 Metodología

Para el desarrollo del *ALDICAM-CM* tal y como se ha descrito se necesita el trabajo colaborativo e interdisciplinar de filólogos, paleógrafos, historiadores, archiveros y especialistas en Humanidades digitales. La metodología para la confección del corpus y para parte de los desarrollos informáticos cuenta con el respaldo de toda la labor ya realizada en el marco de la red internacional *CHARTA*, de la que la mayoría de los integrantes del equipo investigador formamos parte desde sus inicios.

3.1 Documentatión de archivos

El primer paso y fundamental, también por la ingente cantidad de horas que requiere, es el trabajo de campo en los archivos, entre los que existe gran disparidad en cuanto a la riqueza de los fondos, su organización, accesibilidad y facilidades para obtener una reproducción digital de los documentos seleccionados. Hay que tener en cuenta que, a excepción del Archivo Regional de la Comunidad de Madrid, que reúne una abrumadora y valiosísima documentación de carácter exclusivamente histórico, los archivos municipales son archivos vivos, en los que su personal ha de estar dedicado principalmente a las solicitudes diarias del propio ayuntamiento o de los ciudadanos. Esto hace que en muchos no existan catálogos del fondo antiguo, y que incluso los sucintos inventarios no estén a disposición de los investigadores, con lo que se requiere la generosa colaboración de los archiveros responsables para conocer la naturaleza de los fondos que custodian.

Hasta el momento han sido algo más de una veintena los archivos trabajados, sitos en las localidades que se muestran en el mapa accesible en el enlace http://aldicam.blogspot.com.es/p/archivos-utilizados.html. En Madrid capital se hallan los más ricos en diversidad tipológica y cronológica, el ya citado Archivo Regional y el Archivo de Villa, pero todos los municipios visitados ofrecen materiales valiosos. Por otra parte, aunque el grueso de la documentación de un corpus de estas características se halla necesariamente en los fondos municipales (los de los pueblos más pequeños o sin archivo propio, depositados en el Archivo Regional), también son de interés los de algunas instituciones religiosas o benéficas, como las cartas de profesión de religiosas del convento de las Bernardas de Alcalá de Henares, o las notas de abandono de niños, certificados de defunción e informes médicos de la Inclusa de Madrid y la Hermandad del Refugio, fundada en 1615.

En cuanto a la selección de los documentos, son varios los criterios que se tienen en cuenta. Por ejemplo, la escasez de documentación medieval ha determinado que incluyamos la mayor parte de los manuscritos conservados de este periodo, redactados, claro está, dentro de los límites de la actual Comunidad de Madrid[4]. De manera similar, procuraremos que estén representados el mayor número de municipios, aunque de algunos apenas se conserven escritos antiguos. Pero es precisamente esta desigualdad geográfica y cronológica en el volumen de los fondos la que obliga a aplicar unos criterios de selección a medida que avanzamos en el tiempo, por hacerse la documentación mucho más abundante y variada.

Interesan los textos de carácter más local, que traten de aspectos de la vida cotidiana, de los medios de subsistencia, del entorno inmediato, de las fiestas y costumbres, de los motivos de litigio entre los vecinos…, porque en ellos también podremos encontrar muestras de la lengua y del léxico autóctonos.

3.2 Criterios de transcripción

El segundo paso es la transcripción de los documentos definitivamente seleccionados para su inclusión en el corpus, labor para la que se aplican unos criterios comunes largamente discutidos y ensayados en la red internacional *CHARTA* (http://www.redcharta.es/criterios-de-edicion/). De cada documento se ofrece 1) una transcripción paleográfica, que respeta los usos gráficos del manuscrito, la unión y separación de palabras, la puntuación o el uso de mayúsculas y minúsculas, y que es idónea para estudiar los aspectos gráfico-fonéticos, así como cuestiones tales como los procesos de formación de palabras o la puntuación antigua; 2) una presentación crítica, que, fundamentada en el conocimiento de la historia de la lengua y de la escritura, normaliza usos gráficos e interviene en la unión y separación de palabras, puntuación o acentuación, facilitando así los análisis morfológicos, sintácticos, léxicos y discursivos a partir de la intelección del texto propuesta por el editor; y 3) siempre que los archivos titulares den su consentimiento, la imagen digital del documento, que permite realizar estudios paleográficos y diplomáticos, además de posibilitar la comprobación de la corrección de las transcripciones.

4 Desgraciadamente, en muchos lugares de los tiempos más remotos solo se han conservado los documentos debidos a la autoridad real, por ser considerados los más (o únicos) importantes.

3.3 Metadatos

Cada documento va acompañado de una cabecera descriptiva en la que, aparte de los datos puramente identificativos (corpus, número de identificación dentro del corpus, archivo de procedencia, signatura archivística), proporciona información precisa sobre el lugar de emisión, la fecha, materia o tema, tipología documental y paleográfica, posible elaboración femenina, escribano y regesto, para lo que se requiere un conocimiento completo del documento en cuestión, de su tenor y de su contexto geohistórico.

La autoría y responsabilidad de las ediciones de cada unidad documental y de su cabecera recae en el transcriptor y en los revisores, cuyos nombres aparecerán siempre tras la cabecera.

3.4 Base de datos relacional

Los documentos ya preparados en su triple presentación y con cabecera serán incorporados a una base de datos relacional elaborada expresamente para el proyecto, accesible en abierto a través de una página web que contendrá otros diversos apartados relevantes sobre el proyecto. Los textos podrán leerse en su forma íntegra o se podrán hacer búsquedas de secuencias sin límite de extensión, que nos devolverán los resultados encontrados en forma de concordancias, esto es, con el contexto anterior y posterior a la secuencia buscada.

Los campos señalados para la cabecera descriptiva de cada documento serán los que sirvan de parámetros y filtros para las consultas en el corpus, con la combinación de cuantos deseemos. De esta forma, se podrán acotar las búsquedas por la cronología (arco cronológico o año exacto); la geografía (una, varias o todas las localidades de la Comunidad de Madrid con representación en el corpus); la materia o tema; la tipología documental; el canon paleográfico; el escribano; o si es mujer la autora material o, al menos, la otorgante del documento.

No obstante las posibilidades y ventajas que estos recursos ofrecen para la explotación del corpus, la innovación mayor, hasta la fecha no ensayada, es la proyección directa e inmediata de los resultados a un mapa de la Comunidad de Madrid, desarrollos informáticos encomendados a expertos en Humanidades digitales.

4 Resultados y perspectivas futuras

En los más de dos años de desarrollo del proyecto han sido muchos los avances conseguidos, de los que se va dando debida cuenta en la web del *ALDICAM-CM* (http://aldicam.blogspot.com.es/), aunque no es poca la tarea que nos queda por

delante. La intención es ir subiendo lotes de documentos al corpus, de manera que a mediados del año 2018 pongamos a disposición de los usuarios un número significativo que permita ensayar las búsquedas, número que iremos incrementando en sucesivas entregas.

La visita a los archivos ha permitido la elaboración de una completa tabla de los fondos documentales de los archivos municipales y del Archivo Regional, con el cuadro de clasificación temática de cada centro y las fechas de los documentos más antiguos de cada sección (http://aldicam.blogspot.com.es/p/material-de-trabajo.html). En cuanto a los documentos ya transcritos, en el momento de la redacción de este trabajo contamos con más de 700, cuya relación puede consultarse en la tabla facilitada en el enlace http://aldicam.blogspot.com.es/p/material-de-trabajo.html.

Remitimos igualmente a la web del *ALDICAM-CM* para conocer las publicaciones científicas y presentaciones en congresos a las que va dando lugar el proyecto, así como los numerosos cursos y actividades de formación y de divulgación, o la sección «Historias en los documentos», en la que se comentan textos de muy variada naturaleza cuyo contenido puede resultar especialmente curioso para el público general.

No es este el lugar para detallar los resultados de las investigaciones ya publicadas o en prensa, pero no estará de más reproducir aquí las palabras de Pedro Sánchez-Prieto (2017) sobre lo que el *ALDICAM-CM* está permitiendo avanzar en el conocimiento de la norma madrileña y su aportación a la configuración del español moderno:

> A través de los escritos de la administración, que llegan a todo el reino, de la imprenta, y más tarde de la prensa y los medios de comunicación, Madrid expande sus usos lingüísticos. El leísmo, no es, desde luego, una innovación madrileña, pero su arraigo en Madrid contribuyó en gran medida a que triunfara dentro de la norma culta peninsular. Lo mismo sucedió con variantes como *ahora* frente al antiguo *agora*. Y otro tanto cabe decir del superlativo en *-ísimo* (*excelentísimo*, *serenísimo*) para los tratamientos, pues quienes venían a Madrid a pedir favores o empleos, los usaban en sus solicitudes como una forma de captar la benevolencia de sus destinarios. En los documentos hemos comprobado que el yeísmo estaba extendido entre las clases populares en el s. XIX (*apeyido*, *Viya*). El habla de Madrid se convierte en un modelo a imitar, primero en el espacio de lo que hoy es la Comunidad, y luego en todo el ámbito de la nación. No es exagerado decir que la formación del español moderno no se entiende sin el habla de Madrid.

A continuación, como ejemplo de los documentos que integrarán el *ALDICAM-CM* y de las posibilidades que ofrecen para el estudio del léxico y su evolución histórica, presentaremos una breve investigación basada en la comparación de los fueros de Alcalá de Henares, separados entre sí por más de 250 años.

5 Los fueros de Alcalá de Henares como fuente para el estudio del léxico

Los fueros de Alcalá de Henares, el llamado *Viejo* (*FVA*), de la primera mitad del siglo XIII, y el *Nuevo* (*FNA*), de comienzos del siglo XVI[5], son documentos muy singulares tanto en su individualidad, como en su comparación. Conservados ambos en los manuscritos originales firmados de propia mano de sus respectivos otorgantes (el arzobispo Jiménez de Rada y el cardenal Cisneros), son una fuente privilegiada para el estudio de la historia de la lengua y de los procesos de reelaboración discursiva, dado que el *FNA* es en realidad una actualización lingüística del *FVA*, ordenada por el cardenal Cisneros para facilitar su comprensión y consiguiente aplicación y cumplimiento. Como se dice en su preámbulo, informado el cardenal de que en el *FVA* «avía muchas leyes e ordenanças que non se usan ni guardan e otras que están escriptas por tales palabras o vocablos que non se pueden bien entender e han menester declaración», de lo que se derivaban serios perjuicios para sus vasallos, ordena remediar la situación, para lo que, dice, «mandamos quitar y quitamos las leyes del fuero que non eran usadas nin guardadas, y algunas corregimos y enmendamos y otras declaramos así en las cantidades de las monedas como en muchas palabras escuras y non usadas que las leyes del dicho fuero tenía» (Torrens/Sánchez Moltó 2011: 142)[6].

Que a comienzos del siglo XVI se otorgue un fuero municipal es una rareza que solo se explica por el carácter señorial del territorio de Alcalá y sus aldeas, pertenecientes al señorío arzobispal de Toledo hasta fecha tardía. De hecho, las villas de Alcalá y Santorcaz permanecieron bajo la jurisdicción de los arzobispos hasta la extinción de los señoríos, esto es, hasta el final del Antiguo Régimen (Sánchez Moltó 2011: 107–108), si bien es de suponer que fueran ya pocas las leyes vigentes[7].

5 El manuscrito firmado por Cisneros está fechado en 1509, pero gracias a un traslado hecho para Loeches, sabemos que la primera redacción del *FNA* fue de 1501. Ambos testimonios, el manuscrito de Alcalá y el de Loeches, copian ese original de 1501 hoy perdido (Torrens 2011, Sánchez Moltó 2011).

6 Para los fragmentos reproducidos de los dos fueros seguimos nuestras propias ediciones críticas, Torrens (2002) para el *FVA* y Torrens/Sánchez Moltó (2011) para el *FNA*. Sobre la relación entre estos dos textos, v. también Torrens (2011 y 2012).

7 Todavía en 1646, según informan los *Anales Complutenses*, muchas leyes «se conservan con fuerça y vigor» (*apud* Sánchez Moltó 2011: 107).

En la época de elaboración del *FNA*, hacía tiempo que los fueros locales habían sido sustituidos en los aspectos más importantes por los códigos legislativos de ámbito supramunicipal[8], y en los de carácter más estrictamente local, por las ordenanzas municipales (Pérez Bustamante 1986: 752). De hecho, el que hemos llamado *FNA* en realidad lleva por título «Ordenanças y fuero de la villa de Alcalá de Henares», con unas pocas ordenanzas de nueva creación incorporadas tras el renovado fuero. Este, empero, respeta la tradición discursiva de los fueros semiextensos romances, característicos del siglo XIII (Kabatek 2001 y 2004).

En efecto, el cotejo entre los dos textos muestra que el redactor del *FNA* interviene en distinto grado a la hora de actualizar lingüísticamente el *FVA*, modificando lo necesario en todos los niveles de lengua (gráfico-fonético, morfológico, sintáctico, léxico y discursivo) para que el resultado sea comprensible en el siglo XVI, pero procurando por lo general mantenerse fiel a la versión medieval, lo que no debe interpretarse como simple inercia en la copia, sino como deseo de garantizar que las leyes del *FVA* aún vigentes sean fácilmente reconocibles a pesar de su modernización formal.

Esto hará que en muchos aspectos la lengua del *FNA* resulte arcaizante, cuestión esta que no podremos analizar con el detenimiento que se merece, si bien el *FNA* muestra un discurso claramente más elaborado y técnico que el *FVA*. A este respecto, hay que recordar la naturaleza performativa de estos textos de aplicación del Derecho, en los que el acto mismo de jurisdicción se realiza mediante la formulación de la ley. Son textos redactados por especialistas y escritos por notarios profesionales a instancias de una autoridad reconocida, pero la comunicación no se da exclusivamente entre especialistas, sino que va dirigida también a un público general que ha de entender los preceptos y las sanciones que de su incumplimiento se deriven. Como consecuencia, los fueros están a medio camino entre el discurso ordinario y el de especialidad, pero es evidente que el desarrollo de la disciplina jurídica y la práctica de la jurisprudencia tienen un reflejo directo en la nueva redacción del *FNA*, que presenta un discurso más especializado y con un número mucho mayor de tecnicismos.

El *FNA* muestra cierta preferencia por los sustantivos para la expresión de conceptos que en el *FVA* se manifestaban mediante oraciones con verbo conjugado. Veamos un par de ejemplos:

8 Alonso Romero (1982), nos recuerda que el Derecho común romano-canónico no llega a Castilla hasta mediados del siglo XIII, con Alfonso X, pero hasta finales del siglo XV no adquirió sus perfiles definitivos y pleno arraigo en la práctica.

FVA	*FNA*
196. Júdez o fiadores, si les viniere mandado d'aldeas d'Alcalá *que mataron omne o firieron, o que levaron muger rábida o forzada*, e allá foren e quisieren *prender omnes o aver* o prisieren e alguno lo emparare, o lo toliere o lo sobrelevare e negárelo, con III bezinos lo firmen, e si non pudiere firmar jure con II bezinos e pártanse d'él.	XLVIII. Si los alcaldes de Alcalá o alguno d'ellos fuere a algún aldea o logar de la tierra *sobre muerte, o fuerça o otro caso semejante* e quisiere *prender a alguno o secrestar los bienes* e algunos lo contradixeren o lo resistieren, allende de la pena corporal pague cinco mil maravedís, la meitad al señor e la meitad al alcalde.

Las acciones de 'matar a un hombre' o 'llevar a una mujer forzada' de la ley 196 del *FVA* se sustituyen respectivamente por *muerte* y *fuerça*, sin necesidad de complemento del nombre, en la correspondiente del *FNA*. No se sustantiva «prender omnes o aver», pero en la nueva redacción se utiliza el tecnicismo *secrestar* para los bienes, es decir, 'embargar', palabra que, según el *DECH* s. v. *secuestrar* (< lat. SEQUESTRARE 'depositar judicialmente en poder de un mediador'), se documenta por primera vez en Nebrija y, en la forma *secrestar*, en 1503 en la *Nueva Recopilación (apud Diccionario de Autoridades)*. Hoy *CORDE* nos devuelve un primer caso fiable, según el «corderógrafo» de Molina/Octavio (2017), de 1380–85 en el *Libro de Palladio*, con el significado de 'apartarse o retirarse (una persona)', y ya referido a «plata y cobres», en el Pentateuco del Esc. I-j-4, manuscrito del siglo XV, más otras casi 190 apariciones de *secrestar* y familia hasta el año 1500. Parece que la acepción judicial que hallamos en el *FNA*, esto es, la etimológica, fue la que tuvo más éxito, y es la única que recoge *Autoridades*.

Volviendo a los ejemplos de empleo de un sustantivo en lugar de una oración, el segundo caso resulta muy revelador de otra característica que diferencia las redacciones del siglo XIII y del XVI, ya que en el XIII lo habitual es enumerar todos los casos concretos posibles, mientras que en el XVI se tiende al uso de hiperónimos y generalizaciones:

FVA	*FNA*
113. Todo omne qui *dixiere ad otro gafo, o fudud'in culo, o cornudo provado o alevoso provado*, peche I moravidí e jure que lo dixo con saña e con ira e que no lo sabe en él, e si dixiere que no lo dixo jure con II bezinos.	XX. Por *injuria de palabras* ningún vezino de Alcalá o de su término responda nin aya pena si non quexare el *injuriado*; e si el *injuriado* quexare, el alcalde aya información de su oficio e sin escripto nin pleito castigue al *injuriador* sin que aya pena de dineros nin lleven derechos por ello.

Asimismo, en el *FNA* el término *injuria* aparece acompañado de los derivados *injuriado* e *injuriador* para referirse a los participantes, voces todas que adquieren un sentido técnico en el contexto jurídico de la ley sancionadora del delito. Es esta una diferencia significativa, pues, como señalan Henríquez Salido/Alonso-Misol (2010: 59) en su *Historia del léxico jurídico*, en los primeros siglos son escasas las palabras derivadas y sufijadas, formaciones que empiezan a presentar cierta frecuencia a partir de los siglos XV y XVI.

Si bien en la historia del léxico es muy frecuente que palabras del ámbito especializado pasen al lenguaje común, en el caso de los textos de aplicación del Derecho se observa la dirección opuesta del cambio, con palabras del discurso ordinario que adquieren una función especializada en el contexto jurídico-legislativo y se convierten, así, en verdaderos tecnicismos. No faltan tampoco voces nacidas en el ámbito religioso que se extienden al derecho civil o penal. El siguiente cuadro recoge varios de estos casos:

FVA	*FNA*
14. Qui so *enemigo* matare	IV. Quien matare a su *enemigo que por tal fuere dado por sentencia*
68. Toda bona de mueble o de raíz que ganaren o compraren marido e mulier por medio lo partan.	XXXVII. Toda cosa qu'el marido e la muger ganaren en uno o compraren, mueble o raíz, áyanlo de por medio, *ansí en la propiedad como en la posesión.*
78. el padre o la madre que fuere vivo per su palabra seya creído de todo cuanto dieren	XXXVIII. E el padre o la madre por su *juramento* o si en su testamento lo ovieren *declarado con juramento* sean creídos de todo cuanto dieren si non oviere instrumento o testigos.
2. […] e d'estos C moravidís sean partidos per III tercias: la tercia part al señor, e la otra tercia part a *parientes del muerto* e la otra tercia part a los fiadores, e el omezillo sea del señor e esto esca primero.	I. […] los cuales sean partidos en tres partes: la tercia parte para el señor, e la otra tercia parte a *la parte querellante* e la otra tercia parte para el juez que lo condenare a pena de muerte. E si non oviere *condenación de muerte,* non aya omezillo.

El carácter plenamente jurídico de *enemigo* se hace explícito en *FNA* mediante la apostilla «que por tal fuere dado por sentencia». En el *FNA* se diferencia claramente entre los conceptos de *propiedad*, 'derecho o facultad de poseer alguien algo y disponer de ello dentro de los límites legales' (*DLE* 2014, s. v.) y *posesión*, 'situación de poder de hecho al que se le otorga protección jurídica'. Se prefiere el *juramento* a la expresión «ser creído *por su palabra*»; los «parientes del muerto»

se convierten en la *parte querellante*, locución nominal que, según los datos que nos proporciona el *CORDE*, parece utilizarse por primera vez un par de décadas antes de la redacción del *FNA*, en el *Ordenamiento de las Cortes de Toledo* de 1480. La palabra *condenación* se recoge por primera vez en aragonés, concretamente en la traducción de Fernández de Heredia del *Breviarium ab urbe condita*, de Eutropio, realizada en el último cuarto del siglo XIV, pero no se hace frecuente hasta el siglo XV, inicialmente con sentido religioso (condenación frente a salvación de las almas), y después aplicada también a las sentencias por distintos delitos.

La comparación de los dos fueros también nos proporciona, como no podía ser de otra manera, innumerables casos de sustituciones de un término que ha quedado obsoleto —o simplemente se ha perdido— por el correspondiente de la lengua de comienzos del siglo XVI. Curioso, por la incomprensión que revela, es el siguiente:

FVA	*FNA*
14. Qui so enemigo matare e después que fore morto *lo estemare o lo robare*, peche C e VIII moravidís si jo provaren e pártanse las caloñas como lo ál, a tercio, e non esca enemigo, pero si en feriendo lo estemare solo que non lieve miembro non peche caloña, e si non salve·s con XII bezinos.	IV. Quien matare a su enemigo que por tal fuere dado por sentencia e después de muerto *lo robare o tomare alguna cosa*, torne lo que tomó con el doblo e aya de pena de quinientos maravedís para el recurso, demás de las penas del derecho.

El verbo *robar* 'quitar con violencia', que no aparece más veces en los fueros de Alcalá que en la señalada, proviene (*DECH*, s. v.) del lat. vulg. **RAUBARE*, y este del germ. **RAUBÔN* 'saquear, arrebatar', y como tecnicismo jurídico mantiene hasta el presente su diferencia con *hurtar* 'quitar subrepticiamente': *FVA* 20 Qui *furtare* bezino a bezino → *FNA* V Todo omne que alguna cosa *hurtare*; *FVA* 290 A ladrón que tomaren con *furto* en casa → *FNA* CII Al ladrón que tomaren con el *hurto* en casa; etc.

Lo llamativo es que en el caso recogido en el cuadro anterior, el redactor de *FNA* IV parece estar usando *robar* con el sentido de *hurtar*: «lo robare o tomare alguna cosa» (al muerto), lo que, además, obliga a interpretar el pronombre *lo* como un caso de loísmo. No solo resultaría llamativo que tal indistinción se produjera precisamente en un texto jurídico, sino que adelantaría su documentación en más de un siglo: el *DECH* nos informa de que *Autoridades* ya admite la posibilidad de que *robar* valga para 'tomar para sí o hurtar', con ejemplos en el

XVII. Tampoco se puede descartar que se considerara necesariamente como violenta la acción de sustraer un bien a una persona, aunque esta estuviera muerta.

La posibilidad de cotejar el *FNA* con su fuente permite, por fortuna, entender debidamente las razones de este resultado en apariencia anómalo para un texto jurídico de comienzos del siglos XVI. En efecto, en *FVA* 14 «lo estemare o lo robare» tenemos el verbo *estemar*, que desaparece del *FNA*, y un «lo robare», con un pronombre de CD cuyo referente es el muerto. Esto se comprende bien si sabemos que *estemar* es una forma propia del siglo XIII que significa 'mutilar' < STIGMARE[9], por lo que lo que la ley sanciona es el robo del cadáver o su mutilación. Ya sea por el simple desconocimiento por parte del redactor del *FNA* del significado de *estemar*, ya porque la mutilación y el robo de algún miembro o del cuerpo entero no era una práctica que tuviera cabida en la sociedad del siglo XVI, o seguramente por ambos motivos, el redactor mantiene el pronombre *lo* del *FVA* y añade al lado de *robar*, mediante la misma coordinación disyuntiva que relacionaba *estemar* y *robar* en el *FVA*, la acción sinonímica que le parece más probable: *tomar alguna cosa*.

Sirvan estas breves notas léxicas para mostrar el interés de la comparación entre estos dos fueros, así como de las posibilidades de estudio que el *Altas Lingüístico Diacrónico e Interactivo de la Comunidad de Madrid* ofrecerá en un futuro próximo.

Referencias bibliográficas

ALDICAM-CM = Atlas lingüístico diacrónico e interactivo de la Comunidad de Madrid. <http://aldicam.blogspot.com.es> [último acceso: 15/11/2017].

Alonso Romero, María Paz (1982): *El proceso penal en Castilla, siglos XIII–XVIII*. Salamanca: Universidad.

Bustos Gisbert, Eugenio/Ramón Santiago (2002): «Para un nuevo planteamiento de la llamada 'norma madrileña' (siglos XVI y XVII)», en María Teresa Echenique Elizondo *et al.* (eds.), *Actas del V congreso internacional de historia de la lengua española*. Madrid: Gredos, 1123–1136.

Cestero Mancera, Ana María/Isabel Molina Martos/Florentino Paredes García (eds.) (2015): *Patrones sociolingüísticos de Madrid*. Berna: Peter Lang.

9 Geográficamente es de adscripción oriental, pues se halla en el Vidal Mayor, en los Fueros de Navarra, Fueros de Aragón, además de en los de Sepúlveda, Soria y Brihuega, todos relacionados con el de Alcalá, en Berceo y en el Fuero Viejo de Castilla, que trae también el sustantivo *estemamiento*, al igual que en las Siete Partidas (manuscrito de finales del XV).

CHARTA = *Corpus Hispánico y Americano en la Red: Textos Antiguos*. <http://www.corpuscharta.es> [último acceso: 15/11/2017].

CORDE = Real Academia Española: Banco de datos (*CORDE*) [en línea]. *Corpus diacrónico del español*. <http://www.rae.es> [último acceso: 10/07/2017].

DECH = Corominas, Joan/José Antonio Pascual (1980–1991): *Diccionario crítico etimológico castellano e hispánico*, 6 vols. Madrid: Gredos.

DLE = Real Academia Española/Asociación de Academias de la Lengua Española (2014[23]): *Diccionario de la Lengua Española*. Madrid: Espasa. Actualización 2017: <http://dle.rae.es/> [último acceso: 15/12/2017].

Fernández-Ordóñez, Inés (dir.) (2005–): *Corpus Oral y Sonoro del Español Rural (COSER)*. <http://www.corpusrural.es> [último acceso: 15/12/2017].

García Mouton, Pilar/Isabel Molina Martos (2009): «Trabajos sociodialectales en la Comunidad de Madrid», *Revista de Filología Española* LXXXIX, 1, 175–186.

García Mouton, Pilar/Isabel Molina Martos (2015): *Atlas Dialectal de Madrid (ADiM)*. Madrid: CSIC. <adim.cchs.csic.es> [último acceso: 15/12/2017].

García Mouton, Pilar/Isabel Molina Martos (2017): *Las hablas rurales de Madrid. Etnotextos*. Berna: Peter Lang.

Henríquez Salido, Maria do Carmo/Enrique de No Alonso-Misol (2010): *Historia del léxico jurídico*. Madrid: Editorial Aranzadi.

Kabatek, Johannes (2001): «¿Cómo investigar las tradiciones discursivas medievales? El ejemplo de los textos jurídicos castellanos», en Daniel Jacob y Johannes Kabatek (coords.), *Lengua medieval y tradiciones discursivas en la Península Ibérica*. Fráncfort/Madrid: Vervuert/Iberoamericana, 97–132.

Kabatek, Johannes (2004): «Tradiciones discursivas jurídicas y elaboración lingüística en la España Medieval», *Cahiers d'Études Hispaniques Médiévales* 27, 249–261.

Pérez Bustamante, Rogelio (1986): «Pervivencia y reforma de los derechos locales en la Época Moderna. Un supuesto singular: el Fuero Nuevo de Alcalá de Henares de 1509», *En la España Medieval* 5, 743–760.

Rodríguez Molina, Javier/Álvaro Octavio de Toledo y Huerta (2017): «La imprescindible distinción entre texto y testimonio: el *CORDE* y los criterios de fiabilidad lingüística», *Scriptum Digital* 6, 5–68. <http://scriptumdigital.org/numeros.php?num=23&lang=es> [último acceso: 15/11/2017].

Sánchez González de Herrero, Nieves/Juan Sánchez Méndez/Ingmar Söhrman/María Jesús Torrens Álvarez (2013): «La Red Charta: objetivos y método», en Emili Casanova Herrero y Cesareo Calvo Rigual (eds.), *Actes*

del 26é Congrés Internacional de Lingüística i Filologia Romàniques, vol. VII. Berlín: W. de Gruyter, 263–274.

Sánchez Moltó, Vicente (2011): «Del Fuero Viejo al Fuero Nuevo de Alcalá: estudio comparativo», en José Luis Valle (ed.), *Fuero Nuevo de Alcalá. Estudios y edición*. Alcalá de Henares: Institución de Estudios Complutenses, 77–121.

Sánchez-Prieto Borja, Pedro (2017): «Más allá de hablar castizo: la lengua de Madrid en los documentos», *TextoRblog*. <https://textorblog.wordpress.com/2017/10/13/mas-alla-de-hablar-castizo-la-lengua-de-madrid-en-los-documentos/> [último acceso: 15/11/2017].

Sánchez-Prieto Borja, Pedro/Ana Flores Ramírez (2005): *Textos para la Historia del español, IV. Archivo Regional de la Comunidad de Madrid*. Alcalá de Henares: Universidad de Alcalá.

Sánchez-Prieto, Pedro/Delfina Vázquez Balonga (2018): «Toledo frente a Madrid en la conformación del español moderno: el sistema pronominal átono», *Revista de Filología Española* XCVIII, 1, 185–215.

Torrens Álvarez, María Jesús (2002): *Edición y estudio lingüístico del* Fuero *de Alcalá (Fuero Viejo)*. Alcalá de Henares: Fundación Colegio del Rey.

Torrens Álvarez, María Jesús (2011): «La transmisión textual de los fueros de Alcalá», en José Luis Valle (ed.), *Fuero Nuevo de Alcalá. Estudios y edición*. Alcalá de Henares: Institución de Estudios Complutenses, 41–76.

Torrens Álvarez, María Jesús (2012): «Un caso singular de reelaboración lingüística y discursiva: los fueros de Alcalá de Henares», en Emilio Montero Cartelle (ed.), *Actas del VIII Congreso Internacional de Historia de la Lengua Española*, vol. II. S.l.: Meubook, 2441–2452.

Torrens Álvarez, María Jesús/Vicente Sánchez Moltó (2011): «Fuero Nuevo de Alcalá de Henares. Edición crítica», en José Luis Valle (ed.), *Fuero Nuevo de Alcalá. Estudios y edición*. Alcalá de Henares: Institución de Estudios Complutenses, 123–176.

José R. Morala y Mª Cristina Egido

El proyecto *CorLexIn* y la variación diatópica en el léxico del Siglo de Oro[1]

Resumen: El *Corpus Léxico de Inventarios* (*CorLexIn*) es un corpus textual formado por documentación notarial procedente de archivos de todo el ámbito hispánico, realizado con el objetivo principal de estudiar el léxico de la vida cotidiana en el Siglo de Oro, con la posibilidad añadida de tratarlo desde una perspectiva diatópica. En este trabajo utilizamos los datos del corpus como fuente para analizar diacrónicamente algunos localismos (*sesos*, *toña*, *mancaje*) y voces con una distribución geográfica interesante (*trébedes*, *cetra*) para, finalmente, estudiar ciertos nelogismos que nuestro corpus da a entender que aún no se han asentado en el registro léxico más general, como *bagatela* o *valija*.

Palabras clave: Léxico, Diacronía, Variación diatópica, Siglo de Oro

Abstract: The Lexical Corpus of Inventories (*CorLexIn*) contains notarial records from all over Spain; its primary goal was the study of everyday-life vocabulary, taking also into account diatopic variation. This survey uses *CorLexIn* data to study the diachronic changes in a few localisms (*sesos*, *toña*, *mancaje*) and in a number of words with interesting geographical distribution (*trébedes*, *cetra*); besides, certain neologisms which, according to the information in the corpus, were not yet settled in the general register (*bagatela*, *valija*) are also analysed.

Keywords: Lexicon, Diachrony, Diatopic Variation, Golden Age

1 El corpus *CorLexIn*

El *Corpus Léxico de Inventarios* (*CorLexIn*) comenzó a definirse a partir de un proyecto de investigación coordinado (*Diccionario del español de los siglos de oro. Sus bases filológicas y lingüísticas*) realizado en el trienio 2006–2008, en el que el equipo de la Universidad de León se encargó de estudiar textos diatópicamente marcados procedentes de esa época. Con posterioridad, el proyecto ha funcionado de manera autónoma –centrándose en un tipo concreto de documentación

1 Para la realización de este trabajo se ha contado con la financiación del Ministerio de Economía y Competitividad al proyecto con número de referencia FFI2015-63491-P (MINECO/FEDER).

notarial del siglo XVII– y al equipo original se han incorporado investigadores de la Universidad de Burgos y de la de Oviedo.

1.1 Objetivo del corpus

El objetivo de nuestro trabajo a lo largo de estos años ha sido el de crear una base de datos textual que, con criterios muy específicos, nos permitiera analizar el léxico de la vida cotidiana y hacerlo, además, desde el punto de vista de la variación diatópica. Para ello nos servimos de textos notariales, manuscritos e inéditos, procedentes de toda la geografía hispanohablante. Los documentos elegidos están datados mayoritariamente en el siglo XVII y, en menor medida, en el siglo XVI o primeras décadas del siglo XVIII.

Por lo que respecta al contenido de los textos, dado que nuestro interés era expresamente el estudio del léxico, se ha restringido a aquellos documentos que presentan una mayor riqueza léxica. En concreto, el corpus está compuesto por lo que genéricamente puede denominarse como *relaciones de bienes*, es decir, en su inmensa mayoría, inventarios y tasaciones de bienes, partijas y repartos de herencias, cartas de dote y arras, almonedas, testamentos, etc.

En la actualidad, el *CorLexIn* abarca documentos de todos los archivos históricos provinciales españoles[2] y, en cuanto a América, disponemos de documentos de México, Bolivia, Colombia, Chile, Panamá, Guatemala, Venezuela, El Salvador, Perú y Puerto Rico. Fruto de la colaboración con el equipo que está redactando el *Nuevo Diccionario Histórico*, los documentos ya editados están accesibles para su consulta pública en la web de la Fundación Rafael Lapesa[3]. A finales del año 2016, el recuento de datos disponibles en la web es de 494 documentos transcritos a partir de 3868 imágenes que, a su vez, contienen 1 187 888 palabras[4], unos materiales que esperamos seguir incrementando, pues no consideramos aún cerrado el corpus.

2 Aún no se han incorporado textos transcritos procedentes del archivo de Gran Canaria. Quedan al margen, lógicamente, los archivos procedentes de Galicia y el área catalanohablante, que no se incluyen.

3 La consulta del corpus en línea puede hacerse en la web del *NDHE* <www.frl.es>. La información relativa al proyecto, con indicación de los trabajos desarrollados, archivos visitados, legajos vaciados, un índice de voces estudiadas y la versión en línea de los trabajos publicados, puede verse en <http://corlexin.unileon.es/>. Como el corpus no se considera cerrado, el número de ejemplos que utilizamos en el trabajo es más amplio que el que puede encontrarse en la versión en línea.

4 Este número se refiere a los textos publicados en el corpus. Como puede suponerse, los trabajos que hemos ido publicando los miembros del equipo se sirven habitualmente de un número más amplio de ejemplos del que, para un caso concreto, puede obtenerse

1.2 Materiales y resultados

El tipo de documentos que seleccionamos presenta una escasa complejidad sintáctica pues —salvo en el protocolo y en el escatocolo, a su vez plagados de formulismos notariales— lo habitual es que, en el cuerpo del documento, nos encontremos con meras enumeraciones o descripciones de objetos a los que, en el mejor de los casos, se añade su estado de uso, su aspecto o su valoración económica. A cambio, estas relaciones de bienes atesoran una gran riqueza léxica, pues el texto viene condicionado por la obligación del escribano de turno de registrar e identificar —con la mayor fidelidad y detalle posibles— todos los bienes de una persona o institución a partir de la declaración de los dos o más individuos que hacen la tasación.

No obstante, el corpus aporta también información interesante en ámbitos ajenos al del léxico, especialmente cuando contrastamos un corpus de origen notarial, como es este, con los datos extraíbles de corpus generales como el *Corpus Diacrónico del Español* (*CORDE*) o el *Corpus del Diccionario Histórico* (*CDH*) para esa misma época. La diferencia de registro que uno y otros presentan nos ha permitido, por ejemplo, localizar y situar geográficamente el uso del neutro de materia en el Siglo de Oro (Morala 2015b), fijar la existencia de algunas construcciones sintácticas de interés (Pérez Toral 2014 y 2015; Egido 2016) o comprobar la escasa implantación de algunas novedades morfológicas que son habituales en textos coetáneos del registro literario, como ocurre con el plural *quienes* (Morala 2016: 383–387) o con el superlativo sintético en *–ísimo* (Morala 2014: 24–25). Del mismo modo, no es difícil encontrar datos de interés sobre aspectos gráficos o fonéticos (Morala y Egido 2010; Morala 2016) o disponer de una gran cantidad de información novedosa sobre el proceso de formación de palabras, con creaciones léxicas no siempre documentadas por otra vía (Morala 2012, 2015a; Perdiguero 2012).

Bien es verdad que la mayor utilidad del *CorLexIn* —e insistimos que con ese fin se ha diseñado el corpus— se da en el campo del léxico: de un lado, se trata de textos con una gran diversidad léxica; de otro, al tratarse de textos notariales, están estrictamente fechados y localizados. Por último, por su propia naturaleza, registran una buena muestra del léxico cotidiano en distintos niveles sociales y registros lingüísticos. Si unimos todos estos criterios, podríamos decir que, adecuadamente analizados, los materiales del *CorLexIn*

---

en la web. Estos ejemplos son el resultado de la lectura de documentos de los archivos visitados que aún no han pasado el filtro de la transcripción y edición en el corpus publicado.

nos proporcionan unos datos equiparables a los de los modernos atlas lingüísticos, con la diferencia de que esa imagen la obtenemos para el siglo XVII, un momento en el que, en sentido amplio, se están escribiendo las grandes obras literarias de la época áurea, cuyos autores van a ser tiempo después el modelo sobre el que se configura el *Diccionario de Autoridades*. En este sentido, nuestro corpus, con las diferencias de registro lingüístico que implica, sirve de fuente alternativa a la hora de describir la lengua del Siglo de Oro y especialmente el léxico que utiliza.

Como muestra del interés que para el léxico tiene este tipo de textos, analizamos a continuación varios ejemplos representativos de lo que puede aportar un corpus de las características del *CorLexIn*. Lo ejemplificaremos con alguno de los fenómenos que con mayor frecuencia encontramos en el campo del léxico, como ocurre con los vocablos de un ámbito geográfico restringido —unas veces invisibles para el repertorio académico y otras tratados como si fueran de uso general— o con las voces que presentan una variación formal o un uso diatópicamente marcado. Finalmente, veremos que un corpus realizado con estos criterios no solo es útil por la información que contiene, sino también por la que no es posible registrar en él.

2 Documentación de localismos léxicos

En un corpus con tanta variedad léxica, no es difícil localizar ejemplos de palabras que pertenecen al ámbito local y que los diccionarios no han sido capaces de registrar. Los ejemplos que hemos tratado son múltiples, pero ahora nos centraremos en *toña* y en *mancaje*. Del mismo modo, voces que el repertorio académico recoge sin marca alguna —como ocurre con *seso*—, se muestran en nuestro corpus con todas las características de lo que debería considerarse únicamente como un localismo léxico.

2.1 *Seso*

El *Diccionario de la Lengua Española* (*DLE*) define la segunda entrada de *seso* como 'piedra, ladrillo o hierro con que se calza la olla para que asiente bien', manteniendo la definición que ya registró en *Autoridades* (1739). Con anterioridad, el término había aparecido en algunos diccionarios bilingües como el de Oudin (1607) 'vn accotepot' o el de Vittori (1609) 'vn accotepot, apoggio di pentola' (*Nuevo Tesoro de la lengua Española*, *NTLLE*, s. v. *seso*). En ningún caso figura con marca diatópica en los repertorios lexicográficos, dando a entender, por tanto, que se trata de una voz de uso general.

En el *Diccionario Crítico Etimológico Castellano e Hispánico* (*DECH*, s. v. *sentar*) se considera que es una variante mozárabe[5] del resultado del latín SĔSSU 'asiento' y se adjunta una cita de Azorín[6] referida a la localidad de Maqueda, en Toledo. Con este sentido, *seso* aparece en los siguientes casos en el *CorLexIn*[7]:

> en la cocina se alló lo siguiente: unos moricos, dos *sesos*, unas trévedes y un badil (Cortes, Na-1645)[8]
> un par de morillos de yerro, torneados, buenos; dos *sesos* y dos pares de tenazas de yerro; y un badil y badilexa y unas tréuedes (Soria, 1646)
> tres *sesos* para las ollas; vnas tenazas, vna boluedera, vn badil y un tostador; vn morillo de cozina (Lumbreras, LR-1685)
> tres coberteras, tenazas dos, tres asadores, su gato, vn *sesso*, todo de cozina (Lumbreras, LR-1688)

El panorama que presentan los inventarios del siglo XVII apunta a que, más que ante un término general, estamos ante una voz regional. Teniendo en cuenta que el vocablo se registra junto a otros objetos usados en el fuego del hogar, sería impensable que, si fuera de uso general, no apareciera en otras zonas, como ocurre con las *tenazas* o los *morillos*, registrados una y otra vez entre el ajuar doméstico de inventarios procedentes de toda la Península.

Sin embargo, el *seso* 'asiento' solo figura en textos procedentes del área formada por Navarra, La Rioja y Soria. Es posible que en otras zonas haya decaído su uso, pero la imagen que nos dan los inventarios para el siglo XVII es que, en ese momento, *seso* no se empleaba ya de modo general y solo mantendría su vigencia en el área indicada, lo que debería ayudarnos a explicar el uso histórico de este término en castellano.

---

5 La procedencia no estrictamente castellana permitiría justificar la falta de diptongación de /ĕ/, proceso que sí se cumple en el resultado patrimonial castellano SĔSSU > *sieso* 'ano'.

6 Teniendo en cuenta el buen uso del diccionario académico del que hace gala Azorín, no sería extraño que tomara la voz directamente del *DLE*, aunque en la localidad a la que se refiere fuera desconocida.

7 Aparece un ejemplo más pero, por su contexto, no parece relacionado con el entorno del hogar: «vna cama del sesso de este valle, entera, ya bieja, de pluma» (Guriezo, S-1669). La mención a que está hecha «de pluma» ni siquiera permite interpretar *cama* como 'pieza curva', valor con el que sí podría encajar en el sentido con el que aquí se analiza *seso*.

8 Los ejemplos del corpus se acompañan del nombre de la localidad en la que se ubica el documento, seguido del indicativo provincial y el año en el que se data el texto.

2.2 *Mancaje*

En este caso, la voz no aparece ni en el *DLE* ni en el *DECH* y únicamente figura en una de las papeletas del *Fichero General* de la RAE, en la que se toma la referencia del diccionario de Alcalá Venceslada 'especie de almocafre de cabo largo para desembarazar las regueras', quien lo cita para la comarca de Las Alpujarras. En el *NTLLE* (s. v. *mancaje*), el único diccionario que lo incluye es el de Zerolo (1895) con el sentido de 'escardillo'. En cuanto a los datos históricos, el *CORDE* lo registra en una sola ocasión, en este caso con la variante *mancax* («un mancax de hierro»), en un inventario de bienes pertenecientes a moriscos del Reino de Granada datado en 1567.

El término está, sin embargo, bien representado en el *Atlas Lingüístico y Etnográfico de Andalucía* (*ALEA*, mapa 33), donde las respuestas para 'escardillo, herramienta para escardar' registran *mancaje* en los puntos del área central y meridional de la provincia de Granada, así como en zonas aledañas de la de Almería. Del mismo modo, diversos vocabularios locales reunidos en el *Tesoro Léxico de las hablas andaluzas* (*TLHA*) recogen en esta área tanto el sustantivo *mancaje* como el verbo *mancajar* (*TLHA*, s. v. *mancajar* y *mancaje*). Fernández Sevilla (1975: 317–318) la califica de voz típicamente granadina y almeriense y apunta la posibilidad de que su origen tenga que ver con *mano*.

En el *CorLexIn* figura *mancaje* en varias ocasiones al lado de otros aperos agrícolas en un legajo correspondiente a la notaría de Albuñol, con inventarios fechados en la localidad de Narila, en la Alpujarra granadina[9]:

> vn açadón nuebo, dos *mancages* y vna hacha (Narila, Gr-1697)
> vn hoçino y un cañón de un *mancaje* (Narila, Gr-1697)
> tres arados aperados, dos açadones, un *mancaje* (Narila, Gr-1697)
> dos açadones y una hacha; dos *mancajes* y una sartén mediana; y una orca mediana (Narila, Gr-1697)
> un *mancaje* en tres reales; un azadón en beinte reales (Narila, Gr-1699)

Seguramente, si se hiciera una lectura más exhaustiva en vez del muestreo documental que en aras de la operatividad utilizamos para el *CorLexIn*, se encontraría también en otros documentos de la zona. En cualquier caso, son ejemplos que nos permiten constatar el uso histórico de esta voz de ámbito local de la que no abundan los testimonios antiguos.

---

9 Fernández Sevilla, que recoge las referencias aportadas por Alcalá Venceslada, le corrige tanto en la extensión del uso de la palabra –no solo es una voz alpujarreña– como en la definición que le adjudica: el *mancaje* sirve para escardar, más que para limpiar las regueras (1975: 317).

2.3 *Tuña/toña*

En el *DECH*, se indica que *tonel* es voz tomada del francés antiguo *tonel*, diminutivo de *tonne* 'tonel grande', y este del latín tardío TŬNNA, a su vez tomado del céltico TŬNNA 'piel', de donde 'odre' y después 'cuba' (*DECH*, s. v. *tonel*). El seguimiento de los resultados de TŬNNA se revela complejo en el caso del iberorromance: J. Corominas y J. A. Pascual entienden que no hay resultados patrimoniales del TŬNNA del que deriva *tonel*, si bien aportan otros resultados con significados secundarios, además de derivados verbales como *entoñar* 'enterrar' y *atoñar* 'atollar', ejemplos todos ellos localizados en la franja occidental (*DECH*, s. v. *tonel* y *toña*).

Además de registrarse en gallego[10], los repertorios léxicos referidos al dominio asturleonés registran algunos datos de interés al respecto: junto a derivados verbales como *entoñar*, *atoñar* (Le Men 2005: 334), el sustantivo *tuña* se localiza en puntos aislados del occidente de Asturias, del noroeste de León y del occidente de Zamora. En todos ellos, el tenor general de la definición es el de 'arcón o depósito de madera para guardar el grano' (Le Men 2012: 729).

Los registros que localizamos en nuestro corpus, referidos en este caso a los archivos de León y Zamora, atestiguan el uso histórico de *tuña* en el occidente de Zamora y el de la variante *toña* en una amplia zona de León, que contrasta con el único caso en la montaña noroccidental en la que figura *tuña* en los repertorios dialectales modernos[11]:

> en vna *toña* que está en la dicha casa de la dicha difunta aber diez cargas y media de trigo (León, 1643)
> otra quadra biexa de la *toña* de la cevada que se a de medir (Villacelama, Le-1638)
> una *toña* grande y dos escriñas baçías (Bonillos, Le-1680)
> una *toña* que tendrá asta otras çinco cargas de arina de zenteno (Santa Marina de Somoza, Le-1680)

10 La base de datos *Recursos Integrados da Lingua Galega* (*RILG,* s. v. *tuña*) localiza *tuña* —en ocasiones con la variante *tulla*— en diversos diccionarios de gallego con las acepciones de 'troj', 'sitio donde se guarda la cosecha de pan', 'parte del hórreo' o 'depósito de madera donde se almacena el grano'. No se registran, sin embargo, ejemplos en textos históricos en los corpus que incluye esta herramienta.

11 A estos datos, cabe añadir que, en información que hemos podido constatar en Val de San Lorenzo (León), aún se recuerda vagamente *tuña* con el sentido genérico de 'panera, local en la parte alta de la casa', en lo que parece un desplazamiento semántico del recipiente y el uso que se le da al lugar en el que se encuentra y que tiene la misma función de guardar el grano.

quatro *toñas* de paxa y en vna se alló cosa de media carga de arina de centeno (Astorga, Le-1648)
más dos *toñas* que llaman *furonas*, vna mediana y otra grande […] *tres toñas* grandes de paxas y en ellas se alló dentro de la vna cosa de vn quartal de salbados más en la otra, cosa de vna carga de arina de trigo […] en vna *toña* de tabla asta siete v ocho cargas de arina de centeno más, dentro de las *paneras*, seis cargas de çenteno (Brimeda, Le-1648)
más se alló en vna *toña* de la panerica dos cargas de trigo; más se alló en otras dos toñas de la d*i*cha panera y çillero diez cargas de centeno; más se alló en vn escriño biexo vn quartal de linaça (Brimeda, Le-1648)
dos *tuñas* (Alcañices, Za-1669)
quatro cargas y media de zenteno en una *tuña* (Villarino de Manzanas, Za-1683)
un arca; una *tuña* en seis ducados; tres zestos (Villarino de Manzanas, Za-1683)
una *tuña* en sesenta reales (Villarino de Manzanas, Za-1683)

El sentido de estos textos parece claro que se refiere a un recipiente en el que se guarda bien el cereal, bien el grano ya molido. El hecho de que pueda aparecer junto a arcas, cestos o escriños, redunda en esta misma idea, al tiempo que, por las indicaciones que ofrecen los inventarios, sabemos que, además de madera, podía ser también de paja, como el escriño. Este sería a su vez el significado básico y antiguo del término, coincidente con el étimo latino-celta, del que estos ejemplos representarían el resultado patrimonial.

Los datos históricos que aportamos, muestran que *tuña* o *toña* es voz antigua en el dominio asturleonés[12] y que los derivados verbales citados procederían de esta forma simple[13]. Por otra parte, indican que, al menos en León, se trata de una voz en retroceso pues en la documentación alcanza a zonas del área centro-oriental de la provincia, incluida la ciudad de León, en la que hoy no se conoce. Finalmente, cabe indicar que, para un étimo como TŬNNA, con /ŭ/ tónica, la solución esperable sería *toña*, de la que aportamos también registro histórico. La solución en /u/, en vez de la esperable en /o/, probablemente haya que ponerla en relación con otras palabras que presentan esta misma anomalía evolutiva y que son especialmente frecuentes en el ámbito occidental (Lloyd 1993: 51–52 y 296–300). En cualquier caso, la documentación localizada en

12 Aunque excede de los límites temporales del *CorLexIn*, como testimonio antiguo en el área leonesa puede aducirse una cuba «ataunnada» que se registra en un texto del monasterio de Sahagún del año 1221 (doc. nº 1623) cuya fecha ha de tomarse con alguna precaución pues procede de una copia muy tardía.

13 En el *DECH* (s. v. *toña*), ante la falta de datos concluyentes del término simple, se plantea la posibilidad de que *toña* sea un postverbal a partir de *entoñar* o *atoñar*.

León y Zamora contribuye a configurar la historia de un término del que no contábamos con muchos datos.

3 La variación diatópica en el léxico

Como se indicó arriba, hay un cierto paralelismo entre los datos que proporciona el *CorLexIn* —con documentos necesariamente ubicados en una localidad concreta— y los que obtenemos a partir de los atlas lingüísticos. En este caso, veremos la utilidad del corpus para establecer isoglosas léxicas con los términos *trébedes* y *acetre*.

3.1 *Trébedes*

Trébedes, del latín TRĬPĔDES, es voz de uso general. Como se trata de un objeto habitual en la mayoría de los hogares, se registra en múltiples inventarios, lo que nos permite analizar las variaciones que presenta a lo largo de toda la zona estudiada. Lo interesante en este caso es la variación formal que presenta el término como consecuencia de las distintas opciones elegidas para solucionar el grupo romance [b̓d] que se originaría —cuando esto ocurre— por la pérdida de la vocal postónica, una vez que se ha sonorizado la /-p-/ intervocálica[14].

La solución más general a esta secuencia anómala [b̓d] implica el mantenimiento de la vocal postónica, con lo que no llegaría a formarse el grupo consonántico, dando lugar al resultado normativo *trébedes*. No obstante, otra posibilidad es la pérdida de dicho elemento, con un resultado inicial *trebdes*[15] que se va a resolver de formas distintas. Otro elemento que añade una nueva variación formal es la frecuencia con la que la voz toma un prefijo *es-*, dando como resultado variantes del tipo de *estrébedes*[16].

Para analizar los resultados que el término presenta en el *CorLexIn*, dejaremos fuera las formas normativas (*trébedes* y ocasionalmente *trébede*) presentes

14 El grupo lo encontramos igualmente en *capitale* > *cabdal* > *caudal* y —ya con /b/ en el étimo— en *debita* > *debda* > *deuda*, *civitate* > *cibdad* > *ciudad*.

15 Esta forma, con las variantes *trebdes* o *trevdes*, se registra ocasionalmente en textos medievales del *CORDE*, pero no figura ya en nuestros documentos del Siglo de Oro.

16 Pese a que se trata de una forma relativamente conocida, esta forma prefijada no es habitual en los documentos de nuestro corpus. En el *CorLexIn* figura *estreudes*, pero no *estrébedes*. La voz, sin embargo, está documentada ocasionalmente en inventarios de la época: en documento de Yecla (Murcia) de 1568 figura «una estrévades» (Puche Lorenzo 2002: 141); Bastardín (2015: 51), que la considera una voz de uso meridional, cita un *estrébedes* en documento de Medina Sidonia (Cádiz) de 1799.

en prácticamente todos los archivos estudiados, para centrarnos en aquellas que presentan algún tipo de variación formal en su evolución.

Al margen de *trébedes*, el resultado más frecuente es el que, a partir de *trebde*, soluciona el grupo consonántico con la vocalización de la /b/ en /u/, llegando a la solución *treudes*[17]. Esta forma es la que encontramos en la documentación del área navarro-aragonesa[18], en la que aparecen tanto *treudes* como las variantes prefijadas *estreudes* y *estreudas*:

vnas *trehudes* (Zaragoza, 1603)
vnas *treudes*; vn asnico para asar en el fuego (Sos del Rey Católico, Z-1684)
un caldero y unas *treudes* (Teruel, 1625)
cinco asadores, dos *estreudes*, quatro sartenes (Loscertales, Hu-1653)
dos pares de *estreudas*; quatro assadores (Quicena, Hu-1656)
tres cucharas de yerro, vnas *treudes* y tres asadores (Tudela, Na-1641)

Como continuación hacia el Sur del resultado aragonés, aparece igualmente *treudes* en el área suroriental de la Península, tanto en la zona castellano-hablante de Villena (Alicante) como en el extremo oriental de Albacete (Almansa), donde se usa regularmente este resultado que, sin embargo, no figura en el resto de Albacete o en Murcia, áreas en las que se usa la forma normativa:

dos *treudes* en dos reales (Sax, A-1639)
vnas *treudes* en quatro reales (Sax, A-1639)
vnas graíllas, vnas tenaças, vnas *treudes* (Sax, A-1685)
unas *treudes* en siete reales (Almansa, Ab-1639)
vnas *treudes* y tenaças (Almansa, Ab-1640)
más *treudes* y tenaças y dos asadores y una sartén (Almansa, Ab-1640)
unas *treudes* y unas tenaças en quatro reales (Almansa, Ab-1641)

Hay un segundo foco, sin continuidad geográfica con el anterior, en el que se localiza un número representativo de ejemplos de *treudes* —a veces con la

17 El resultado aparece registrado en varios diccionarios bilingües de los siglos XVI–XVII, mientras que en el repertorio académico solo lo hace en la edición de 1803, que lo califica de antiguo y remite, sin más, a *trébedes* (*NTLLE*, s. v. *treudes*). En el *CORDE*, *treudes* aparece en una docena de textos, principalmente de los siglos XV al XVII, entre los que se encuentra el uso de este ejemplo en Nebrija y en Correas como referencia gramatical a un grupo de sustantivos que carecen de la forma de singular. Quizá como un cruce entre *treudes* y *trébedes*, aparece en nuestro corpus algún caso aislado de *truébedes* que tal vez solo represente un error de grafía: «vnas truébedes quemadas» (Alaejos, Va-1630); «unos lares y unas truébedes» (Pamplona, 1640).

18 Ortiz Cruz (2015: 53) registra igualmente *estreudes* en inventarios de bienes del siglo XVIII y considera la voz como aragonesa.

variante gráfica *treodes*—, aunque convivan con el resultado normativo *trébedes*. En este caso, aparecen en el área suroriental de la Península, con ejemplos en documentos de los archivos de Badajoz, Huelva, Cádiz y Sevilla y, como ocurre con frecuencia, este mismo resultado suroccidental alcanza también a Canarias:

unas *treudes*, muelles y asadero (Mérida, Ba-1642)
dos assadores y un badil y unas muelles y unas *tréudes* (Mérida, Ba-1657)
unas *treodes* (Torre de Miguel Sesmero, Ba-1658)
dos *treodes*, un morillo, çinco asadores, unas muelles y un vadil (Segura de León, Ba-1659)
vnas parrilas y unas *treodes* (Bollullos Par del Condado, H-1650)
vnas parrillas y vnas *treodes* (Bollullos Par del Condado, H-1657)
vnas *treudes* (Medina Sidonia, Ca-1603)
vna sartén, y vnas *treudes*, y dos asadores (Medina Sidonia, Ca-1603)
vnas *treudes* y vn asador (Medina Sidonia, Ca-1603)
dos asadores, y una paleta y unas *treudes* (Lebrija, Se-1641)
unas *treodes*, en tres reales y medio (Gerena, Se-1658)
doz *truedes* y dos exparrillas (Garachico, TF-1695)

Bien como extensión hacia el oeste de la forma navarro-aragonesa *treudes*, bien como solución propia, aparecen varios ejemplos en la Rioja Baja en la que el antiguo *trebdes* se ha reducido por la pérdida de /b/[19], con resultado final *tredes*, que encontramos en documentos de Alfaro[20]:

vnas *tredes* de la lumbre y otras más pequeñas […] vn trasfuego de yerro y dos moricos y una orquilla y un badil y unas *tredes*, todo de yerro (Alfaro, LR-1646)
y unas *tredes* pequeñas y un caballuelo […] dos coçinos y dos calderos de colar y unas *tredes* y dos tinajas de agua (Alfaro, LR-1646)

Frente a estos resultados, en la zona occidental peninsular, de influencia leonesa, el grupo /-b̓d-/ puede solucionarse con un cambio de /l/ por /b/ para facilitar la pronunciación del grupo consonántico[21]. En ese caso, la solución de *trebdes*

19 La solución con pérdida de /b/ no es extraña en castellano: el *DECH* (s. v. *ciudad*) recuerda que entre los resultados del latín CĪVĬTATE ha de contarse el antiguo *cidat*, presente, por ejemplo, en el *Alexandre*.

20 Alejado de esta zona, en documentos de Ciudad Rodrigo (Salamanca) se localiza un ejemplo de *tredes*, con la particularidad de que aquí figura en masculino (Perdiguero 2016: 140), variante que quizá haya que relacionar con las formas habituales en Cáceres.

21 Borrego (1999: 145), si bien incluye este rasgo entre los característicos del leonés (*trelde, caldal, julgar*), considera que se trata de un elemento presente sobre todo en los textos antiguos pero escasamente representado en la actualidad. De hecho el *trelde*, habitual en la zona de Toro (Zamora) en el siglo XVII, no se reconoce en la actualidad en esa misma zona.

sería *treldes*. Esta variante (*treldes*, *estreldes*), que no aparece en los documentos analizados de Asturias o de León[22], figura sin embargo en alguno de Salamanca y, sobre todo, en documentos del área oriental de Zamora, curiosamente la menos propicia para presentar formas de raíz leonesa:

vnas *treldes* de hierro de tres esquinas (La Alberca, Sa-1669)
vna caldera con sus *treldes* (Toro, Za-1607)
vna caldera grande de bodega y unas *estreldes* (Toro, Za-1607)
vna caldera con sus *estreldes* (Toro, Za-1665)
adjudicósele unas *estreldes* tasadas en doze reales (Morales de Toro, Za-1675)
unas *estreldes* tasadas en doce reales (Morales de Toro, Za-1678)
una caldera y *treldes* (Fuentesaúco, Za-1671)
unas *treldes* y dos morillos de yerro (Fuentesaúco, Za-1671)

Finalmente, alternando con la forma normativa *trébedes*, aparecen en la documentación procedente de Cáceres algunas variantes diferenciadas de todas las anteriores, con el añadido de un apreciable grado de variabilidad interna (*treuces*, *treoces*, *trences*, *treunces*):

unas tenazas y badil; unas *treoces* (Cáceres, 1657)
vnas *treuces* […] vnas *treuços*, vn morillo (Logrosán, Cc-1675)
unas *trevços* pequeñas […] unas *trevços* grandes […] unas *treuços* […] unas *treuços* (Logrosán, Cc-1675)
unas *treunces* con un pie menos (Logrosán, Cc-1687)
unas *trences*, en cinco reales (Jaraíz de la Vera, Cc-1663)
un asador; unas *trençes*; un sobellón (Cañamero, Cc-1668)

El hecho de que se citen estos objetos en contextos similares a los que hemos analizado arriba conduce a interpretarlos en el mismo sentido que *trébedes*. En los repertorios dialectales del área se recogen voces semejantes —siempre con el prefijo *es-* que hemos visto en *estrébedes*— como *estreoces*, *estreuce*, *estreocis* (Viudas 1980: 71).

Como puede verse, el hecho de que estemos ante un objeto citado con mucha frecuencia en inventarios y tasaciones, añadido al dato de que presente una apreciable variabilidad formal, nos permite establecer con bastante precisión isoglosas y áreas con las variantes de los resultados de TRĬPĔDES a lo largo de todo el territorio analizado. Ejemplos como este son los que nos permiten hablar del *CorLexIn* como una especie de atlas lingüístico para el siglo XVII.

22 En consonancia con estos datos, los repertorios léxicos modernos del área asturleonesa (*DGLA*, Le Men 2005) no registran tampoco variantes del tipo de *treldes* sino que solamente parten de *trébedes*.

3.2 *Acetre/cetra*

Procedente del latín SĬTŬLA, pero llegado al castellano por intermedio del árabe *satl*, tenemos en castellano *acetre* (*DECH*, s. v. *acetre*), voz que registra ya Nebrija y es general en los diccionarios. El *DLE* da dos acepciones: una general, 'caldero pequeño con asa que sirve para sacar agua de las tinajas', y otra con la marca *religión*: 'caldero pequeño con asa en que se lleva el agua bendita para las aspersiones litúrgicas'. Ambas están presentes desde el primer repertorio académico, aunque con la particularidad de que on *Autoridades* —y así permanece hasta la edición de 1817— se anota que la primera acepción se usa «en el Reino de Granada y en otras partes» (*NTLLE*, s. v. *acetre*). La documentación del término es antigua y, como indican J. Corominas y J. A. Pascual, aparece ya en la documentación altomedieval leonesa bajo formas como *azeptre* o el diminutivo *azetrelio* (*DECH*, s. v. *acetre*).

Para las referencias relativamente abundantes en la documentación leonesa ha de tenerse en cuenta, sin embargo, que se trata de una voz que habitualmente no aparece en contextos que podríamos definir como laicos, sino que lo hace mayoritariamente en las relaciones de bienes con las que se dota un monasterio o entre los bienes de una iglesia, por lo general junto a otros objetos litúrgicos de un cierto valor, que generalmente aparecen solo en esos contextos[23]. Es muy probable, por tanto, que se trate de objetos foráneos que constituyen parte del tesoro de una iglesia, pero que se usan escasamente fuera de este ámbito y que, en consonancia, la palabra pertenezca más al registro eclesiástico que al general.

Por lo que se refiere a la documentación del siglo XVII que manejamos para el *CorLexIn*, la distribución de resultados de *acetre* presenta dos áreas bien definidas. Una en el Norte que incluye documentos de Álava y del enclave burgalés del Condado de Treviño y en la que, al menos en un caso, apunta hacia *acetre* como objeto de uso eclesiástico:

23 En la documentación de la catedral figura, por ejemplo, en «aquamaniles cum suos concos pares II, *acetre* ereo I» (León, año 967), «conko ereo et aquamanile ereo, *azetre* ereo uno, uasos argenteos duos» (León, 1042), «mortarios et conquos II, aquamanile I, coginas II^{as}, *azetrelios* II» (León, 1038), todos ellos correspondientes a dotaciones fundacionales de monasterios. Mucho menos frecuente es que se registre entre los bienes de seglares, aunque es posible hallar algún caso como el que aparece en un testamento dentro de una enumeración de recipientes metálicos: «II^{as} calderas, II^{os} conquos, I^{o} aquamanil, I^{a} quodma, I^{o} *azetre*» (Otero de las Dueñas, doc. nº 345, 1150?). La forma en diminutivo aparece en una serie similar en testamento del siglo X, conservado en copia de mediados del XI: «I^{o} conquo;/et I^{o} aquamanile; et I^{o} *acdrelio*» (Otero de las Dueñas, doc. nº 50).

vn *acetre* de cobre, viejo (Vitoria, 1638)
un *açetre* de cobre (Vitoria, 1638)
dos *azetres* de azófar para agua bendita (Puebla de Arganzón, Bu-1628)

Frente a esta área norteña, en la que el término aparece escasamente representado, figura de forma habitual en los inventarios de la zona suroccidental de la Península, con especial incidencia en la documentación de Huelva y algo menor en la de Sevilla, además de casos más aislados en Granada, en Badajoz e incluso en Canarias. Pese a esta localización en el occidente de Andalucía y en Canarias, no lo encontramos, sin embargo, en nuestros documentos de América[24]:

vn *asetre*, quatro reales (Huelva, 1617)
vnas trébedes, parrillas, y vn *azetre* y vn pailón (San Juan del Puerto, H-1632)
vn perol; vn *asetre* (Bollullos Par del Condado, H-1657)
vn *asetre* y una coladera grande y un casillo (Niebla, H-1660)
vna caldera mediana y un *asetre* (Rociana del Condado, H-1660)
un *asetre* de alatón (Trigueros, H-1633)
vn *asetre* (Almonte, H-1650)
vn *acetre* de cobre nuevo (Bollullos de la Mitación, Se-1634)
un *azetre*, quatro ducados (Sevilla, 1640)
vn *azetre* y un perol (Sevilla, 1640)
vn cubo y vn *acetre* de cobre nueuo (Sevilla, 1650)
dos *açetres*, vno grande y otro pequeño, de asófar (Gerena, Se-1651)[25]
vna caldera y un *azetre* y vn chocolatero y dos peroles (Sevilla, 1679)
vn *açetre*; vna caldera mediana (Segura de León, Ba-1659)
un *azetre* pequeño, de cobre (Montefrío, Gr-1661)
un *azetre* de plata, pesó quize onzaz (La Orotava, Tf-1663)

De lo que no cabe duda es de que en esta zona *acetre* se usa con la primera acepción que proporciona el *DLE*, constituyendo un objeto de uso cotidiano que está

24 No solo no aparece en nuestro corpus sino que tampoco lo hace en el *Corpus Diacrónico y Diatópico del Español de América* (*CORDIAM*). Únicamente tenemos documentado un caso en América que se sitúa fuera del rango cronológico que usamos en el *CorLexIn*: en un inventario de los bienes que los Jesuitas tenían en Mizque (Bolivia), realizado en 1771, se registra «un azetre de cobre por ocho pesos», lo que no hace más que confirmar que, fuera del área indicada para el suroeste peninsular, la voz se usa únicamente en el ámbito religioso y no para los efectos habituales de una hacienda.

25 Este ejemplo está en el inventario de una iglesia. De ahí quizá la diferencia del material del que está hecho, azófar, frente a los de uso doméstico, generalmente de cobre. En cualquier caso, demostraría que en la zona *acetre* se usa tanto con el sentido de objeto religioso como con el de objeto de uso cotidiano.

presente en el ajuar doméstico de muchas de las haciendas inventariadas o en las cartas de dote.

Si la zona suroccidental de la Península es el área en la que más vivo se conserva el uso de *acetre* en la lengua del siglo XVII, en el área suroriental encontramos otra variante de esta misma voz procedente del árabe, con la diferencia de que ahora estamos ante un orientalismo de procedencia catalana usado en la documentación de Murcia[26], Almería[27] y en partes de Albacete y Alicante. Se trata de *cetra*, del que abundan los ejemplos en la zona mencionada:

vna tinaja de agua y una *çetra* (Sax, A-1639)
vna *çetra* y una jarra de aranbre (Sax, A-1639)
vna *cetra*, vna tinaxa de poner agua y unas parrillas (Orihuela, A-1717)
una tenaja de tener agua con su *cetra* (Hellín, Ab-1636)
una tenaja de tener agua con su *cetra* (Hellín, Ab-1647)
una tinaxa de tener agua con su *çetra* (Vera, Al-1648)
una tenaxa grande de agua con su *cetra* (Cuevas de Almanzora, Al-1649)
una tenaxa con *çetra* (Alcantarilla, Mu-1613)
dos tinajas de agua con vna *zetra* con sus tapadores (Moratalla, Mu-1637)
tres tinajas de tener agua con sus tapadores y una *çetra* (Alhama de Murcia, Mu-1648)
dos tenajas de tener agua con *cetra* y tapador (Caravaca de la Cruz, Mu-1654)
quatro tenajas de tener agua, con *zetra* (Murcia, 1657)
dos *zetras* a medio traer (Mazarrón, Mu-1659)

No solo se trata de una variante formal sino que también implica diferencias de significado: mientras que el *acetre* se inventaría entre los útiles de la cocina, la *cetra* aparece, cuando se especifica, como complemento de las tinajas usadas para contener agua.

El *DLE* registra *cetra* 'escudo de cuero usado en la antigüedad', voz que el *NDHE* (s. v. *cetra*) considera tomada del latín *caetra* y «una voz antigua recogida en obras históricas y en los repertorios lexicográficos», sin referencias aún a *cetra* con la acepción que aquí interesa. En cambio, en el *DECH* (s. v. *acetre*) se cita como regionalismo murciano equivalente a *acetre*. Es justamente en los vocabularios murcianos donde encontramos referencias de este *cetra*: García Soriano (1980: 37), por ejemplo, la define como 'jarrito, generalmente de cobre, con un asa larga vertical para sacar agua, vino, etc., de las tinajas', al tiempo que ofrece

26 En inventario de Yecla (Murcia) de 1568 aparece también el apunte de una *cetra* citado entre un caldero y una tinaja (Puche 2002: 141).

27 Vivancos (2013: 126–127) localiza media docena de ejemplos de *cetra* en cartas de dote e inventarios de la comarca de Vera (Almería) de los siglos XVI–XVII, pero el término no aparece ya registrado en el *TLHA*.

algunos testimonios históricos y la relaciona con el catalán *cetra* y el valenciano *citra*. El catalán *sitra*, procedente del mismo étimo árabe que el castellano *acetre*, según el *Diccionari etimològic i complementari de la Llengua Catalana* (*DECat*, s. v. *setra*) se usa en la actualidad principalmente en la zona norte del dominio del catalán, pero testimonios como los citados, documentados en el área suroriental peninsular, evidenciarían que la voz se extendió mucho más al Sur, entrando como préstamo en el castellano regional de Murcia —como indican los diccionarios dialectales—, y también en Almería, en el área de Villena, en Alicante, o en la cercana Hellín, en Albacete.

En definitiva, lo que nos muestran los inventarios del siglo XVII, es un área quizá residual en el Norte peninsular y dos zonas diferenciadas en el Sur en las que los resultados del latín SĬTŬLA, si bien llegados por distintas vías, presentan una apreciable vitalidad en la lengua de la época.

4 Voces no documentadas

Otra de las utilidades que se puede obtener de este corpus no estriba tanto en la información que nos ofrece como en aquellas voces que es incapaz de registrar. Nos referimos ahora a innovaciones como los préstamos léxicos que el registro literario no tarda en reflejar pero que, en el campo documental, muestran a veces un grado de expansión mucho más lento. Por la época que tratamos, los préstamos más interesantes son los galicismos, americanismos e italianismos. A voces de esta última procedencia, como *bagatela* y *valija*, nos vamos a referir, aún corriendo el riesgo, claro está, de que más adelante localicemos un documento en el que se contradiga lo dicho aquí.

4.1 *Bagatela*

Uno de los italianismos que se señalan para esta época es la voz *bagatela* 'cosa de poca importancia o valor'. El término aparece por primera vez en el *NTLLE* en *Autoridades*, donde se define como «cosa menúda, de poco provecho, sin substáncia ni valor» y ahí se indica ya que procede del «toscano *bagatelle*», con idéntico sentido. Su inclusión en el primer diccionario académico se apoya en un fragmento de la *Gatomaquia* (1634) de Lope de Vega[28], en el que se ironiza sobre lo español y lo italiano (*NTLLE*, s. v. *bagatela*):

28 Con idéntica referencia al italiano, el *CORDE* nos ofrece un ejemplo anterior en otra obra de Lope: «Dame, Celia, el escritorcillo de los embustes. No os haga escrúpulo el nombre, que en verdad que no soy hechicera; que le llamo así por las *bagatelas* que

> Pero ¿dónde me llevan *niñerías*,
> que en Italia se llaman *bagatelas*
> ingiriendo novelas
> en tan funestos casos
> más dignos de Marinos y de Tasos
> (que de Helicona son solos y soles)
> que de mis versos rudos españoles?

La consulta de corpus históricos, como el *CORDE* o el *CDH*, nos muestra que *bagatela* o *vagatela* empieza a ser usado en castellano justamente a comienzos del siglo XVII, si bien el ejemplo tomado de Lope indica que se considera aún una palabra marcada como extranjerismo. En cualquier caso, hay ya una decena de ejemplos en este siglo, que aumentan sensiblemente desde comienzos del siguiente, probablemente, según J. Corominas y J. A. Pascual, por la influencia del italiano, pero también por intermedio del francés *bagatelle*, igualmente un italianismo (*DECH*, s. v. *bagatela*). Sea de una u otra forma, lo cierto es que el número de ejemplos en el castellano del siglo XVII parece suficiente como para deducir que el término llegado del italiano está ya razonablemente asentado en castellano (Terlingen 1960: 304).

Sin embargo, si repetimos la búsqueda en el *CorLexIn*, nos encontramos con que, entre los textos reunidos en este corpus[29], no se registra ningún ejemplo de *bagatela* y ello a pesar de que el concepto al que hace referencia esta palabra figura con frecuencia en los inventarios transcritos: minuciosos como son, al llegar en el inventario a un conjunto de objetos de escaso valor, no es raro que los escribanos cierren con lo que esperaríamos fuera —al menos en alguna ocasión— «otras bagatelas». Sin embargo, cuando necesitan inventariar objetos de escaso valor sin especificar a cuáles se refieren, no utilizan el término *bagatela* sino que, una y otra vez, los amanuenses se sirven de las voces patrimoniales con las que entraría en conflicto este préstamo. Así, junto a denominaciones genéricas como *cosillas* o expresiones como *de poco valor*, figuran en este contexto términos como *menudencias* —sin duda el más frecuente— junto a otros como *bujerías* —este a veces alterado en *brujerías*—, *baratijas* o el *niñerías* citado por Lope, pero nunca lo hace *bagatela*:

---

tiene, vocablo de un señor italiano, que me le ferió a un instrumento que yo tenía y que él codiciaba», *La Dorotea* (1632).

29 Algo similar ocurre en otro corpus documental como el *CORDIAM*, en el que *bagatela* aparece en una única ocasión, pero lo hace ya en un texto del siglo de XIX datado en Chile.

vna zestilla con una túnica y otra *cosillas* (Santander, 1676)
vna arquilla y quatro tocados y otras *cosillas* que tiene en ella (Cuenca, 1630)
vn aseo y una gargantilla y otra *buxerías* en veinte reales (Escalona, To-1644)
vna caxa y en ella vnas tocas, guantes, anus, abanico y otras *bruxerías* de muxeres (Argamasilla de Calatrava, CR-1661)
vna sartén y otras *buxerías* de por cassa (Jerez de los Caballeros, Ba-1661)
más unas *baratixas* de hierro en un arquita viexa (Alburquerque, Ba-1645)
vna arquimessa grande con *baratijas* […] vnas *baratijas* en ellas (Zaragoza, 1603)
seis herraduras con otras *baratijas* de yerro en un cajón de un bufete (Melgar de Tera, Za-1675)
unas *baratixas* de yerro […] vn cántaro con vnas *niñerías* (Valdelaguna de Abajo, Áv-1651)
una caxa de pino andada con tropillos y otras *niñerías* (Arnedo, LR-1639)
adereços de caveça y valonas y otras *niñerías* (Mérida, Ba-1642)
vna balona y vn estuche y vna toca y vn Anus y otra *niñerías* (Pinto, M-1653)
las *menudenzias* que están en el sobrado (Carbajales de Alba, Za-1653)
de *menudençias* de seruiçio de cosina, dies y ocho pessos (Cádiz, 1655)
otras *menudencias* de madera (Antequera, Ma-1628)
calderos, sartenes, candiles, assadores, asnillos de fuego y otras *menudençias* (Teruel, 1625)
vna caxita de zedro, de la costura, con *menudencias* de este efecto (Garachico, Tf-1695)
ottras surtijas, y asientos, y rostrillos de tocado de oro y *menudencias* (Puebla de los Ángeles, Puebla, Mx-1622)
vna poca de pita y otras menudençias que lleuó (Panamá, 1608)
una cajuela con su candadito de golpe y dentro de ella vnas *menudensias* de trapos (La Plata (Sucre), Bo-1703)
otras cosas de omenaje de cassa de *poco balor* (Ciudad de México, Mx-1622)
estas menudensias abía dejado por çer cosas tan de *poco balor* (Ciudad de México, Mx-1622)

Contamos incluso con un documento en el que el escribano no parece estar muy convencido del término apropiado que debe utilizar en este contexto y, después de tachar sucesivamente *brugerías* —confusión por *bujerías*— y *niñerías*, opta finalmente por *cosillas*, al que se le añade la expresión «de poco valor» que acompaña con frecuencia a estos asientos en los inventarios:

una arca de pino con sus cerraja y llaue con un gergonillo en ella y otras (tachado: *brugerías niñerías*) *cosillas* de poco ualor (Teruel, 1652)

4.2 *Valija*

Encontramos una situación similar para *valija*, término del que no parece haber duda respecto a su procedencia del italiano *valigia* (Terlingen 1960: 295; *DECH*, s. v. *valija*). La palabra la usa Cervantes en varias ocasiones (Bucalo 1998: 76)

y el *CORDE* aporta varias decenas de ejemplos (*balija*, *valija*) a lo largo de los siglos XVI y XVII. No obstante, el término no debía estar aún lo suficientemente asentado fuera de determinados círculos pues en el *CorLexIn*, pese a la frecuencia con la que los escribanos inventarían *bolsas*, *cofres*, *baúles, cajetas*, *carteras* o *mochilas*, no se registra ni un solo ejemplo de *valija* —lo mismo sucede en el *CORDIAM*—, algo que no ocurre con otros préstamos como el galicismo *maleta*, introducido ya en la Edad Media, o, lo que es más significativo, uno de entrada más reciente como es el náhuatl *petaca*, que es ya común en los textos del corpus procedentes de América y alcanza incluso a algunos de la Península y Canarias:

dos *petacas* de camino (Ciudad de México, Mx-1623)
quatro *petacas* de cuero con sus candados y llaues (Potosí, Bo-1677)
van en estos dichos ocho vaúles, y los dos fardos y dies *petacas* (Panamá, 1608)
dos *petacas* pequeñas […] dos *petacas* grandes; otras dos *petacas* con candados (Choconta, Co-1636)
la *petaca* aforrada en cuero de vaca […] una *petaca* vieja retobada con cuero de vaca […]
otra *petaca* de cuero de vaca con su cadenilla y armella (Santiago de Chile, Ch-1668)
dos *petacas* de mimbres forradas de badana con sus llaues […] dos *petacas* de minbre forradas en badana (Zaragoza, 1646)
una *petaca* aforrada, con su cadena (Pedroso, LR-1676)
dos candados de las dos *petacas* (Adeje, Tf-1695)
dos *petacas* (Garachico, Tf-1695)

La interpretación de estos datos no puede ser otra que constatar que voces como *bagatela* o *valija*, usadas principalmente en fuentes literarias, no han conseguido aún, fuera de este ámbito, desplazar a las formas con las que tradicionalmente se identifica su significado en castellano, a diferencia de lo que ocurre en el último caso con el americanismo *petaca*, que se extiende rápidamente no solo por América sino también por España[30].

5 Conclusiones

Como puede verse, un corpus específico como el que manejamos nos aporta, de un lado, un importante volumen de datos capaces de documentar históricamente variantes no habituales y, de otro, la posibilidad de disponer de una especie de atlas lingüístico del pasado, útil para poder fijar el área de expansión

30 De estos dos ejemplos no ha de deducirse que los documentos notariales representen necesariamente un registro que pueda considerarse conservador. Hay campos, como el de la vestimenta y el del textil, en los que las innovaciones léxicas se ven rápidamente reflejadas en los inventarios.

de una palabra o sus variantes, datos ambos imprescindibles para la lexicografía diacrónica.

La ventaja de disponer de corpus específicos como este estriba, sobre todo, en la posibilidad de completar y servir de apoyo a los grandes corpus históricos, que necesariamente han de tener un contenido más convencional. Su principal valor es el de ofrecernos una perspectiva distinta de la lengua histórica que, en cualquier caso, nos permite hacernos una idea de la complejidad diacrónica, diatópica o diastrática que encierra cualquier estadio lingüístico.

Referencias bibliográficas

ALEA = Alvar, Manuel (1991): *Atlas lingüístico y etnográfico de Andalucía*. Madrid: Arco Libros.

Bastardín Candón, Teresa (2015): «Léxico de la vida cotidiana en las particiones de bienes del AHP de Cádiz (s. XVIII)», en Teresa Bastardín y Mª del Mar Barrientos (eds.), *Lengua y cultura en el Archivo Histórico Provincial de Cádiz*. Cádiz: Universidad de Cádiz, 39–56.

Borrego, Julio (1999): «Leonés», en Manuel Alvar (dir.), *Manual de dialectología hispánica. El Español de España*. Barcelona: Ariel, 139–158.

Bucalo, Mª Grazia (1998): «Los italianismos léxicos en las *Novelas Ejemplares* de Miguel de Cervantes Saavedra», *Cuadernos de Filología Italiana* 5, 29–80.

CDH = Instituto de Investigación Rafael Lapesa de la Real Academia Española (2013): *Corpus del Nuevo diccionario histórico*. <http://web.frl.es/CNDHE> [último acceso: 20/09/2017].

CORDE = Real Academia Española: Banco de datos en línea *Corpus diacrónico del español*. <http://www.rae.es> [último acceso: 20/09/2017].

CORDIAM = *Corpus Diacrónico y Diatópico del Español de América*. <http://cordiam.org> [último acceso: 20/09/2017].

CorLexIn = Morala Rodríguez, José Ramón (dir.): *Corpus Léxico de Inventarios*. <http://web.frl.es/CORLEXIN.html> [último acceso: 20/09/2017].

DECat: Corominas, Joan (1980–1991): *Diccionari etimològic i complementari de la Llengua Catalana*. Barcelona: Curial Edicions Catalanes.

DECH = Corominas, Joan/José Antonio Pascual (1980–1991): *Diccionario Crítico Etimológico Castellano e Hispánico*. Madrid: Gredos.

DGLA = *Diccionario General de la Lengua Asturiana*. <http://mas.lne.es/diccionario> [último acceso: 20/09/2017].

DLE = Real Academia Española: *Diccionario de la Lengua Española*. <http://www.rae.es> [último acceso: 20/09/2017].

Egido Fernández, Mª Cristina (2016): «América vs. España: contrastes gramaticales y léxicos en documentación del siglo XVII», en Marta Fernández Alcalde, Elena Leal Abad y Álvaro S. Octavio de Toledo y Huerta (eds.), *En la estela del Quijote. Cambio lingüístico, norma y tradiciones discursivas en el siglo XVII*. Fráncfort: Peter Lang, 189–213.

Fernández Sevilla, Julio (1975): *Formas y estructuras en el léxico agrícola andaluz*. Madrid: CSIC.

Fichero general = Real Academia Española: *Fichero General*. <http://web.frl.es/fichero.html> [último acceso: 20/09/2017].

García Soriano, Justo (1980): *Vocabulario del dialecto murciano*. Murcia: Editorial Regional de Murcia.

Le Men, Janick (2005–2012): *Léxico del leonés actual*, vols. III (D–F) y VI (R–Z). León: Centro de Estudios e Investigación «San Isidoro».

Lloyd, Paul M. (1993): *Del latín al español. Fonología y morfología históricas de la lengua española*. Madrid: Gredos.

Morala Rodríguez, José Ramón (2012): «Datos sobre interferencias entre los sufijos *-dor* y *-dero* en un corpus del Siglo de Oro», en Mar Campos Souto, Ramón Mariño, José Ignacio Pérez Pascual y Antonio Rifón (eds.), *«Assí como es de suso dicho»: Estudios de morfología y léxico en homenaje a Jesús Pena*. San Millán de la Cogolla: Cilengua, 237–254.

Morala Rodríguez, José Ramón (2014): «El *CorLexIn*, un corpus para el estudio del léxico histórico y dialectal del Siglo de Oro», *Scriptum Digital* 3, 5–28.

Morala Rodríguez, José Ramón (2015a): «Derivados en *-dor* en la documentación del Siglo de Oro. Voces escasamente documentadas», en José María García Martín (dir.), Francisco Javier de Cos Ruiz y Mariano Franco Figueroa (coord.), *Actas del IX Congreso Internacional de Historia de la Lengua Española*, vol. II. Madrid: Iberoamericana-Vervuert, 1503–1519.

Morala Rodríguez, José Ramón (2015b): «Datos para la historia del *neutro de materia* en castellano», *Revista de Filología Española* XCV, 2º, julio–diciembre, 307–337.

Morala Rodríguez, José Ramón (2016): «Fuentes manuscritas del siglo XVII e Historia de la Lengua», en Marta Fernández Alcalde, Elena Leal Abad y Álvaro S. Octavio de Toledo y Huerta (eds.), *En la estela del Quijote. Cambio lingüístico, norma y tradiciones discursivas en el siglo XVII*. Fráncfort: Peter Lang, 373–388.

Morala Rodríguez, José Ramón y Mª Cristina Egido Fernández (2010): «Variantes formales en hiatos y diptongos en textos notariales del siglo XVII», en Rosa Mª Castañer y Vicente Lagüens Gracia (eds.),

De moneda nunca usada. Estudios dedicados a J. Mª Enguita Utrilla. Zaragoza: Institución Fernando el Católico, 423–435.

NDHE = Instituto de Investigación Rafael Lapesa de la Real Academia Española (2013–): *Nuevo diccionario histórico de la lengua española*. <http://web.frl.es/DH> [último acceso: 20/09/2017].

NTLLE = Real Academia Española: *Nuevo Tesoro Lexicográfico de la Lengua Española*. <http://www.rae.es> [último acceso: 20/09/2017].

Ortiz Cruz, Demelsa (2015): «Los inventarios de bienes en el norte peninsular: el caso de un inventario de un mercader zaragozano (1748)», *Res Diachronicae* 13, 49–57.

Perdiguero Villarreal, Hermógenes (2012): «Palabras en *-ero/-era* en protocolos notariales de Castilla», en Mar Campos Souto, Ramón Mariño, José Ignacio Pérez Pascual y Antonio Rifón (eds.), *«Así como es de suso dicho»: Estudios de morfología y léxico en homenaje a Jesús Pena*. San Millán de la Cogolla: Cilengua, 381–192.

Perdiguero Villarreal, Hermógenes (2016): «Peculiaridades léxicas en un inventario mirobrigense de 1632», *Scriptum Digital* 5, 135–145.

Pérez Toral, Marta (2014): «Huellas de lo oral en documentos notariales turolenses del Siglo de Oro», *Archivo de Filología Aragonesa* 70, 145–168.

Pérez Toral, Marta (2015): «Las expresiones *mesa de manteles* y *cama de ropa* en el siglo XVII: ¿ropa de hogar o mobiliario?», *Anuario de Letras. Lingüística y Filología* III/1, 131–167.

Puche Lorenzo, Miguel Ángel (2002): *Documentos jurídico-notariales del siglo XVI (1534–1590) del archivo de protocolos notariales de Yecla*. Murcia: Real Academia Alfonso X el Sabio.

RILG = *Recursos Integrados da Lingua Galega*. <http://sli.uvigo.es/RILG> [último acceso: 20/09/2017].

Terlingen, Juan (1960): «Italianismos», en M. Alvar (ed.), *Enciclopedia lingüística hispánica*. Madrid: CSIC, 263–305.

TLHA = Alvar Ezquerra, Manuel (2000): *Tesoro léxico de las hablas andaluzas*. Madrid: Arco/Libros.

Viudas Camarasa, Antonio (1980): *Diccionario extremeño*. Cáceres: Servicio de Publicaciones de la Universidad de Extremadura.

Vivancos Mulero, Mª Esther (2013): *La lengua del repoblador. Estudio histórico-lingüístico y tipología documental en el oriente del Reino de Granada. La Tierra de Vera (siglos XVI–XVII)*. Tesis doctoral, Univ. de Granada. <http://digibug.ugr.es/bitstream/10481/31703/1/22706021.pdf> [último acceso: 20/09/2017].

Miguel Ángel Puche Lorenzo

Estudio del léxico castellano a través de fuentes medievales murcianas

Resumen: Para profundizar y comprender la evolución del léxico español, pretendemos mostrar fuentes medievales inéditas en el contexto de los estudios lingüísticos. Son textos de naturaleza jurídica, conservados y redactados en la ciudad de Murcia durante el s. XIV, donde se inserta, a causa de las necesidades sociales, un amplio caudal léxico que no se ha puesto de relieve hasta el momento. Los datos obtenidos no solo nos descubrirán voces desconocidas, sino que también nos ayudarán a conocer cómo fue el contacto lingüístico entre castellano, catalán y árabe, principalmente, además de completar la historia del léxico español a través de la variedad que caracterizaba a las zonas periféricas, integrantes del mismo reino.

Palabras clave: Historia de la Lengua Española, Historia del Léxico, Léxico medieval, Murcia

Abstract: This paper analyses unpublished medieval sources in the context of linguistic studies in order to deepen and understand the evolution of the Spanish lexicon. The investigation takes into consideration legal texts, preserved and written in the city of Murcia during the 14th century, in which a wide volume of lexicon, not highlighted in the foregoing research, was inserted due to social needs. The data not only reveals unknown voices, but explores what the contact between Spanish, Catalan and Arabic was like, and provides valuable information to complete the history of the Spanish lexicon through the variety that characterized the peripheral zones of the same kingdom.

Keywords: History of the Spanish Language, History of Lexicon, Medieval lexicon, Murcia

> «Pero para poder nombrar cada cosa con su nombre… debemos saber los nombres de las cosas… Echemos una mirada por la casa y por el campo; a centenares se nos ofrecerán las cosas, los detalles, los particulares, las faenas y operaciones que no sabemos nombrar. Y, sin embargo, todo eso tiene o *ha tenido* su nombre; debemos conocer y usar esos nombres» (Azorín 1999 [1913]: 943).

1 Introducción

Conscientes de que la elaboración de una Historia del léxico español debe incluir a todos los actantes que participan de una misma lengua, en el espacio, en el tiempo y en la más diversa producción escrita, abordaremos en este

trabajo la aportación realizada desde territorios periféricos, en este caso Murcia, a la configuración del castellano en el periodo medieval. Para llevar a cabo tal proceso, será necesario comprender la situación lingüística, que implicaba necesariamente al plano léxico, en este espacio geofísico, los condicionantes que produjeron alternancias y sustituciones en épocas posteriores y la conservación de aquellas voces que, como testimonio fidedigno de un pasado, han pervivido hasta tiempos recientes. Tales ejemplos, junto a los procedentes de otros lugares, nos hacen ver la variación léxica de un reino, el castellano, factor que sustentó la clásica afirmación de García de Diego (1950: 104 y ss.) acerca de que nuestra lengua era rica en dialectalismos y pobre en dialectos.

Desde un punto de vista histórico, ha sido estudiado en profundidad cómo el antiguo Reino de Murcia, fundado por los árabes, se incorporó a la Corona de Castilla por parte de Alfonso X el Sabio y cómo fue repoblado por un importante contingente aragonés y catalán que, junto al castellano, se sumó a la numerosa población árabe que allí permaneció, además de otros de procedencia ultrapirenaica. Esta convivencia repercutió en la lengua de esa comunidad de hablantes y en propiciar una serie de características dialectales que han pervivido hasta tiempos recientes, en mayor o menor medida. Durante ese primer momento, se produjeron interferencias y se introdujeron rasgos fonéticos de procedencia catalana, sobre todo, que fueron perdiéndose con el paso del tiempo y por la progresiva castellanización, tal como estudió Díez de Revenga Torres (2008)[1]. Este hecho ha sido observado desde el punto de vista de la onomástica (Díez de Revenga/Puche Lorenzo 2002), así como la confluencia de las diversas situaciones lingüísticas en el campo de la toponimia (Díez de Revenga Torres 2008). A pesar de haber quedado muy bien descritos estos niveles lingüísticos, otros, sin embargo, no han recibido la misma atención, como sucede con el léxico. En este sentido, como afirmó Líbano Zumalacárregui (2012: 95), «advertiremos que no todas las diatopías han sido tratadas de la misma manera», de hecho, poco sabemos del léxico medieval del sureste peninsular, del solar ocupado por

1 Dado que son numerosos los trabajos que publicó esta autora sobre los temas referidos desde al año 1986, hemos preferido citar el volumen que de manera más reciente los recopiló. En cuanto a los rasgos fonéticos analizados, destaca la temprana datación, del siglo XIII, que hizo del seseo, en este caso de origen catalán, que le llevó a afirmar que «tal fenómeno se habría iniciado en la periferia por un problema de lenguas en contacto (provenzal/catalán/castellano)» (Díez de Revenga Torres 2008: 95).

el Reino de Murcia que se caracterizó por no haber dispuesto, históricamente hablando, de unos límites precisos y continuados en el tiempo. De hecho, en el plano léxico y en el periodo medieval[2], son escasos los trabajos que encontramos. Entre ellos debemos destacar la descripción realizada por García Soriano (1980 [1932]: XXV–LXIV) en el estudio preliminar al *Vocabulario del dialecto murciano*, las anotaciones sobre el léxico vinculado al campo semántico del agua por Díez de Revenga Torres (2014) o los cambios que se produjeron en el tránsito del siglo XV al siglo XVI, documentados en las fuentes notariales (Puche Lorenzo 2012).

En virtud del vacío existente, se hace necesario abordar el estudio del léxico medieval en fuentes conservadas y redactadas en Murcia, mediante el que se podrá advertir en mayor o menor medida la pervivencia de elementos árabes y los efectos de una mayor población catalana o aragonesa, llegada a partir de las repoblaciones del siglo XIII. No se puede olvidar, en ese contexto, que este territorio se conformó como una frontera con otro reino, el de Valencia, cuyo contacto fue, y es, duradero en el tiempo. Dado que existe un amplio corpus documental, hemos seleccionado el primer documento que emanó del concejo murciano, a partir de las necesidades sociales. Nos referimos a las *Ordenanzas municipales*, redactadas a lo largo del siglo XIV que, con toda seguridad, nos ofrecerán mayor riqueza léxica, similar a la que poseen los inventarios de bienes en años posteriores, que son los más antiguos conservados en esta zona. Un estudio de estas características se situaría al lado de «quienes entienden que la evolución lingüística es una manifestación del espíritu y vida de las comunidades de hablantes; a la cabeza de ellos figura Guillermo de Humboldt, el idealismo de Vössler, la neolingüística italiana y la integración histórico-lingüística de Menéndez Pidal», en palabras de Lapesa (1978: 17–18).

2 Se hace constar esta apreciación porque, desde un punto de vista dialectal, se han realizado numerosos estudios entre los que destacan los de Muñoz Garrigós (2008), principalmente. Así mismo, y con referencia expresa al periodo que pretendemos abordar, se incluyó documentación procedente y/o referente a Murcia en el *Diccionario* elaborado por Sánchez González de Herrero (2000).

2 Fuentes, objetivos y características o sobre qué estudiar

Con los datos expuestos, resulta fácil intuir que los primeros textos escritos en castellano en el Reino de Murcia datan de mediados del siglo XIII[3], y que su estudio pone de manifiesto la complejidad dialectal de la Corona de Castilla y la aportación que, desde la periferia, se realizó a su configuración, especialmente léxica. De manera que, dentro de las fuentes medievales, hallamos documentación notarial, de gran interés para observar la convivencia lingüística, el bilingüismo y la influencia fonográfica de estos fenómenos en ellos, el *Libro del Repartimiento*, que nos permite realizar un estudio de los pobladores o colonizadores de este territorio en esa primera época (principalmente catalanes, aragoneses y castellanos), accediendo a la toponimia y la onomástica, tal como realizamos con anterioridad (Díez de Revenga/Puche Lorenzo 2002), aunque para el estudio del léxico de la vida cotidiana se obtenga mayor riqueza en los documentos, también notariales, pero relativos aún oficios o a inventarios de bienes. En este caso, se suelen datar ya en el siglo XV, aunque también existe un amplio corpus del XIV que no dispone de tanta riqueza como los de esa tipología surgidos en el siglo posterior. Sin embargo, aun en fecha tan tardía en un principio, se llega a vislumbrar cómo podría ser la conciencia y habilidades lingüísticas en esta época a través de las lecturas que realizaban algunos de sus habitantes. De hecho, en el «Inventario de bienes *post mortem* de Antón Pérez de Valladolid, maestrescuela de la Iglesia de Cartagena», fechado en 1483[4], se crea un apartado para describir los libros del difunto, entre los que se encuentran:

> Un libro de dichos de alcaldes […] scripto en papel *en romançe*
> Un confesional scripto *en latin*
> Un formulario. Otro scripto *en latyn* con cubiertas de pargamino
> Un libro de medicina scripto en papel *con lengua catalana*
> Un libro de leys scritpo en papel las cubiertas blancas

3 Nos encontramos en el periodo comúnmente aceptado como «castellano de los siglos XIII y XIV», continuando con un concepto clásico de periodización dentro de la Historia de la Lengua. Las propuestas de fragmentación temporal han sido revisadas a lo largo del tiempo con la consiguiente variación de enfoque entre la historia interna y externa del idioma, tal como ha puesto de manifiesto Quilis Merín (2010: 50–54).

4 Se localiza este documento en el AHPMU, Fondo Notarial, Prot. 363, fols. 82r-85v. Al ser inédito, la transcripción es nuestra, así como las cursivas, empleadas aquí para poner de manifiesto las diferentes lenguas en que estaban escritos los libros enumerados, hecho este que podríamos vincular a la temática o área de conocimiento al que pertenece cada obra. No obstante, sí se advierte que no deberían resultar extrañas las lenguas que allí aparecen: romance (castellano), latín y catalán.

Un libro de títulos de la tercera partida
La quinta partida.

Para el objetivo que perseguimos en esta ocasión, nos acercamos a un texto nuevo que vendría a completar esas fuentes medievales, aunque, probablemente, su valor sería mayor que todas las anteriores, desde el punto de vista léxico, por la extraordinaria riqueza que posee y porque no ha sido estudiado desde esta perspectiva hasta ahora, salvo las referencias realizadas por Díez de Revenga (2013). Nos referimos al conocido como *Libro de ordenamientos municipales* (1305–1375), que se completó con unas *Ordenanzas para la guarda del campo* en el siglo XV.

El texto presentado es un manuscrito original redactado a lo largo del siglo XIV (y completado con unas ordenanzas del XV, como hemos indicado), porque iba ampliándose con el paso de los años, y que fue editado por Torres Fontes (1956, 1975, 1978, 1983 y 1985) en el contexto y como apéndice de diversos estudios históricos[5]. El manuscrito reúne todos los ordenamientos realizados por el concejo de la ciudad durante el reinado de Alfonso XI y el inicio del de Pedro I. Los acuerdos adoptados y allí expuestos se refieren no solo a la propia labor del concejo, sino también a la función que cumplían los corredores y los almotacenes, las medidas concernientes a los aspectos más diversos referidos a la vida urbana y al uso y disfrute de las tierras que configuraban la huerta, principalmente. Se cumple, de esta manera, un doble objetivo, por un lado, la recopilación de todos esos acuerdos y, por otro, el mantenimiento de su vigencia. Al abarcar toda esta normativa un periodo amplio, casi medio siglo, hay ciertas cuestiones repetidas en algunos de ellos (Torres Fontes 1983 y 1985).

Un aspecto que se debe tener en cuenta, referido a la tipología textual en la que se inserta, es que no estamos ante unas «ordenanzas», porque estas «alcanzan una categoría superior, un rango que no solo las hace obligatorias para todos tras su aprobación, sino que también significan una permanencia e inalterabilidad»[6]

5 Hemos de indicar que, a pesar de su publicación, se ha consultado el original en todo momento con el fin de comprobar la veracidad de algunas grafías que demostrarían la confluencia lingüística en este territorio. El manuscrito se conserva el Archivo Municipal de Murcia, Serie 3, nº 32 y posee en la descripción una datación diferente a la que ofrecemos aquí, sustentada la nuestra en los estudios históricos realizados sobre él. Del mismo modo, se encuentra disponible en línea una copia digitaliza en <http://www.regmurcia.com/servlet/s.Sl?METHOD=FRMSENCILLA2&sit=c,373,m,139,serv,Carmesi>.

6 Unas de las primeras conocidas bajo este nombre se redactaron bajo el reinado de los Reyes Católicos (Martínez Martínez 2011) o en 1536 (González Arce 2000).

(Torres Fontes 1985: 243), sino ante unas «ordenaciones» u «ordenamientos», aunque es preferible utilizar la primera denominación porque así es como se expresa el propio concejo, además de constituirse en una «suma de acuerdos, a veces dispares, fruto de criterios distintos en épocas diferentes, y en otras de confirmación o aprobación de acuerdos semejantes… para corregir acciones perjudiciales para el común ciudadano o para los dueños de las heredades» (Torres Fontes 1985: 243). En cierto sentido, esta recopilación viene a suplir la ausencia de una legislación y pone de manifiesto el aislamiento que, en este sentido, tenía el reino murciano. Si en los primeros años de la Reconquista se aplicaron los modelos sevillanos al respecto[7], después esto no era posible porque la distribución territorial era diferente, así como las necesidades que emanaban de la población. Uno de los actantes con mayor peso fue el almotacén, creado por Alfonso X el Sabio, puesto que fue el encargado de velar por el orden en todo lo concerniente a la vida cotidiana. Era este un continuador de lo establecido bajo la gobernación árabe y constituye una muestra de la convivencia religiosa, étnica y cultural que comenzó a partir de ese momento y que fue necesario legislar u ordenar, con el fin de evitar altercados o malas prácticas sociales. Por ello, sus principales funciones consistían en velar por el correcto uso en campos como la sanidad y la limpieza, el abastecimiento y la venta y el control de pesos y medidas (Torres Fontes 1983: 71–88).

En virtud de la descripción que hemos realizado del manuscrito objeto de estudio, tanto por el contenido que reúne, como por su datación, podemos afirmar que constituye un excepcional testimonio para el análisis del léxico, concerniente a un territorio específico, aunque no necesariamente exclusivo. A ello se une el hecho de que posee una extraordinaria validez para nuestro trabajo porque cada uno de estos acuerdos está perfectamente fechado, lo que nos proporciona cierta garantía de autenticidad además de la posibilidad de poder datar ciertos fenómenos con precisión, aparte de estar localizado en el espacio, aspectos estos que han sido puestos de relieve en multitud de ocasiones en cuanto al estudio de la documentación notarial. En consecuencia, el objetivo de estudiar el léxico medieval murciano está plenamente justificado. No obstante, ante los numerosos testimonios encontrados, muchos de ellos con una sola ocurrencia en el documento, nos lleva a tener en consideración la interesante apreciación

7 Pensemos que en 1266, el rey concedió a Murcia el fuero de Sevilla, es decir, el *Fuero Juzgo*, aunque veinte años después la ciudad no disponía de un ejemplar del código. Las reclamaciones realizadas y las dudas que planteaban algunas leyes, en cuanto a su vigencia, parecen indicar que, sin embargo, algún ejemplar debían tener a pesar de que no tuviera las suficientes garantías legales (Pérez Martín 2002: 57).

de Müller (2004: 69–70), al diferenciar en el campo de la lexicología histórica entre palabras puntuales y palabras generales, es decir, las que aparecen en un solo autor y una sola obra, como muestra de idiolectismos, y las que superaron el umbral de la Edad Media, clasificación esta que llevará a plantearnos si realmente pervivieron o no, o si, por otra parte, constituyen ejemplos puntuales fruto de la convivencia o la interferencia lingüística, principalmente.

3 La documentación de Murcia digitalizada y en la red

Antes de proceder al análisis léxico, objeto de este estudio, resulta imprescindible conocer el acceso que disponen los investigadores a la documentación murciana en la actualidad. La publicación, en papel, de fuentes documentales redactadas en el espacio geográfico que nos ocupa ha sido constante, en virtud de la cantidad de testimonios conservados, así como por el buen estado que, de forma general, han mantenido. Así mismo, su localización no se encuentra exclusivamente en archivos regionales, sino que es necesario acudir a otros, de ámbito catedralicio o nacional, que permita reunir todo ese material textual de este espacio geográfico concreto.

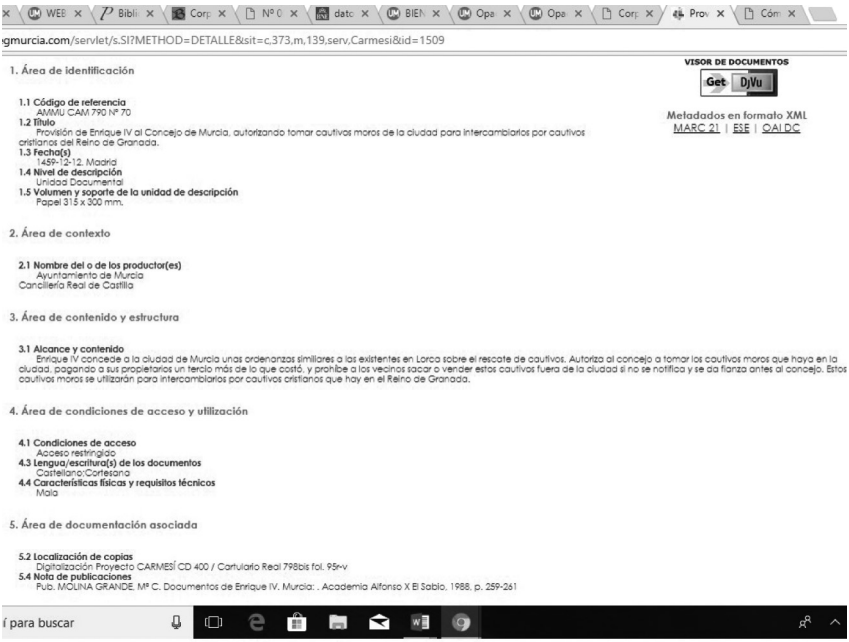
En cuanto a su datación, debemos tener en cuenta que, como afirmó Díez de Revenga (2008: 41), «la documentación más antigua que conocemos en romance data de la conquista cristiana en 1243». Este hecho no implica que no deba ser considerada en el entramado de la evolución del castellano medieval, puesto que su estudio reflejará la complejidad dialectal de la Corona de Castilla. De hecho, Menéndez Pidal incluyó ocho ejemplos en los *Documentos Lingüísticos de España I. Reino de Castilla* (1919). La preocupación por conservar y permitir el acceso a la documentación llevó a que se formara, bajo el sello de la Real Academia Alfonso X el Sabio y dirigida por Juan Torres Fontes, la *Colección de documentos para la historia del Reino de Murcia* (*CODOM*), que publicó veinticuatro volúmenes (1963–2006) y propició la participación de numerosos autores. En ellos se recopila un interesante material desde la Edad Media hasta el siglo XVI. Como continuidad de ese periodo, principalmente, surgió la colección *Fuentes Históricas de la Región de Murcia*, editada por la Universidad de Murcia, que ha reunido hasta la fecha (2000–2015) textos de diversa naturaleza a partir del siglo XVI, sobre todo[8]. Las transcripciones o ediciones críticas ofrecen criterios diversos que obligan al filólogo, en algunas de ellas, a revisar el soporte original del documento.

8 Al margen de estas colecciones, han ido viendo la luz numerosas transcripciones de textos que no incluyo por no ser ese nuestro objetivo en estos momentos.

La digitalización de fuentes y las transcripciones disponibles en la red nos ponen de manifiesto cómo las nuevas tecnologías nos proporcionan nuevos métodos de estudio. En este sentido, el mundo digital nos indica la escasa atención que la documentación murciana ha recibido, por una parte, pero nos facilita, por otra, los medios para iniciar una tarea de investigación lingüística de gran interés. Si se accede a corpus de una relativa reciente creación, como el elaborado por la Red Internacional *CHARTA*, se observa que, de los 2076 documentos, transcritos paleográficamente y presentados críticamente, solamente 25 son redactados en Murcia y fechados, casi todos, en el periodo alfonsí. Queda patente, por tanto, la baja representatividad del territorio[9], es decir, constituyen un material valioso pero consideramos que insuficiente. Para el siglo XVII, y con el objetivo de estudiar el léxico español en su variedad geográfica, nació el corpus *CorLexIn*, dirigido por José Ramón Morala. En aquel se incorporaron 106 inventarios de bienes murcianos, de diversas localidades, que proporcionan una magnífica panorámica del léxico español en la época que abarca el corpus.

Al lado de la representatividad que tienen las fuentes documentales murcianas en los corpus digitales, elaborados estos con criterios fiables, debemos indicar la digitalización de esas fuentes y su disponibilidad en la red. La digitalización de los documentos conservados en los archivos históricos de Murcia fue la finalidad del proyecto Carmesí (<http://www.regmurcia.com>), que, en la actualidad, ha conseguido que todos los documentos del periodo medieval estén disponibles a través de su página, así como otros de siglos posteriores, principalmente Actas Capitulares. A través de un visor que permite su acceso en PDF o DjVu se puede consultar, con gran calidad de imagen, el texto buscado. Además, se han incorporado algunas transcripciones incluidas en el *CODOM* o, si esto no sucede, se remite al lugar donde está publicada la transcripción del documento, tal como se observa en la siguiente imagen:

9 Excepto uno, son todos documentos de la cancillería real y se encuentran en el subcorpus *CODECAR*, dirigido por Mª Nieves Sánchez González de Herrero. Si se consulta en la web de la Red la distribución geográfica, realizada a partir del número de formas, el 1.22 % son murcianas.



A través de lo expuesto resulta fácil deducir la falta de fuentes transcritas digitalizadas y el valioso material disponible en la red, donde se tiene libre acceso a una documentación que no debe olvidarse para el mejor conocimiento del castellano medieval. Hemos consultado, mediante este recurso, el texto que nos proponemos estudiar, lo que indica el nuevo cauce de investigación que nos ofrecen las Humanidades digitales.

4 Palabras y más palabras. Análisis léxico

Junto a la idiosincrasia que representa este tipo textual, bastante extraño dentro de todo lo que se ha venido estudiando o se conoce hasta ahora, se encuentra la enorme variedad y riqueza léxica, ya que queda constancia de un inventario extraordinario de voces que afectan a la vida cotidiana y que, difícilmente, habrían encontrado reflejo en otros documentos o textos de la época referidos al territorio que nos ocupa. Dada la amplitud de todo lo que se podría exponer aquí, solo me referiré a tres grandes campos léxico-semánticos (el agua y el riego, el vestir y la alimentación) que ilustrarán a la perfección todo lo que venimos

argumentando. Se podrían haber creado otros pero, dada la extensión del trabajo, hemos creído conveniente mostrar los que consideramos más significativos en esta ocasión.

4.1 El agua y el riego

Las disposiciones vinculadas al agua y riego de la huerta son las ordenaciones con las que se inicia la obra que estudiamos. Conocemos cómo el agua, durante la Edad Media, proporcionaba distribuciones de territorio dada la importancia que poseía este bien tantas veces escaso (Puche Lorenzo 2010: 110). El agua y su disfrute mediante el riego debían legislarse adecuadamente en el contexto de la huerta murciana, con el fin de evitar fraudes y abusos[10]. Dejando al margen la polémica entre el origen romano o árabe del sistema de riego murciano, lo que sí está claro es la influencia de esta última tanto en la toponimia, como en los nombres de acequias. Si nos detenemos en la partición del agua de las acequias de allende del río (fols. 9v-12r), encontramos las denominaciones de *Adava*[11], *Alcantarilla, Menjalfaco, Beninabia, Benihalel, Tell açengur, Almohaja, Tell Alquibir, Albadell, Alcatell e Erardor, Alquibla, Alguasça, Alfandech, Alfanchec, Alhariella, Almunna, Alfariella, Crepas, Barralhariella, Beniaçot* y *Benihazeran*. Aquende el río localizamos las de *Carabixa, Nelba, Alfatego* o *Çaheriche el chico*. Al lado de ellas, se pueden encontrar otras de origen latino o castellano, como *Villanueva, Quixanos, Chopo, la Molera* o *Montagudo*.

La presencia del árabe, en consecuencia, se dejaría notar en el modo de denominar la acción y el proceso del riego. El agua se extrae de un *azud*[12], que es una especie de presa:

10 En Díez de Revenga (2013) se encuentran referencias a algunas de las voces que, en este campo, mencionaremos.

11 Se incluye en esta ocasión una muestra de la convivencia cultural que se produjo: «los otros algaidoneros que son en esta dicha acequia de la Adava de christianos e de moros» (fol. 9v). Recordamos que, aunque existe transcripción del texto, la que aquí aparece es nuestra.

12 La voz aparece recogida en el *Diccionario de Autoridades* con la marcación diatópica de voz usada en los reinos de Aragón, Murcia y Valencia. A través del *CORDE*, comprobamos que los ejemplos se encuentran localizados en el *Vidal Mayor* (s. XIII) y en documentación notarial de Navarra (s. XIV) (s. v. *açut*). Indicamos en esta nota que la consulta de los diccionarios que se mencionen a partir de ahora se ha realizado desde el *Nuevo Tesoro Lexicográfico de la Lengua Española* (*NTLLE*).

> que es de la açequia mayor Alquiblia del açud fasta Aljuçer que monden e ayan mondado… (fol. 1v).
> Iten son las de Turbedal dos mil tafullas e an de aver quatro partes quando viniere el agua del açut mayor pagando en ello su parte e su derecho commo los otros (fol. 9v).

De ahí se distribuye a través de *acequias* o *azarbes*. La primera voz dispone de una mayor extensión que la segunda si consideramos la información lexicográfica[13] pues, según el *Diccionario de Autoridades*, era propia de la huerta de Murcia:

> Et el açarbe mayor de parte del acequia Aljeufia deven fazer mondar todas las otras de Beniajam fasta el Javali. Et el otro açarbe mayor de parte del açequia Alquibla deven fazer mondar todas las otras de Beniporche fasta el Alcantarilla (fol. 4r).
> Otrosí todos aquellos que desfizieren açarbe o acequia después que la tierra fue de cristianos fasta agora (fol. 5r).

La repartición del agua se hacía también mediante *rafas* o *paradas*, es decir, aberturas para obtener ese recurso o represas transitorias que desviaban la dirección de la corriente. La primera voz es también un arabismo, no documentado en *CORDE*, pero presente en *Autoridades* con este significado, al haberse obtenido el ejemplo de los *Ordenamientos de Lorca*, posteriores en el tiempo al que estamos analizando ahora:

> De los que fizieren parada o rafa en el acequia mayor (fol. 1v).

En el desarrollo de los ordenamientos siguientes observamos que el binomio desaparece y se muestra una preferencia léxica por *parada* en vez de *rafa*:

> De commo deven regar e de commo alinpien el lodo de la parada (fol. 2v).
> E si es el dia o a la noche que se deviere desfazer la parada ovieren a regar dos o tres omnes o mas el postrimero de todos sea tenudo de desfazer la parada e de alinpiar el acequia commo dicho es (fols. 2v-2r).

Aunque, si el ordenamiento va dirigido hacia el disfrute del agua por parte de la población árabe, el término utilizado es *rafa* en vez de *parada*, como se desprende de la partición del agua de la acequia Aljufía:

> Que los moros puedan fazer rafa en la çequia mayor […] por ende los dichos omnes buenos ordenaron que los dichos moros puedan fazer la dicha rafa pagando cada año çient maravedís para obra del acequia (fol. 80v).

---

13 En el *CORDE* no hemos hallado ningún ejemplo en el intervalo comprendido entre 1200 y 1500, que es el que utilizamos con el fin de comprobar la localización de una voz en la Edad Media. Corominas y Pascual (*DECH*) (1980–1991: s. v. *azarbe*) describen también su presencia casi exclusiva en territorio murciano.

Respecto a la forma de desplazar el agua, de elevarla, se contaba con *annoras*[14] (norias) y *çenias*, rueda hidráulica característica de las zonas de huerta, de hecho, la primera voz ha mantenido la marcación lexicográfica de murcianismo, mientras que la segunda lo ha hecho con la región valenciana:

> Contrasto en lo de las annoras e acequias […] todos los pleitos en contrastos que acaesçieren por fecho de aguas asi de las annoras e de Sangonera como de las acequias (fol. 6r).
> Primeramente se falla que son todas las tafullas de las dichas acequias a menos la heredat de las çenias de Miguel de Rallat e de Turbedal (fol. 9v).

Pero existían otros sistemas de riego que, probablemente, no eran tan generales. Las formas que los denominan son de origen árabe y no las hemos encontrado recogidas en ningún otro tipo de corpus textual o lexicográfico. Solamente se incluyen en el *Vocabulario del comercio medieval* formado a partir del legado de Gual Camarena (en línea), que utiliza como fuentes los estudios historiográficos de esta zona, aunque no se mencionan fuentes documentales. Nos referimos a *alfayt*[15], que se correspondería con un riego por crecida:

> deven pagar el açequiage de la huerta de Murçia que riegan del agua de Segura a diez maravedís una tafulla e de la que riega de Sangonera e del alfait de Tinnosa e de las tierras del río a cinco maravedís una tafulla (fol. 3v).

Esta voz se usaría junto a *algaidón* o *dalgaidón*, forma de riego por elevación. Quienes realizaban estas funciones o los lugares donde se llevaban a cabo eran conocidos como *algaidoneros*:

> Et fallase por christianos e moros sabidores Dalgaydones que diez Algaydones cada uno con un capaço pueden tirar una parte de las dichas partes e mas. Et este mas es por razon que riegan con gran trabajo. Et porque maguer que toman el agua non pueden regar tanto commo farian de pie llano (fol. 9v).
> e desto an gran mejoría porque el verano se uso más el agua que en el invierno. Et son diez y ocho oras de dia e seys de noche. Et asi viene a los dichos heradieros que puedan

14 A propósito de la evolución de esta voz, se puede consultar Muñoz Garrigós (2008: 85–93).

15 Bajo la forma *alfaide*, encontramos en diccionarios no académicos, desde Castro hasta Rodríguez, la acepción de «marea viva», alejada del significado que posee en este texto. Sin embargo, Zerolo incluye esta voz como un topónimo menor referido al nombre de una acequia valenciana. A pesar de no contar con otros testimonios lexicográficos sobre ella, sí aparece recogida en glosarios elaborados en el contexto de análisis histórico sobre el agua en Murcia (Martínez 2013: 136). Esta misma anotación resulta válida para la voz que se analiza a continuación de *alfayt*. En *DECH* se considera este arabismo propio del andaluz.

> tirar cada dia que se quisieren cinco algaydones de los sobredichos e no mas de la dicha çequia de la Dava. Et partense desta guisa e manera todos los otros algaidoneros que son en esta dicha açequia de la Adava de christianos e de moros (fol. 10r).

Observamos, como muestra de la integración de esos arabismos, que se crean nuevas voces a partir de ellos según los procesos de derivación propiamente castellanos. Lo hemos advertido en *algaidonero* pero también es fácil comprobarlo con los resultados que se constatan a partir de *acequia*. De ese modo nacen creaciones léxicas como *acequieros* o *çequieros* o *sobreaçequiero*[16], para designar a quienes se ocupan del mantenimiento y guarda de las acequias, brazales y azarbes:

> Del sobreçequiero. Ofiçio del sobreçequiero que guarde e procure en quanto pueda el pro e el bien de las açequias et que afinque a los çequieros que fagan tener las açequias e los braçales et los açarbes mondadas (fol. 1r).

En el contexto de la huerta, se debía pagar un impuesto en relación con la conservación de estos cauces, al que se denominaba *cequiaje* o *acequiaje*. No hemos hallado ejemplos en *CORDE*, pero lexicográficamente, desde que Salvá introdujo *cequiaje*, se consideró valencianismo, hasta 1936, momento en el que desapareció su marcación diatópica. Después Alemany y Bolufer incorporó *acequiaje*, esta vez como murcianismo, y así permaneció en el diccionario académico a partir de 1925, edición en la que se introduce la voz con la *a* protética. En los *Ordenamientos* se constata que ambas variantes se utilizan de forma indistinta:

> Et en qual manera deven fazer pagar el açequiage los que lo non quisieren pagar (fol. 3v).
> Otrosí los çequieros por fazer mondar las açequias e las fronteras dixessen de palabra e diesen por escripto que fallavan gentes rebeldes en tres maneras… de que deven pagar el açequiage de la huerta de murçia (fol. 3v).
> De commo deven coger el çequiage. Los cogedores de los çequiages comiencen de coger a la Sant Johan en la manera que ya es dicho e todos los herederos quel non avran pagado XV días (fol. 4v).

Si estos arabismos lograron una expansión geográfica considerable en la zona oriental peninsular, los vocablos creados con posterioridad no corrieron la misma suerte pues unos sí lo consiguieron, mientras que otros se han quedado bajo la denominación de regionalismos. P*artidor* se convierte en un claro ejemplo de ello. Se refiere a una obra que se destina a distribuir, por medio de

16 Ninguna de estas voces presenta ejemplos en el *CORDE* (recordemos que el intervalo temporal en el que nos movemos es el comprendido entre 1200 y 1500), y solo la primera, *acequiero*, recibió atención lexicográfica, en la segunda edición de *Autoridades* a partir de las *Ordenanzas de Granada*.

compuertas en diferentes conductos, las aguas que corren por un cauce. Todo parece indicar que son creaciones deudoras de la influencia catalana, a través del uso sufijal que se desprende, según García Soriano (1980 [1932]: XCIV).

> Et la açequia Aljeufia fasta la prima fila de Beniscornia. Et debe yr con cada uno dellos un omne de pie por que si algún embargo fallaren en la açequia o en los partidores que lo alinpie el omne con un gancho (fol. 2v).
> Almunna que toma al partidor de Alhariella son mil e çiento e cuarenta tafullas e toma dos partes e quinta e media e toma de los VII días e noches (fol. 11v).

Resulta de interés, igualmente, la presencia del verbo *escorrer*[17], considerado como variante de *escurrir* y ese podría ser el significado que aquí se recoge:

> Otrossi ordenaron e pusieron para siempre que cada unos puedan fazer en lo suyo dentro lo vedado balsas quanto para sus linos tan solamente e que non las alquilen nin presten a otros pero estas balsas no escorran en el rio (fol. 8v).

Debió adquirir otros matices semánticos que permitieron la derivación a *escorredor*[18], refiriéndose a la compuerta que permite administrar las aguas a una acequia o canal o a un cauce, de pequeñas dimensiones, utilizado como desagüe.

En el campo léxico del riego, encontramos otros testimonios destacables formados mediante procesos de prefijación, como sucede en formaciones del tipo *sorregar*, *sonregar* y *sorregamiento.* Si la presencia de la *n* epentética proporcionó que Salvá y otros diccionarios no académicos la consideraran una voz anticuada[19], la no inclusión de ese sonido introdujo en las páginas del diccionario académico un murcianismo efímero (desde *Autoridades* hasta la edición de 1791). Con respecto al sustantivo deverbal, no hemos hallado otros testimonios que indiquen su inclusión en repertorios de estas características:

> De los que sonregaren tierra a vezino. E otrosi sy sonregaren tierra a vezino que gelo quiera demandar aya esta misma pena mas si non gelo quiere demandar non aya calopnia por el sorregar (fol. 2r).

17 Así la hemos leído en el *Fuero de Teruel* (*CORDE*).

18 Ambos significados han quedado registrados en repertorios lexicográficos posteriores. De hecho, Terreros y Pando le otorga una marca diastrática al uso como desagüe, mientras que el diccionario académico, tras introducirla en 1925, la considera diatópicamente como murcianismo. A pesar de estar registrada en el texto que analizamos, sí constatamos su presencia a través de documentación notarial del siglo XIV (Puche Lorenzo 2010).

19 En García Soriano (1980 [1932]) se encuentran definidas *sonregar* y *sonriego*.

> Et las mayores dende arriba que se partan según los ordenamientos primeros pero del sorregamiento que aya el señor querelloso el terçio de la calonna e emienda del danno (fol. 7r).

4.2 El vestir

Este campo está plenamente relacionado con el ámbito comercial y de transacciones durante la Edad Media, convirtiendo esta labor en objeto de la legislación del momento. Los abusos que se podían producir en la compra y venta de las más diversas mercancías o mercadurías eran patentes y, en ocasiones, traslucían ciertas desavenencias entre los diversos grupos étnico-religiosos que convivían. Por ello, el ordenamiento de 1351 afirma, una vez reunido el consejo de *albalanes*[20], que «Et sobre que fue demostrado en consejo que los corredores jodios fazian en su ofiçio muchas cosas con enganno. Et commo quier que gelo provavan con christianos…» (fol. 12v). La limpieza del oficio culmina con un apartado dedicado a «que ningún corredor que non vista nin use para si vestidos nin ninguna otra cosa quel sea librada para vender en su casa nin fuera de su casa nin lo preste a otrie que lo use nin lo vista» (fol. 13v). Tras esto, se distribuyen y enumeran las diversas mercaderías que son objeto de mediación. Las que se refieren a este apartado son *paños* (fols. 17r y 17v) que se describen, principalmente, a partir de su origen: *paño de Françia, de Perpiñan, de Narbona, lonbardizco, cubiertas de ypre*[21] o, por último, de *blanqueta*, que era un tejido basto realizado con lana. En la misma medida se encuentra la *grana* (fol. 17v) que puede ser *grana apurada, grana verde* o *grana mustia*; la corambre (fols. 18r y 18v), donde se incluyen *cordovan, badanas bermejas, cuero vacuno, cueros de parage, boquinas, moltoninas*[22] *añina*[23], *pieles de cabritos* y *vestido de coranbre de los conejos*; la brunetería (fol. 19r)

20 Arabismo que se utiliza para designar una función y no un tributo o carta. La lexicografía nos indica que no tenía casi uso en Castilla, pero sí alcanzó difusión en la Corona de Aragón. En el *DECH* tampoco se recoge esta acepción.

21 Se refiere esta voz a la ciudad flamenca.

22 No hemos encontrado ningún repertorio castellano que recoja esta palabra, pero sí lo hacen los diccionarios del catalán donde se describe como «pell de moltó», cuyo equivalente castellano sería *zamarra* o *zalea* (*Diccionari enciclopèdic de la llengua catalana* 1938). La variante *moltolina*, por otra parte, se encuentra definida en los diccionarios de Domínguez y Zerolo como «piel de cordero que se estrae de Levante», y los datos de la geografía lingüística indican que era utilizada en zonas aragonesas colindantes con Cataluña o catalanohablantes (Alvar 1991a: 95–96).

23 Este término solo está presente en diccionarios no académicos, desde que la introdujera Domínguez, con el significado de «lana del cordero que se esquila por primera vez».

donde se detallan *varas de picoteño, de cañamos o de las marragas, pieza de fustán, pieza de fustán de seda, paño de paño de leyda lobuno o de valençia* o *de Segovia*; se hace mención a cómo regular el precio de las *venderías*, incluso se introducen apartados referidos a los *espadadores* (fols. 47r y 47v) y *tejedores* (fol. 38r). En este último caso se pone de manifiesto la ausencia de legislación, pues lo que hace el Concejo es trasladar lo que se había establecido en la ciudad de Sevilla:

> Este es traslado de un capítulo de los usos de Sevilla que dize así. Estos almotacenes an de reconosçer todos los pesos e todas las medidas de toda la çibdat a quien quier que las tenga tres vezes en el año. Et en todas an de poner su señal coñosçida. Iten ay otro capitulo que dize así. Todos los texedores que son en Sevilla que tienen varas e pesos para requerir que paguen cada una al almotaçen cada uno doze sueldos (fol. 38r).

Esto nos hace ver que las medidas también eran objeto de supervisión de la figura del *almotacén* y se refleja, por tanto, aquellas que se convierten en usuales dentro de la ciudad y el precio que se obtiene de cada una de ellas (fol. 21r). Estas son la *cantara, media fanega, vara, çelemin, medio çelemin, medio açumbre* y *taferia*[24].

Como queda constancia de la variedad en este campo seleccionado, nos detendremos en el apartado dedicado a la seda. La industria sedera tuvo gran importancia en Al-Ándalus que, ya en el siglo VIII, disponía de numerosos talleres cuya producción competía con los de Bagdag o Bizancio. En ese contexto, Murcia contribuyó con elaboraciones de gran calidad, aunque sin alcanzar la fama que adquirieron los vecinos productos de Almería. Después de su incorporación a la corona castellana, decayó la producción serícola porque no se aprovechó la herencia árabe, como había sucedido con el sistema de riego, por ejemplo, pero tampoco se introdujo ninguna nueva. A pesar de ello, sí continuó, de forma familiar y artesanal, una producción modesta que se correspondía con el mercado de seda en bruto. Esto ha llevado a afirmar que el único centro de producción de seda en Castilla, durante el siglo XIV, era Murcia, aunque existan débiles argumentos que sostengan tal afirmación (Eiroa 2017: 24). Conociendo, por tanto, este contexto, podemos situar el documento integrado en los ordenamientos analizados, donde queda constancia de las tarifas que tenían los productos realizados con seda de baja calidad. Está fechado este en 1313 y se indica lo siguiente:

El *Fichero general* de la Real Academia tiene depositadas numerosas fichas referidas a ella y sus variantes de género, aunque casi todas las fuentes donde se localiza son altomedievales. En el caso del *CORDE*, solo hay registrado un ejemplo del siglo XVI, fuera del intervalo temporal que habíamos acotado.

De cada libra de seda fina de cada una de las p*a*rtes dos din*e*ros
De cada libra de azache de cada una de las p*a*rtes VII din*e*ro
De cada libra de parval de cada una de las p*a*rtes medio din*e*ro
De la libra de ma*n*chapa de cada una de las p*a*rtes medio din*e*ro
De la libra de cadars de cada una de las p*a*rtes medio din*e*ro
De la libra de filadis de cada una de las p*a*rtes medio din*e*ro (fol. 17v).

De las voces aquí contenidas, observamos que *seda* y *filadiz* no resultan extrañas en el contexto del castellano. *Azache*, por su parte, es un arabismo estudiado por Serrano Niza (2007: 564). Junto a las apreciaciones sobre su etimología, nos interesa ahora su datación y su origen. Está recogida en el ámbito lexicográfico desde *Autoridades*, utilizando la fuente de las *Ordenanzas de Granada*, de hecho, las primeras dataciones de esta seda basta son del siglo XVI o finales del XV y dentro del ámbito andaluz. Nuestro ejemplo es bastante más antiguo y ello nos lleva a indagar algo más sobre su localización. En las ordenanzas granadinas[25] leemos: «Item que qualquier maestro oficial que texere en paño de seda y cebare en ellas atanquía o azache o aduque o cadarço o seda de Murcia o otra seda basta semejante», epígrafe que se repite casi de forma idéntica en las *Ordenanzas de Málaga*: «Yten, que qualquier maestro o oficial que texiere en paños de seda atanquias o açache o ad(a)[u]que o cadarce o seda de Murcia o otra seda basta semejante» (Martín Acosta 2010: 256). La presencia en el texto de la procedencia murciana de la seda parece indicar que estamos ante un arabismo originado y difundido a partir de ese territorio. En la actualidad ha perdido vigencia y se mantiene en sentido metafórico, como cosa o persona descuidada o desastrosa, en una parte pequeña de la población, al igual que sucede con *azogue*, arabismo procedente del ámbito de la minería en este caso. Con respecto a los términos restantes, hemos de señalar que no hemos hallado ningún tipo de documentación sobre *manchapa* (o *machapa*)[26] y *parval*, que Martínez (1988: 154) define como «seda parval, pequeña o de hebras de escasa longitud; machapa o chapa[27],

24 No hemos encontrado ningún repertorio lexicográfico ni corpus textual que recoja esta palabra. Torres Fontes (1983: 89) da la equivalencia entre *taferia* y *cántaras*.

25 Citamos a partir del ejemplo recogido en el *Fichero general* de la RAE.

26 Es necesario apreciar que en el manuscrito se realiza un trazo sobre la vocal que indicaría la abreviación de una nasal, aunque los estudios realizados sobre la historia de la seda se refieren a la forma léxica sin la nasal.

27 Bajo esta forma se encuentra la acepción siguiente en *Autoridades*: «Se llama también el capullo de seda mal formado, ú desmedrado. Es voz de Murcia», aunque solo apareció en esa edición. En el *Fichero general* (RAE), se registra otro ejemplo de esta voz, procedente de un documento notarial de D. Juan Manuel: «las cuerdas de las tres (cartas) son tajadas de cintura de seda e dotra cosa que fuera puesta en lauor de seda

que era la procedente de "capullos parches", con poca seda a consecuencia de la prematura muerte del gusano, floja y de poca consistencia». El último término que ocupa nuestra atención es *cadars*[28] que equivale a *cadarzo*, considerado arabismo en *Autoridades* y después helenismo. A partir de su forma gráfica, se puede considerar que se introdujo a través del catalán[29].

4.3 La alimentación[30]

Otra de las funciones que tenía el almotacén consistía en supervisar los productos que se vendían en la ciudad destinados a la alimentación de sus habitantes. A quien ostentaba este cargo se le requería que fuera «un omne bueno que sea vezino e leal e de buena fama que jure en poder de los jurados usar en ello bien e lealmente» (fol. 22r). Por ello, uno de los ordenamientos que regulan sus funciones va destinado para que «Ningun omne mercadero nin otro alguno christiano moro nin jodio vezino nin estraño non venda nin tenga a vender grana pobre, çafran nin otras mercadurías nin averes que fuesen encamarados con engaño» (fol. 23r), lo que nos indica la importancia adquirida por aquellos productos que tenían ese destino. La alimentación se trasluce, por tanto, en una necesidad fisiológica cuyos hábitos pueden ofrecer variaciones y variantes entre las diferentes comunidades sociales en función, muchas veces, de las disposiciones comerciales y de las disponibilidades que ofrecía el terreno. A raíz de ello, se ha contemplado esta parcela del léxico como un vocabulario especializado que puede estudiarse a través de tratados y recetarios de cocina, obras anteriores al siglo XVI como el *Arte cisoria* o *Manual de mugeres en el qual se contienen muchas y deversas reçeutas muy buenas*, tratados médicos, «literatura contable», obras literarias, el refranero (Eberenz 2014). Tras comprobar el interés que ha

verde e vermella cardena e amarilla e la cuerda dellotre es de filo o machapa cardena bermeja e blanca». Puede ser una alteración de *pacha*, documentada en los derechos de la seda de Granada de 1787 (Gallardo Fernández 1805: 299): «y que de la seda en rama llamada pacha, procedente de los desperdicios de las demás clases, que se emplea en labores bastas para usos mas comunes».

28 Las transcripciones posteriores reflejan *cadarzo*, aunque la lectura *cadars* es clara y no ofrece ninguna duda. En este caso, como en algún otro comentado con anterioridad, se advierte la introducción de catalanismos, entendidos como voces catalanas que se transcriben sin la consiguiente adaptación gráfica al castellano.

29 Bajo esa forma se recoge en diccionarios catalanes como el de Fray Magín Ferrer (1854).

30 Conocemos la diferencia entre alimentación y culinaria, con sus respectivas interpretaciones desde el punto de vista diacrónico (Eberenz 2016: 83).

despertado en la actualidad el léxico de este campo semántico, aportamos un nuevo tipo de fuentes de estudio a partir de las que observaremos tanto un interesante caudal léxico de productos alimenticios, fiel reflejo de una sociedad multicultural de principios del siglo XIV, como de determinadas apreciaciones que van destinadas a la elaboración de esos productos. En definitiva, lo que se utiliza para comer y cómo se cocina.

Son numerosos los ordenamientos que intentan legislar los productos alimenticios más diversos como las frutas y las verduras. Ocupan, por consiguiente, un lugar especial los *figos* (fol. 23r), el *agraz*[31] (fol. 24r), la *uva* (fol. 24r), el *aceite* (fol. 31r), el *vino* (fol. 32r), la *cebada* (fol. 33r), los *espárragos* (fol. 41v), la *turma*[32] (fol. 41v), la *hortaliza* (fol. 42v) y la *fruta* (fol. 43r), en la que se mencionan los *priscos*[33]. En estos apartados, el almotacén debe velar para que no se produzca ningún engaño en el precio, en las medidas de capacidad, en la higiene de su transporte, en que estén maduros los productos, en el caso de la fruta, o que no se establezca timo con el peso, como sucede con las hortalizas pues «Todos aquellos que vendiesen hortaliza que la tengan fresca e linpia e non la remogen nin echen agua por razón de mas pesar» (fol. 42v), todo ello junto con el importante lugar que ocupa el pan, como producto elaborado y básico en la alimentación. A pesar de que se ha afirmado que el consumo de frutas y hortalizas no era estrictamente alimentarse y podía denotar pobreza o cierta frugalidad (Eberenz 2016: 82), pensamos que en el ámbito social que nos ocupa estas disponían de una representación considerable, pues su cultivo era prioritario en todo el dominio de la huerta, sin que ello implicara necesariamente las circunstancias antedichas.

En estrecha relación con el comercio se encuentran las especias que se utilizaban para sazonar y elaborar otros productos, aunque algunos de ellos también podían tener un uso medicinal. Bajo la denominación de *especiería* (fol. 18r) se introduce un ordenamiento que establece el precio que debían alcanzar: *girofre*[34],

31 Entendido como la uva que todavía no está madura, de ahí el vínculo con *agrio*, *DECH* (s. v. *agrio*).

32 Raíz o tubérculo, conocido como turma de tierra, de color rojo o negro.

33 Especie de durazno o de melocotón.

34 Se corresponde con el castellano *girofe* o *giroflé*. Aunque está documentada en las obras lexicográficas españolas del siglo XVIII, la localización de este galicismo se extiende, sobre todo, por el oriente peninsular. Esta especia era parecida al clavo y la nuez moscada y dudamos de su equivalencia con el *jengibre*, pues este último está también incluido en la enumeración.

medio dinero; *çafran*, siete dineros; *pimienta*, medio dinero; *jengibre*, *canela*, medio dinero; *greda*[35], *cominos*, *alcaravea*[36], *matahalua*[37], *alegría*[38], *xenus*[39], *culantro*, *almenlones*[40], medio dinero; *pan de açucar*, medio dinero; *polvo de açucar*, tres dineros; *gala*[41], *regalicia*[42], *xabon de losa, alum de roca*, siete dineros; y *sal de conpas*, medio dinero.

Otro de los productos que debe ser ordenado es la carne y todos los aspectos concernientes a su venta y su procesamiento, razón por la que, en 1348, se incluye un apartado especial para el oficio de carnicero, en el que se percibe una extrema preocupación por la limpieza e higiene y porque no se produzca engaño alguno en la venta de estos productos, además de asuntos relativos a cuestiones religiosas: «Los carniçeros non fagan degollar en la carneçeria a judíos nin a moros vacas nin carneros nin otras reses ningunas so la dicha pena» (fol. 27r). En el procesamiento de la carne se indica que vendan «las telas del puerco e del cabrito con las frexuras» o «que el quarto del cordero en VI dineros e la cabeça e la corada con la tela por un quarto» (fol. 26r bis). *Frexura*[43] y *corada*[44]

35 Esta sustancia mineral tenía, entre otros usos, la capacidad de aclarar el vino.

36 Se utilizaba como sazonador, al ser aromática y de sabor picante. Este arabismo está documentado en *CORDE* en tratados patológicos y en el *Arte cisoria*, ya en el siglo XV.

37 Esta forma gráfica se refiere a *matalahúva*, *anís*.

38 Es esta el *sésamo* o *ajonjolí*.

39 Con la forma gráfica aquí representada solamente hemos encontrado dos ejemplos fuera de nuestro texto y se localizan en un *Libro de recetas*, consultado a través del *CORDE*. Se corresponde esta voz con *ajemuz*, arabismo que en castellano equivale a *neguilla*. Es esta una simiente olorosa utilizada para amasar el pan.

40 El trazo sobre la *e* pretónica nos hace pensar en *almenlones* y no *almelones* como transcribe Torres Fontes (1978: 259). Constituiría una variante de *almidón* influida por el árabe y considerada un catalanismo por Alvar (1991b: 22).

41 No hemos hallado referencia lexicográfica alguna respecto a esta voz, aunque sí hemos advertido su presencia en algunos textos incluidos en el *CORDE*, del siglo XIII, como los ordenamientos de Jaime I y los peajes de Zaragoza. Ello nos hace pensar que nos hallamos ante un orientalismo que encontraría su origen en el griego γαλα, gala, con el que se hace referencia a la *galactita* o *piedra de leche*.

42 *Regaliz* y *regaliza* son las formas más frecuentes en castellano. Aunque se encuentra recogida junto a ellas *regalicia*, los datos del *CORDE* nos indican que es más frecuente en el oriente peninsular.

43 La consulta de diccionarios de la lengua catalana corrobora esta afirmación.

44 La palabra se encuentra incluida en la tradición lexicográfica española, aunque en las obras académicas tuvo marcación de orientalismo a partir de 1992. Con anterioridad, desde 1925, y de forma intermitente, se consideró como diatopismo de *Ast*[*urias*]. Por ello, la consideración como orientalismo debe ser revisada puesto que se encuentra documentada en gallego y en zonas leonesas.

se corresponden con el castellano *asadura* pero, mientras que la primera es una voz catalana, la segunda es característica del dominio aragonés o, mejor dicho, oriental. Del mismo modo, se especifica que «de los puercos que tajaren e vendieren la carneçeria que non aparten los enplexes» (fol. 27r) y que la carne en venta esté en perfecto estado, no sea mortecina, por lo que se introduce «que los carniçeros den tablas a çiervos e a carnes rafalinas» (fol. 28r), «sisa de buey o vaca rahalín quatro dineros» (fol. 34v bis) o «et por buey o vaca rafalina» (fol. 58r). Las voces que nos interesan son *enplexes*, de la que no hemos encontrado ninguna documentación, y *rafalina*, de la que no existen tampoco otros testimonios disponibles, aunque está documentada por García Soriano (1980 [1932]) y Griffin (1964: 110) considera que este arabismo es un préstamo del catalán, presente también en tierras valencianas[45].

Por último, cabe señalar la importancia que adquiere en el seno del comercio y, en consecuencia, para los hábitos alimenticios de la población el pescado. El ordenamiento realizado para la pescadería se sustenta en parámetros similares a los establecidos para la carne (higiene, venta por separado del género del día anterior, penalización de la reventa, etc.), sin olvidar aquello que se refería a la convivencia religiosa en la ciudad: «Que non vendan pescado a moros nin a jodios en días de ayuno. En la Quaresma nin en los días de viernes o de ayunos los pescadores que non vendan pescado fresco a moros nin a jodios fasta pasado medio dia» (fol. 32v). Es este el campo que presenta mayor riqueza léxica, justificada por la cercanía con el mar y por la convivencia de pobladores que lo dotaron de un amplio caudal denominativo. Los bienes que proporciona el mar, y en menor medida el río, aparecen en diversos ordenamientos. El primero de ellos hace mención a los precios, relacionado esto con la medida de capacidad utilizada para la venta, y se enumeran las siguientes voces: *congrio, pixotas, sardina o aletría*[46] *salada,*

45 Probablemente sea un arabismo de difusión oriental y, debido a la temprana datación de nuestro texto junto con los testimonios valencianos, podría plantear la hipótesis de una difusión inversa, es decir, de sur a norte.

46 El término presentado ofrece un valioso testimonio. Posee el significado de fideos, aunque aquí está vinculado con algún pescado pequeño lo que nos permitiría pensar en un uso metafórico, por un lado, o que, dado que se cocinaban esos fideos con pescado, haya adquirido un uso metonímico. Este razonamiento se vería sustentado por otra localización de *aletría* en el texto: «non se entienden en sardina nin en boga nin aletria nin en caramell ni en otros pescados semejantes que sean menudos» (fol. 33v). En cualquier caso, estamos ante un arabismo que ha pervivido en Murcia, así se incluyó en *Autoridades*, pero que dispone de gran difusión en otras lenguas peninsulares como el catalán, en *alatria* (Trias 1996: 43–46), o el portugués (Nimer 2005^{2}: 243).

melva[47], *toñina*[48], *anguillas, arenques, sardina* o *danchova* (fol. 18v).

Más adelante se incluye otro ordenamiento sobre la pescadería (fols. 33r y 33v), copiado en 1373, donde se establecen los precios a que debe venderse el pescado y donde, además, se introduce una división de las diferentes especies a partir de la manera de ser cocinado: «De esa guisa es esclareçido en razón del pescado, qual es de salsa e qual es de freyr». Mientras que la *salsa* guardaría relación con lo *cocido*, *freír* supondrá una transformación del alimento a partir de aceite o grasa a elevada temperatura, acción que quedó definida en la primera parte de la *General Estoria* (Eberenz 2016: 91). Las denominaciones encontradas se recopilan, principalmente, en este fragmento:

> Pescado de tal salvo dalfín e bestina[49], muelle*s*[50] e pajeles[51] q*ue* pesare*n* una libra es de salsa. Et los de menos pesso fuere*n* de freyr. Todo el pescado sobredicho q*ue* troxere*n* fuere sal [peso] es de freyr.
> Mujoles[52], liças, dentoles[53], pagres[54], oradas, saros, corballs[55], scorp*ione*s, arañas, palomidas[56], morrudas q*ue* pesare*n* cada una una libra o más se entendían por pescado de salsa. Et los de menos peso fuere*n* de freyr.

47 Está documentada lexicográficamente en Terreros por primera vez, mientras que la Academia la incorpora en 1925. No existen datos en el *CORDE*, dentro del periodo acotado, y el *Fichero general* recoge algunos ejemplos, la mayoría procedentes de testimonios murcianos. Alvar (1970: 184) la documenta sin embargo en todo el sur y sureste peninsular a raíz de la nomenclatura oficial ictionímica.

48 Voz propia de Andalucía, según *Autoridades*.

49 Solo hemos encontrado un testimonio, dentro del *Fichero general* de la RAE, en el que se especifica su uso en Murcia, Málaga y Almería.

50 No aparece en ninguna de las fuentes de consulta.

51 La voz es común a todo el dominio castellano, pero Alvar (1970: 170) considera que es un catalanismo al igual que se lee en *DECH* (s. v. *pagro*).

52 Presente en la tradición lexicográfica española, a partir del DRAE-1992 se incluyó la información del origen catalán, como así era considerada en el *DECH*.

53 Registrado únicamente en una de las fichas académicas que reproduce un fragmento de la obra de Ruperto Nola.

54 Terreros se hace eco de esta voz y nos indica que es un pescado procedente de la Galia narbonense (también de allí llegaron repobladores), mientras que en el *Fichero general* de la RAE se encuentra la descripción de Máximo Torreblanca que la consideró catalanismo, correspondiente con el *pagro* castellano. En el *DECH* (s. v. *pagro*) leemos que es la voz del catalán de Valencia utilizada para llamar al *besugo*.

55 Hallamos la forma *corval*, documentada por Venceslada, según lo que proporciona el *Fichero general* de la RAE.

56 Registrada en diccionarios bilingües como los de Vittori y Minsheu, junto con el testimonio de Terreros, se considera una voz catalana (Sempere 1995: 264).

> Congrio fresco, sipias, calamares, espetos, sorguers, mármoles, obladas, sardinas e todo otro pescado menudo es de freyr[57].

Esta nómina se completa con las formas *boga*, *caramell* y *sorra* que se incluyen al final de las ordenaciones que constituyen este apartado. La distribución que se efectúa, en virtud de la manera en que deben ser cocinados, se corresponde con el tamaño; de la misma manera, no se realiza ninguna distribución entre pescado y moluscos o cefalópodos, desconocida en aquel momento. El análisis de las voces aquí contenidas nos desvela interesantes datos, a pesar de que haya algunas que resulten comunes o generales como *sardina*, *congrio*, *arenque*, *anguila, congrio, arañas, sepias* o *calamares*. En primer lugar, el arabismo ha perdido la predominancia que tenía en los campos anteriores que hemos presentado y ahora aparece de forma testimonial en *aletría*. El resto de pescados, procedentes en su mayoría del mar Mediterráneo, ofrece en su denominación una huella catalana notable bien porque la voz tenga esa procedencia, bien porque represente una adaptación fonográfica a esa lengua. Se pone de manifiesto, en suma, la temprana datación, y a veces única, de muchas de ellas, la extensión geográfica que tuvieron durante la Edad Media y la intervención o interferencia del catalán, o valenciano, en ese proceso denominativo que nos llevará a plantearnos, en algunos casos, en una influencia inversa, de sur a norte.

5 Conclusión

El estudio del léxico de nuevas fuentes documentales nos proporciona interesantes datos que, en nuestro caso, vienen a completar las lagunas existentes en relación con territorios poco atendidos desde esta perspectiva. De esta manera, se advierte cómo la periferia contribuyó a la configuración del léxico español, atravesando etapas donde las confluencias eran latentes, a pesar de que determinadas huellas de ese pasado desaparecieran por los diferentes procesos de nivelación, mientras que otras pervivieron y han dotado de ciertas peculiaridades idiomáticas a los diversos enclaves de nuestra geografía. En el caso de Murcia,

57 Salvo los términos que hemos citado como generales en castellano, el resto no tiene referencia alguna en la tradición lexicográfica española ni en los corpus que venimos manejando. Sin embargo, el vínculo común con áreas donde se habla catalán o valenciano está demostrada por la recopilación léxica elaborada por Mariano Bru en el siglo XVIII (Corcoll i Llobet 2000), donde se incluyen todas las voces que no hemos localizado. A ello se une la forma gráfica con la que están representadas en nuestro texto. Aunque es necesario indicar que la datación más antigua que hemos podido localizar sobre ellas se encuentra en estas ordenaciones murcianas.

la convivencia entre repobladores de procedencias varias, junto a la población que continuó residiendo en la zona, tuvo efectos que se reflejaron en la lengua y los testimonios escritos dejaron buena cuenta de ello. No obstante, se percibe que, mientras que características que se observaron en otros niveles lingüísticos, como el fonético-fonológico, desaparecieron antes por la progresiva castellanización, aquellas huellas que se insertaron en el plano léxico se mantuvieron vigentes durante más tiempo e, incluso, han perdurado hasta la actualidad. En este sentido, dentro de los objetivos que nos marcamos, hemos conformado un amplio corpus léxico a través de un texto inédito donde se datan en fechas tempranas algunas voces, mientras que otras lo hacen por vez primera a través de este texto, constatando en todo momento la confluencia de procedencias léxicas que se correspondían con el modelo social que imperó durante la Edad Media en Murcia, desde el siglo XIII hasta los umbrales del periodo áureo. Es decir, queda configurado un inventario léxico murciano caracterizado por introducir, de manera notable, arabismos, catalanismos y aragonesismos, aunque, en ocasiones, esta separación no se pueda establecer y sea aconsejable hablar de orientalismos.

A raíz de los datos que hemos presentado en relación con tres campos semánticos concretos, extraemos varias conclusiones. El léxico recopilado a través de las ordenaciones y ordenamientos se caracteriza por disponer de un elevado número de arabismos y orientalismos, aunque no se distribuyen de la misma manera, de hecho, se percibe una disposición inversa. Si en la administración del agua para el riego, casi todas las voces tienen un origen árabe (*acequia*, *azarbe*, *azud*, *alfait*, etc.) y las creaciones que remiten a otras procedencias, como la catalana, son testimoniales (*partidor* o *escorredor*), en el vestir la preeminencia es de voces castellanas con algunos ejemplos de arabismos (*azache*) o voces de aspecto oriental, generalmente catalán (*cadars*), mientras que en la alimentación se perciben algunos arabismos (*rafalina*, *aletría*) frente a todo un enorme contingente oriental referido a las denominaciones de pescados (*corvalls*, *dentol*, *caramell*, *sorguer* o *pagres*).

A pesar de la consecuente y progresiva castellanización del territorio, se ha afirmado y demostrado la constante pérdida de elementos, ya sea fonéticos o léxicos, de otras procedencias lingüísticas, advertido con notable precisión en el paso de la Edad Media al Renacimiento (Puche Lorenzo 2012). No obstante, la mayor parte de las voces que aquí se localizan y datan de forma temprana en el seno del castellano nos permite plantearnos el análisis de este caudal léxico desde la perspectiva de un léxico especializado. En el campo del riego, por ejemplo, hemos comprobado que, aunque se ha considerado gran parte de estas voces como murcianismos, los datos lexicográficos y textuales disponibles nos hablan de una extensión mayor, restringida, casi siempre, a zonas donde existía huerta,

además de continuar vigentes hasta la actualidad. Por ello, si no se difundieron más allá de esos enclaves geográficos fue porque no existía el modelo de distribución del agua que mediante esos sistemas se planteaba. Este hecho nos permite pensar en el concepto de tecnicismo histórico (Puche Lorenzo 2015) que se sustentó, en este caso, a partir de la aportación del árabe. Este mismo proceso podría avalar el uso de orientalismos, principalmente de vínculo catalán o valenciano, en las denominaciones de los pescados, cuya transmisión se ha mantenido hasta nuestros tiempos tal como han demostrado los estudios de ictionimia. Un caso híbrido estaría representado por lo referente al vestir que, en la seda concretamente, muestra el uso de arabismos (*azache*) y de orientalismos (*cadars*). El primero se conservó a lo largo del tiempo; el segundo acabó siendo sustituido por su correspondiente castellano (*cadarzo*). Pensemos que la castellanización de un orientalismo implica una sencilla, y en ocasiones analógica, adaptación fonográfica, pero la de un arabismo se reflejaría en un proceso de sustitución léxica.

A todo lo anterior, es necesario reflexionar sobre la datación de los términos reflejados en este trabajo. La temprana aparición de todos ellos en documentos murcianos, como el que aquí hemos presentado, ofrece la posibilidad de pensar en el nacimiento de esas voces en territorios periféricos y que luego se integraron, dependiendo de los casos, en el seno del castellano o, incluso, en el catalán. Si hubo fuerzas centrípetas que operaron para el nacimiento de algunos fenómenos fonéticos, propuestas por Corominas y continuadas o avaladas por Díez de Revenga Torres (2008: 95 y ss.), también podría deberse a esa tendencia el nacimiento de algunas denominaciones. En consecuencia, este tipo de fuentes pone de manifiesto la complejidad léxica del castellano en la Edad Media, la importancia que poseen los territorios periféricos, olvidados hasta ahora desde esta perspectiva, y la necesidad de seguir estudiando nueva documentación.

Referencias bibliográficas

A.A.V.V (1938): *Diccionari Enciclopèdic de la Llengua Catalana*. Barcelona: Editions de la Junta d'Exposicions d'Art de Catalanya.

Alvar López, Manuel (1970): «Ictionimia y geografía lingüística», *Revista de Filología Española LIII*, 1/4, 155–224.

Alvar López, Manuel (1991a): «Antigua geografía lingüística de Aragón: Los peajes de 1436», en Tomás Buesa y José María Enguita (coords.), *I Curso de Geografía Lingüística de Aragón*. Zaragoza: Institución Fernando «el Católico», 11–104.

Alvar López, Manuel (1991b): «La falsa convergencia de los derivados de *Amĭndŭla y Amidum», en Roberto Dengler Gassin (ed.),

Estudios humanísticos en homenaje a Luis Cortés Rodríguez, I. Salamanca: Universidad, 19–23.

Azorín (1999): *Clásicos y modernos*, en Miguel Ángel Lozano Marco (coord.), *Obras escogidas, II. Ensayos*. Madrid: Espasa.

CARMESÍ = Catálogo de Archivos de la Región de Murcia en la Sociedad de la Información. <http://www.regmurcia.com/servlet/s.Sl?METHOD=FRMSENCILLA&sit=c,373,m,139,serv,Carmesi> [último acceso: 07/09/2017].

CHARTA = Corpus Hispánico y Americano en la Red: Textos Antiguos. <http://www.corpuscharta.es> [último acceso: 07/09/2017].

Corcoll i Llobet, Antoni (2000): «Una llista de peixos valencians de Mariano Bru (1778)», en Josep Massot i Muntaner (coord.), *Homenatge a Arthur Terry*. Barcelona: Abadía de Montserrat, 5–56.

CORDE = Real Academia Española: *Corpus diacrónico del español*. <http://www.rae.es> [último acceso: 07/09/2017].

CorLexIn = Morala Rodríguez, José R. (dir.): *Corpus Léxico de Inventarios (CorLexIn)*. <http://web.frl.es/CORLEXIN.html> [último acceso: 15/10/2017].

DECH = Corominas, Joan/José Antonio Pascual (1980–1991): *Diccionario crítico etimológico castellano e hispánico*. Madrid: Gredos.

Díez de Revenga Torres, Pilar (2008): *Estudios de Historia de la Lengua Española (desde la Edad Media a nuestros días)*. Murcia: Real Academia Alfonso X el Sabio.

Díez de Revenga Torres, Pilar (2014): «Arraigo y evolución del léxico castellano», en María Bargalló Escrivá, María Pilar Garcés Gómez y Cecilio Garriga Escribano (eds.), *«Llaneza». Estudios dedicados al profesor Juan Gutiérrez Cuadrado*. Anexos *Revista de Lexicografía*, 23. La Coruña: Universidade, 483–492.

Díez de Revenga Torres, Pilar/Miguel Ángel Puche Lorenzo (2002): «Onomástica castellana y onomástica catalana en tierras fronterizas durante la Edad Media», *Estudios de lingüística, ELUA* 16, 309–324.

Eberenz, Rolf (2014): «El léxico español de la alimentación y la culinaria en su historia. Fuentes y líneas de investigación», en Vicente Álvarez Vives, Elena Díez del Corral Areta y Natacha Reynaud Oudot (coords.), *Dándole cuerda al reloj: ampliando perspectivas en lingüística histórica de la lengua española*. Valencia: Tirant lo Blanch, 23–46.

Eberenz, Rolf (2016): «De lo crudo a lo cocinado: sobre el léxico fundamental de la culinaria en la historia del español (siglos XIII a XVII)», *Revista de Filología Española* XCVI, 1, 81–112.

Eiroa Rodríguez, Jorge (2017): «El trabajo de la seda en Murcia durante la Edad Media». *Seda. Historias pendientes de un hilo. Murcia, siglos X al XXI*. Murcia: Editum.

Ferrer, Magín (fray) (1854²): *Diccionario catalán-castellano con una colección de* 1670 *refranes*. Barcelona: Imprenta y librería de Pablo Riera.

Fichero general = Real Academia Española: *Fichero general*. <http://www.rae.es> [último acceso: 07/09/2017].

Gallardo Fernández, Francisco (1805): *Origen, progresos y estado de las rentas de la Corona de España, su Gobierno y Administración*. Madrid: Imprenta Real, T. III.

García de Diego, Vicente (1950): «El castellano como complejo dialectal y sus dialectos internos», *Revista de Filología Española* XXXIV, 107–124.

García Soriano, Justo (1980 [1932]): *Vocabulario del dialecto murciano*. Murcia: Editora Regional.

Griffin, David (1964): «El castellano raleo, ¿arabismo desconocido?», *Boletín de la Real Academia Española* XLIV, 107–111.

González Arce, José Damián (ed.) (2000): *Ordenanzas de la Ciudad de Murcia (*1536). Murcia: Universidad.

Gual Camarena, Miguel (legado) (en línea): *Vocabulario del Comercio Medieval*. <http://www.um.es/lexico-comercio-medieval/> [último acceso: 07/09/2017].

Lapesa, Rafael (1978): «Historia lingüística e historia general», en *Buscad sus pares pocos: tres ensayos*. Madrid: Gredos, Cátedra Seminario Menéndez Pidal-Universidad Complutense, 13–24.

Líbano Zumalacárregui, Ángeles (2012): «Historia y léxico medieval del País Vasco: la tierra, el hombre y su hábitat; transición del latín al romance», en Gloria Clavería Nadal, Margarita Freixas Alás, Marta Prat Sabater y Joan Torruella i Casañas (eds.), *Historia del léxico: perspectivas de investigación*. Madrid: Iberoamericana, 93–125.

Martín Acosta, Mª Dolores (2010): *Ordenanzas de Málaga de* 1611. *Edición y estudio léxico*. Tesis doctoral: <https://riuma.uma.es/xmlui/bitstream/handle/10630/4580/TDR_MARTIN_ACOSTA.pdf?sequence=6> [último acceso: 07/09/2017].

Martínez Martínez, María (1988): *La industria del vestido en Murcia (siglos XIII–XV)*. Murcia: Real Academia Alfonso X el Sabio.

Martínez Martínez, María (ed.) (2011): *Unas ordenanzas inéditas de la Huerta de Murcia durante el reinado de los Reyes Católicos*. Murcia: Junta de Hacendados de la Huerta de Murcia.

Martínez, María (2013): *La cultura del agua en la Murcia medieval (ss. IX–XV)*. Murcia: Editum.

Menéndez Pidal, Ramón (1919): *Documentos lingüísticos de España. I. Reino de Castilla*. Madrid: Centros de Estudios Históricos.

Müller, Bodo (2004): «Aspectos del léxico medieval desde la perspectiva del *Diccionario del español medieval* (*DEM*)», en Jens Lüdtke y Christian Schmitt (eds.), *Historia del léxico español. Enfoques y aplicaciones*. Madrid: Iberoamericana, 61–71.

Muñoz Garrigós, José (2008): *Las hablas murcianas. Trabajos de dialectología*. Edición y estudio de Mercedes Abad Merino. Murcia: Editum.

Nimer, Miguel (2005[2]): *Influências orientáis na língua portuguesa. Os vocábulos árabes, arabizados, persas e turcos*. São Paulo: Edusp.

NTLLE = Real Academia Española: *Nuevo Tesoro Lexicográfico de la Lengua Española*. <http://www.rae.es> [último acceso: 07/09/2017].

Pérez Martín, Antonio (2002): «El *Fuero Juzgo*, código de leyes del Reino de Murcia», en *El Fuero Juzgo. Estudios críticos y transcripción*. Murcia: Fundación Séneca, 41–73.

Puche Lorenzo, Miguel Ángel (2010): «Nombrar el agua en la Edad Media. Del preciado líquido al líquido mortal», *Cuadernos del CEMYR* 18, 105–117.

Puche Lorenzo, Miguel Ángel (2012): «Léxico de la vida cotidiana en la Murcia áurea», *Cuadernos del Instituto Historia de la Lengua* 7, 347–363.

Puche Lorenzo, Miguel Ángel (2015): «¿Dialectalismo y/o tecnicismo? Una mirada al léxico especializado de la minería del siglo XIX», *Etudes Romanes de Brno* 36/1, 103–117.

Quilis Merín, Mercedes (2010): «Fronteras y periodización en el español de Orígenes», en Mónica Castillo Lluch y Mónica López Izquierdo (eds.), *Modelos latinos en la Castilla medieval*. Madrid: Iberoamericana, 43–62.

Sánchez González de Herrero, Mª Nieves (dir.) (2000): *Diccionario español de documentos alfonsíes*. Madrid: Arco Libros.

Sempere Martínez, Juan Antonio (1995): «Apunts d'ictionímia catalana a Múrcia», *Estudis de llengua i literatura catalanes XXXI. Miscel.lània Germà Colón* 4, 259–276.

Serrano-Niza, Dolores (2007): «Arabismos relacionados con el léxico de la seda», *Revista de Filología de la Universidad de La Laguna* 25, 559–566.

Torres Fontes, Juan (1956): «La hacienda concejil de Murcia en el siglo XIV», *Anuario de Historia del Derecho Español* XXVI, 741–756.

Torres Fontes, Juan (1975): *El regadío murciano en la primera mitad del siglo XIV*. Murcia.

Torres Fontes, Juan (1978): «Los corredores del comercio murciano en el reinado de Alfonso XI», *Miscelánea medieval murciana* IV, 239–262.

Torres Fontes, Juan (1983): «Las ordenaciones al almotacén murciano en la primera mitad del siglo XIV», *Miscelánea medieval murciana* X, 71–131.

Torres Fontes, Juan (1985): «Ordenaciones para la guarda de la huerta de Murcia (1305–1347) y ordenanzas para la guarda del campo (siglo XV)», *Miscelánea medieval murciana* XII, 241–274.

Trias y Teixidor, Anna (1996): «El català en el llatì del "Regimen sanitatis ad regem aragonum" d'Arnau de Vilanova», *Estudis de llengua i literatura catalanes XXXII. Miscel.lània Germà Colón* 5, 34–50.

M.ª Ángeles Blanco Izquierdo, Gloria Clavería Nadal y
Enrique Jiménez Ríos

Fuentes lexicográficas y estudio del léxico: el *Diccionario de la lengua castellana* de la Real Academia Española (1817–1852)[1]

Resumen: El proyecto de investigación «Historia interna del *Diccionario de la lengua castellana* de la Real Academia Española en el siglo XIX (1817–1852)» tiene como objetivo fundamental el establecimiento de la enmienda llevada a cabo en cada una de las seis ediciones del diccionario publicadas entre 1817 y 1852 por medio de la identificación de los cambios introducidos en cada una de ellas y en el marco de las humanidades digitales. El objetivo último es averiguar la significación lexicográfica y lexicológica de estas transformaciones, y reconstruir la historia interna del diccionario académico. En esta contribución se exponen los objetivos del proyecto de investigación, sus fundamentos metodológicos, los resultados obtenidos y las perspectivas futuras.

Palabras clave: Lexicografía, Diccionarios académicos, Siglo XIX, Real Academia Española

Abstract: The research project «Internal history of the *Diccionario de la lengua castellana* of the Real Academia Española in the nineteenth century (1817–1852)» aims to establish the amendment performed in the six editions of the dictionary published between 1817 and 1852 by identifying the changes made in each, and also within the framework of the digital humanities. The main goal is to find out the lexicographic and lexicological significance of the changes and to reconstruct the internal history of the academic dictionary. The study exposes the research goals, its methodological foundations, the results and future prospects.

Keywords: Lexicography, Academic dictionaries, Nineteenth century, Real Academia Española

1 Introducción

Las veintitrés ediciones del *Diccionario de la lengua española* de la Real Academia Española, además del *Diccionario de autoridades*, constituyen un excelente

1 Esta investigación ha podido desarrollarse gracias a las ayudas de la DGICYT (FFI2014-51904-P) y al apoyo del Comissionat per Universitats i Recerca de la Generalitat de Catalunya (SGR2017-1251).

corpus para el estudio de la lengua española desde diversas perspectivas. Son una parte esencial de la lexicografía en esta lengua, pues a través de ellas es posible reconstruir una porción muy importante de su historia y, además, conocer el devenir del léxico español desde el siglo XVIII. Con este planteamiento, el objetivo del proyecto «Historia interna del *Diccionario de la lengua castellana* de la Real Academia Española en el siglo XIX (1817–1852)» (FFI2014-51904-P) es investigar las ediciones de la primera mitad del siglo XIX (de la quinta a la décima ediciones), período seleccionado por los motivos que se expondrán más adelante.

1.1 Antecedentes

Tal como señala Blecua (2018), los tres siglos de existencia del diccionario de la Real Academia Española constituyen un «fuerte atractivo intelectual» para los lingüistas. Ello puede observarse ya en el trabajo pionero de Lázaro Carreter (1972 [1980]) sobre el *Diccionario de autoridades*, al que siguen otros muchos como los de M. Alvar, M. Alvar Ezquerra, P. Álvarez de Miranda, D. Azorín o M. Seco. Los estudios de las distintas ediciones del diccionario académico se han multiplicado en las últimas décadas del siglo XX y los inicios del XXI. Estos se han centrado muy especialmente en el *Diccionario de autoridades* y los avatares del diccionario en el siglo XVIII, mientras que a la lexicografía académica del siglo XIX se le ha dispensado una menor atención.

Desde los años noventa del pasado siglo, la difusión de estas obras ha recibido el impulso de la propia corporación con el desarrollo de herramientas para facilitar su consulta: el *NTLLE*, que ofrece lematizados todos los diccionarios académicos, desde *Autoridades* a la vigésima primera edición; el *Mapa de diccionarios*, que recoge seis ediciones representativas (las publicadas en 1780, 1817, 1884, 1925, 1992 y 2001)[2], y una aplicación para la consulta del *Diccionario de autoridades*. A ello hay que unir, desde fecha reciente, la *Biblioteca Virtual de la Filología Española*, dirigida por M. Alvar Ezquerra, amplio repositorio de obras de carácter lingüístico.

1.2 Marco de la investigación

Esta investigación, que se enmarca en la historiografía lingüística y, en particular, en la historia de la lexicografía española, tiene como objetivo primordial analizar las ediciones del diccionario publicadas por la Real Academia Española en

2 Se citan las ediciones del diccionario académico a través del *NTLLE*.

la primera mitad del siglo XIX, fijándose en la evolución de la técnica empleada para su confección; se toma, además, el diccionario como base de estudio del léxico y las sucesivas ediciones del repertorio como una vía de acceso al conocimiento de su historia desde la concepción normativa que tiene la Academia.

No puede concebirse una investigación moderna sin el apoyo de las nuevas tecnologías y de los productos que ellas suministran. El *NTLLE* y otros recursos digitales constituyen un importante apoyo y, como se recoge en el epígrafe 3.4, los resultados del proyecto también habrán de formar parte de lo que hoy se entiende por Humanidades digitales.

2 Objetivos, características y metodología

El objetivo del proyecto es la reconstrucción del modelo de revisión y aumento que se aplicó en cada una de las ediciones del período señalado por medio de la identificación de los cambios introducidos y de la interpretación de su significado; la idea es trazar la «historia interna» del diccionario, siguiendo el planteamiento de Clavería (2016), con un análisis pormenorizado de cada una de las ediciones.

2.1 Historia interna del diccionario

La historia interna del diccionario pretende reconstruir los principios en los que se basó la elaboración de cada una de las ediciones y analizar de manera exhaustiva las modificaciones introducidas en ellas, de lo que resulta el conocimiento de la historia de la lexicografía académica en la etapa analizada y la historia de la recepción del léxico dentro del modelo de la lexicografía académica, con el consiguiente establecimiento de sus bases metodológicas y lexicológicas.

El diccionario se erige de esta forma en fuente de estudio del léxico desde la perspectiva de su recepción en las obras lexicográficas regidas por criterios normativos. Este uso del diccionario permite la observación de la descripción lexicográfica de cada palabra hecha por los académicos de la época. No se olvida, sin embargo, que la elección del diccionario como instrumento de análisis del léxico hace necesaria, para un enjuiciamiento comprensivo, su confrontación con el uso que reflejan los textos.

Partimos del convencimiento de que la reconstrucción de la historia interna del diccionario es la única manera de conocerlo en profundidad y de llegar a comprender sus principios. Para ello se ha decidido estudiar la lexicografía académica del siglo XIX como forma de profundizar en una etapa de la historia de la lexicografía y de la lengua en la que aún queda bastante por investigar.

2.2 Metodología de análisis

Atendiendo a la propia historia lexicográfica académica, se ha dividido el siglo XIX en dos fases (1817–1852 y 1869–1899) y ahora se acomete la primera. Se toma como punto de partida del proyecto la quinta edición, publicada en enero de 1817. La razón se encuentra en los cambios de cierto calado que esta entraña con respecto a la edición inmediatamente anterior, publicada en los primeros años de la centuria (*DRAE* 1803, *cfr*. Clavería 2018). El primer periodo analizado concluye en la décima edición (*DRAE* 1852) porque se considera que las ediciones siguientes (*DRAE* 1869, *DRAE* 1884 y *DRAE* 1899) responden a unas directrices lexicográficas distintas; estas últimas serán motivo de atención en una fase posterior.

La metodología de trabajo se ha fundamentado en el estudio comparativo de cada edición con respecto a la inmediatamente anterior, puesto que este era el método de trabajo que los académicos aplicaban en cada nueva impresión del diccionario.

El cotejo se lleva a cabo de dos formas distintas. En primer lugar y con el auxilio del *NTLLE* se identifican las diferencias en la macroestructura de la edición estudiada con respecto a la precedente; por ejemplo, en el caso de la quinta edición (*DRAE* 1817) se ha tomado como punto de comparación el *DRAE* 1803. El contraste de la macroestructura se realiza a través de los listados de formas[3] de cada edición obtenidos con el *NTLLE* y para la comparación se emplea el programa *COMPADRAE*, una herramienta creada en este proyecto para identificar las formas coincidentes y las distintas. En estas últimas es necesario discriminar los simples cambios formales del propio aumento. Las transformaciones en la forma pueden ser debidas a distintos motivos: uno de los mejores ejemplos se encuentra en la quinta edición del diccionario, en la que se aplicaron varias reformas ortográficas (véase el epígrafe 3.1.1); estos cambios modificaron sustancialmente la macroestructura de esta edición. También provocaron cambios involuntarios en la macroestructura las erratas de los lemas. Dentro de las formas y los lemas añadidos en cada una de las ediciones, algunos son verdaderas adiciones, otros podrían resultar de modificaciones en los criterios de lematización; por ejemplo, el lema *mediacaña* (*DRAE* 1817) figura ya en el *DRAE* 1803 como

3 Los diccionarios académicos pueden tener varias formas dentro de una misma entrada, por ejemplo *agujica, lla, ta* (*DRAE* 1817); del mismo modo *anémona* y *anémone* se incluyen en un mismo lema a partir del *DRAE* 1843. El *NTLLE* permite recuperar las formas de los lemas múltiples individualmente, aunque no son raros los errores de recuperación.

sublema de la entrada *medio*; *con que* se encuentra como lema diferenciado de *conque* entre 1837 y 1852 como 'partícula ilativa de interrogación', aunque esta información ya aparecía desde el *Diccionario de autoridades* en el lema *con*. En fin, las formas y lemas añadidos reflejan, además, la ampliación léxica registrada en el diccionario, fruto de la propia evolución del léxico y también del trabajo lexicográfico académico.

El cotejo de la microestructura de cada edición del diccionario con la inmediatamente anterior lleva al reconocimiento de otras modificaciones. En este caso, la comparación se realiza manualmente. Se observan transformaciones en las marcas, en la redacción de las definiciones, en la adición de nuevas acepciones y de nuevas estructuras complejas; unos cambios son eminentemente lexicográficos y otros tienen significación lexicológica.

Se persigue con todo ello tener una base de análisis suficientemente amplia para poder identificar y reconstruir las directrices lexicográficas y lexicológicas que se siguieron en la enmienda de cada edición. Interesa ante todo el tratamiento del diccionario desde el punto de vista filológico con la determinación de cuáles son las fuentes lexicográficas y textuales que subyacen a la ampliación y a los cambios introducidos. En el caso de las fuentes lexicográficas, cómo se realiza la selección y el aprovechamiento y, para las fuentes de carácter textual, cómo se trasvasa la información de los textos al diccionario. Se ha podido rastrear en algunos casos el origen de la información, pero se trata de un trabajo difícil y costoso (Clavería/Freixas/Torruella, en prensa; Clavería/Paz, en prensa).

Solo el análisis detallado permitirá una reconstrucción del método lexicográfico que dio como resultado los cambios que afloran a través de las páginas de cada una de las ediciones del diccionario.

2.3 Trabajo en equipo y colaboraciones externas

Un proyecto de esta envergadura precisa de un trabajo de equipo y realmente colaborativo. Buena muestra de ello son algunas de nuestras publicaciones (Azorín *et al.* 2017), las jornadas científicas celebradas en la UAB, de las que han resultado aportaciones de mucha utilidad para el conocimiento de las ediciones y el desarrollo del proyecto, y la asistencia a distintos congresos.

Pese a que el equipo de investigación está formado por un buen número de especialistas en historia de la lexicografía, ha sido necesario solicitar la colaboración de investigadores expertos en ciertas áreas del léxico con el fin de profundizar en determinados aspectos de las distintas ediciones del diccionario. Por ejemplo, en el estudio de la quinta edición, se ha recurrido a la colaboración de Gómez de Enterría (2018) para la interpretación de las voces relacionadas con

la historia natural, de Carriazo (2018) para las voces pertenecientes a la náutica, de Garriga (2018) para las voces de la química y de Hoyos (2018) para las de la economía.

3 Resultados y perspectivas futuras

3.1 El diccionario y su relación con otras obras académicas

A lo largo de la historia de la Real Academia Española, los trabajos lexicográficos se superponen a los trabajos ortográficos y gramaticales, necesitados unos de otros. Así, desde la fundación de la corporación hasta que ve la luz el diccionario con el que se inicia nuestro periodo de estudio, el *DRAE* 1817, se publican, además del *Diccionario de autoridades*, cuatro ediciones del diccionario usual (1780, 1783, 1791 y 1803), ocho de la ortografía (1741, 1754, 1763, 1770, 1775, 1779, 1792 y 1815) y cuatro de la gramática (1771, 1772, 1781 y 1796). Por ello, este proyecto plantea el análisis del diccionario no como obra aislada, sino como parte de un plan integral de la descripción y regulación del idioma (Sarmiento 2001).

Las ideas ortográficas y gramaticales se plasman en el diccionario y este, a su vez, las difunde, siendo incluso, en ocasiones, el primer receptor de la doctrina. Uno de nuestros objetivos es comprobar si la evolución de los principios teóricos se manifiesta de forma homogénea en el diccionario, la gramática y la ortografía en el periodo que comprende el proyecto. Además, la investigación se dirige a evaluar el grado de autosuficiencia de las obras académicas, entendida esta como la presencia en el diccionario de las voces técnicas que se utilizan en la ortografía y la gramática, y que el propio diccionario utiliza a veces en sus definiciones.

3.1.1 La ortografía y los cambios ortográficos en el diccionario

En los trabajos iniciales de cotejo entre la cuarta y la quinta ediciones del *DRAE*, se detectaron innumerables enmiendas relacionadas con la ortografía, las cuales atañían a dos aspectos: por una parte, a la ortografía con la que aparecen fijadas las voces en la nomenclatura; por otra, a las entradas relativas a las letras del abecedario español. Los cambios, en ambos casos, se corresponden con lo fijado en la octava edición de la *Ortografía*, de 1815, publicada solo dos años antes que el diccionario.

Así, en el ámbito gráfico, se aplican de forma sistemática las innovaciones ortográficas de 1815 (Terrón 2018), entre las que destacan las relativas a la distribución de los usos de *i* e *y* (*aceyte* > *aceite*), la eliminación de la grafía *x* para representar la consonante velar /x/ (*axedrez* > *ajedrez*) o la prescripción del uso

de la secuencia *qu* solo como dígrafo (*equador* > *ecuador*), lo que, además, afecta al uso de algunos diacríticos (*exôrbitante* > *exorbitante; seqüencia* > *secuencia*).

Por su parte, en la definición de las letras, se reproduce literalmente la parte inicial de los epígrafes dedicados a ellas en la *Ortografía*: información sobre la letra, sobre el sonido que representa (con la presencia de detalladas descripciones articulatorias) e información ortográfica. La ortografía es en este caso la fuente directa de las definiciones del diccionario (Blanco 2018c).

Dentro de los límites cronológicos de este proyecto, se publicaron la *Ortografía* de 1820, reimpresa en 1826, y tres ediciones del *Prontuario de ortografía de la lengua castellana,* de 1844, 1845 y 1850. La investigación debe centrarse ahora en comprobar si los cambios ortográficos de las ediciones del diccionario que se publicaron entre 1822 y 1852 van parejos a los textos ortográficos. A falta todavía de una investigación exhaustiva, puede afirmarse que, en este periodo, es el diccionario el que avanza las novedades ortográficas —de menor calado que las anteriores, pero todavía frecuentes—. Así, la limitación del uso de *g* a las palabras cuyo étimo presenta este grafema (*gerarquía* > *jerarquía*) y la supresión de la grafía *x* con valor de /x/ en posición final (*relox* > *reloj*) quedarán descritas y aplicadas, aunque no siempre de forma coherente en la macroestructura y la microestructura, en el *DRAE* 1837 y en el *DRAE* 1832, respectivamente. Ambas ediciones se anticipan, pues, unos años a la ortografía que expone estas reglas, el prontuario de 1844, que constituye la publicación ortográfica más relevante de la Academia desde 1815. El análisis de los vaivenes gráficos, sobre todo en la elección de *g* o *j* (*gefe/jefe, sugeto/sujeto)*, y de otros cambios que no son de carácter general, sino que afectan a voces aisladas (*zizañero* > *cizañero; móbil* > *móvil*), puede ayudar a trazar la historia de la fijación ortográfica del léxico.

Fuera del análisis de las grafías, queda por estudiar la relación entre la ortografía y el diccionario en lo relativo a la acentuación y a las mayúsculas, aspecto este último que solo es posible considerar en la microestructura, dado que los lemas suelen ser nombres comunes[4].

3.1.2 La gramática y el diccionario

Desde el inicio de las reuniones académicas, la gramática es objeto de debate, tratándose cuestiones fundamentales como el establecimiento de sus partes, el número de clases de palabras, la existencia de declinación en español, etc.

4 Se ha detectado, por ejemplo, una tendencia a la adición de mayúscula en el *DRAE* 1852, tanto en la escritura de los meses del año (*enero* > *Enero, mayo* > *Mayo*) como en la de tratamientos o cargos (*reyes* > *Reyes; san* > *San; papa* > *Papa*).

(Esteve 1982; Martínez Alcalde 2010; Gómez Asencio 2011). Todo ello, de una u otra forma, había ido teniendo su reflejo en el diccionario, de manera evidente en las definiciones de las voces técnicas de la lingüística. No obstante, la influencia de la *Gramática* en los diccionarios analizados en nuestro proyecto es necesariamente menor, puesto que no se publican ediciones de la obra entre 1817 y 1852. El texto vigente cuando ve la luz el *DRAE* 1817 es la cuarta edición de la *GRAE*, de 1796 —se interpone, por tanto, entre ambas obras el *DRAE* 1803—, y habrá que esperar hasta 1854 para que aparezca una nueva edición.

El análisis de las relaciones entre gramática y diccionario se ha detenido hasta el momento en comprobar si el texto gramatical de 1796 y el ortográfico de 1815 están en el origen de los cambios experimentados por las voces de la gramática[5] en la quinta edición del *DRAE,* así como en valorar el grado de armonización entre las ediciones vigentes de los tres códigos académicos (Blanco 2018b). Pese a detectarse un alto grado de coherencia, se quebranta en algunos casos el principio de autosuficiencia en el propio diccionario, y entre este y los textos ortográfico y gramatical[6]. A partir de ahora, la investigación debe centrarse, por una parte, en hacer un seguimiento de la evolución de la terminología gramatical identificada en 1817 y del aumento o supresión en este campo. Por otra parte, habrá que verificar si hay cambios sobre aspectos gramaticales concretos en las ediciones publicadas entre 1817–1852 y si estos tienen repercusión en la gramática de 1854.

3.1.3 Otras obras

Además de la elaboración del diccionario, la ortografía y la gramática, entre los objetivos fundacionales de la Academia se encuentra la divulgación de obras literarias. Obras de publicación cercana a nuestro periodo de estudio son la edición del *Quijote* de 1780, realizada en la imprenta Ibarra, y la del *Fuero juzgo* de 1815. Al rastrear las fuentes textuales de las que beben las entradas del diccionario, habrá de considerarse la relevancia que pudieran tener estas publicaciones en

5 De acuerdo con la clasificación expuesta en la *GRAE* 1796, la disciplina gramatical se divide en cuatro partes: ortografía, analogía, sintaxis y prosodia. Por tanto, nos referimos aquí también a la terminología ortográfica e incluso fonética, dado que los contenidos de esta disciplina actual se reparten en la época entre la ortografía y la prosodia.

6 Por poner un ejemplo: de los ocho tipos de conjunciones definidas en la *GRAE* (1796: 263), el diccionario registra únicamente como voz técnica *adversativo*, mientras que *final, comparativo, causal* y *condicional* solo presentan acepciones generales, y *disyuntivo* y *continuativo* no tienen entrada.

la selección de entradas —especialmente en el caso de los arcaísmos— y en la redacción de definiciones o ejemplos de uso.

3.2 Punto de partida: el *DRAE* 1817

La quinta edición del diccionario marca el inicio del proyecto por confluir en ella al menos dos aspectos que la hacen singular: mientras que las ediciones anteriores (*DRAE* 1780–1803) se elaboraron a la sombra de los trabajos de revisión de la segunda edición del *Diccionario de autoridades*, el *DRAE* 1817 supuso el principio de una ruptura con esta metodología de trabajo, pues, aunque su elaboración se inició poco después de la publicación del *DRAE* 1803, la parte más importante del proceso de revisión y aumento se llevó a cabo a partir de 1814, después de los graves acontecimientos políticos que tuvieron lugar entre 1808 y 1814 (Clavería 2018). La singularidad de la quinta edición se encuentra también en los cambios ortográficos que se verifican en ella (epígrafe 3.1.1).

Aunque en el prólogo de la quinta edición la Academia calificó el aumento de *considerable* (epígrafe 3.3.1), este fue menor al de la edición precedente (Clavería 2016); pese a ello las reformas incorporadas en el *DRAE* 1817 no fueron nada despreciables. Entre ellas destacan, aparte de los cambios ortográficos (Blanco 2018a y 2018c; Terrón 2018), las relativas al tratamiento lexicográfico de las formas complejas, cuya ubicación cambió siguiendo la aplicación de criterios gramaticales (Buenafuentes 2018; Paz 2018); la introducción de algunas reformas en las voces con marcación diacrónica y las abreviaturas utilizadas en este sector del léxico (Jiménez Ríos 2018); la variación en las marcas de carácter diafásico (Azorín 2018); la intensa revisión de las correspondencias latinas (Jiménez Ríos/Clavería 2018) y el amplio repaso de las definiciones (Freixas 2018), relativamente importante en ciertas esferas como la historia natural (Gómez de Enterría 2018) o la química (Garriga 2018); por último, se atendió también a la relación entre remisión y definición (Muñoz 2018).

3.3 La evolución de la técnica lexicográfica: de la macroestructura a la microestructura (1817–1852)

3.3.1 Cambios en la nomenclatura

Los cambios verificados en la nomenclatura se materializan tanto con la adición de nuevas entradas como con la supresión de lemas existentes en la edición anterior. La identificación del aumento en todas las ediciones se ha efectuado a través del *NTLLE* (véase el epígrafe 3.4.1). Los datos numéricos del incremento de lemas y formas, que ofrece Clavería (2017), son los siguientes:

Tabla 1: Aumento de lemas y formas

Edición	Lemas	Formas
1817 (5.ª)	1292	1531
1822 (6.ª)	467	513
1832 (7.ª)	335	387
1837 (8.ª)	272	278
1843 (9.ª)	540	544
1852 (10.ª)	678	749

Destaca sobre las demás la quinta edición (*cfr*. Clavería/Freixas 2015: 1306), cuyo incremento de lemas y formas es considerablemente superior al de ediciones posteriores. Desde la sexta a la octava edición, la disminución del número de adiciones es notable, mientras que en las dos últimas ediciones estudiadas se produce un incremento progresivo de lemas y formas, aunque nunca equiparable al de 1817. Se explican esas cifras al considerar la variedad de ámbitos en los que se producen aumentos, entre los que destaca en el *DRAE* 1817 el de historia natural. Era la consecuencia de que, ya en el siglo XVIII, a raíz de la traducción de textos y del contacto con la Ilustración europea, se sintiera el deseo de ampliar el saber con conocimientos de áreas más específicas (botánica, farmacología, medicina y cirugía, zoología, mineralogía, etc.). A este interés por el conocimiento se unía el deseo de su divulgación, para lo que el diccionario se convirtió en un importante órgano difusor (Gómez de Enterría 2018).

El papel desempeñado por el diccionario en la difusión de las novedades léxicas explica la inserción de este léxico técnico —abundante en esta edición, pero no exclusivo de ella— y de otro perteneciente a otras áreas. Se incorporan otras voces de especialidad, particularmente náutica y química, y esta incorporación está determinada por la concepción que se tenía entonces del léxico de especialidad, del modo como se establecían las áreas (Battaner 1996: 101), y de los criterios utilizados en la aceptación del léxico en el diccionario, así como en la asignación de marcas, pues unas voces están marcadas y otras, siendo técnicas, aparecen sin marcar (Torruella/Huertas 2018)[7].

7 En el *DRAE* 1817 la incorporación de voces y acepciones con marca diatécnica es limitada, si bien la presencia de léxico de especialidad es notable, pues puede aparecer con marca o sin ella: es el caso, por ejemplo, de *economía política* y *edicto pretorio,* pertenecientes a la economía y al derecho respectivamente, pero recogidas en el diccionario sin marca (Paz 2018; Hoyos 2018). Esta inserción de voces técnicas sin marca se observa también en el *DRAE* 1843 (Clavería 2016: 123).

Si esto sucede en la primera edición examinada, en las siguientes hay también inserción de tecnicismos: de medicina en el *DRAE* 1822, *DRAE* 1843 y *DRAE* 1852; voces forenses en el *DRAE* 1843, y voces del comercio, como las que ya aparecían en el *DRAE* 1817, en el *DRAE* 1852. Al lado de todos estos tecnicismos, términos como *absolutista, extradición, insurrección*, resultado de los cambios sociales y políticos producidos en las primeras décadas del siglo XIX, son introducidos en el *DRAE* 1843.

Ciertamente, el léxico técnico sobresale en el aumento de la nomenclatura, pero no es el único que lo conforma. Ligado al desarrollo social que se acaba de mencionar, en el *DRAE* 1852 se incorporan voces relacionadas con cargos públicos como consecuencia del surgimiento de nuevas instituciones políticas, como se indica en el prólogo de esta edición (*cfr.* Córdoba/Terrón 2017). Y a este léxico se unen otras incorporaciones: arcaísmos, por ejemplo en la letra *p–* se incorporan en el *DRAE* 1817 quince lemas con marca *ant.* vinculados probablemente a la revisión del *Diccionario de autoridades* (Clavería 2018); en esa misma edición, léxico dialectal (Varela 2018; Julià 2018), con poca presencia en el *DRAE* 1822, y voces con marcación diafásica y estilística (Azorín 2018); a ellas se unen las de léxico familiar y coloquial (*farolear*, *peripuesto*, y estructuras fraseológicas como *hacer a uno el caldo gordo* o *poner los pies en polvorosa* en el *DRAE* 1822, *chiripa* o *paparrucha* en el *DRAE* 1832, y *miedoso* o *potingue* en el *DRAE* 1843). Estas incorporaciones resultan de la concepción moderna que tiene la Academia de la práctica lexicográfica tendente a registrar los usos generalizados en la lengua, lo que explica la adición en el *DRAE* 1837 de préstamos, incluso sin adaptar al español (*club, fagot, frac*), y de formaciones resultantes de la creación de palabras (*neutralizar, fanatizar, impermeable, intachable, negociable*)[8].

La pertenencia de estas incorporaciones a distintos ámbitos redunda en que muchas de ellas presenten remisión y no definición; de hecho, una cuarta parte de los lemas incorporados tiene remisión, lo que muestra que no se añaden solo nuevos significados, sino significantes con alguna peculiaridad, la que los adscribe a los grupos que se acaban de señalar (Clavería 2018).

Este aumento de lemas en el diccionario refleja el modo de proceder de los académicos revisores al concebir cada edición como una obra suministradora de novedades léxicas, pese a que en estas ediciones la ampliación es ciertamente limitada. Al disponer de la información, la ponen a disposición del usuario inmediatamente en la edición que preparan, recurriendo para ello incluso al suplemento, algo que singulariza al *DRAE* 1837, pues el 60 % de las adiciones

8 Todos estos ejemplos están tomados de Clavería (2016: 94, 106, 112 y 123).

aparece en él (Clavería 2016: 111 y Clavería 2018), por lo que quedan por esclarecer las circunstancias que envuelven la revisión realizada en esa edición.

Todo lo señalado hasta aquí viene a confirmar lo que, en los albores de la lexicografía académica, en la realización del *Diccionario de autoridades*, los estudiosos han destacado como seña de identidad de la corporación, distinguidora de la lexicografía practicada en otros idiomas: la apertura a registrar el léxico de la lengua en toda su variedad. Las ediciones aparecidas en el siglo XIX, por su contacto con la realidad, son buen testimonio de ello.

Ese contacto con la realidad y la propia evolución de la lexicografía académica y de la idea de diccionario hacen que, al lado de las incorporaciones, haya supresiones: la supresión es reflejo de la corrección del diccionario y es muy importante en la sexta y séptima ediciones. En el *DRAE* 1822 se eliminan, principalmente, lemas marcados diacrónicamente y estos superan las dos mil entradas; sobrepasan las seis mil entradas las supresiones del *DRAE* 1832 porque en ellas se encuentran los participios pasivos. Actualmente se están identificando las supresiones de cada una de las ediciones estudiadas con el fin de poder averiguar las razones de un tipo de revisión hasta ahora desatendida. Muestra esta incorporación y supresión de lemas del diccionario la evolución de la lengua con el paso del tiempo, de un tiempo acotado en este proyecto de investigación a la primera mitad del siglo XIX; y muestra también el aprovechamiento del diccionario para conocerla, pues los cambios aquí señalados, más allá de reflejar un modo de practicar la lexicografía, están determinados por los que se producen en la lengua y en la sociedad en ese momento[9]. Actualmente, en esta fase del proyecto, se está investigando la significación tanto del aumento como de las supresiones en la concepción del diccionario.

3.3.2 El encabezamiento de los artículos

El encabezamiento de los artículos es fuente de información no solo léxica, sino también morfológica (moción de género en las categorías que presentan esta variación, *pluralia tantum,* forma pronominal de los verbos, etc.). Los cambios en estos aspectos suelen responder a los criterios lexicográficos aplicados en la edición, pero en algún caso pueden dar pistas sobre cambios en el uso, de ahí que sea interesante señalarlos. Por otra parte, los errores en el lema, frecuentes

9 Se incorporan en el diccionario voces como *vacuna, vacunación* y *vacunar*, fruto del avance de la medicina a finales del siglo XVIII y se suprimen otras, nombres de plantas como *ispida* u *otona*, resultado de la revisión de esta parcela léxica (Clavería 2016: 69 y 87).

en la quinta edición por la celeridad con que se hizo y los abundantes cambios ortográficos que presenta, suelen corregirse en ediciones posteriores, aunque a veces se mantienen durante varias ediciones y se repiten en algunos diccionarios no académicos (Clavería/Freixas 2018b y Freixas/Clavería 2018), lo que permite investigar el *DRAE* como fuente de estos otros diccionarios (*centidonia* por *centinodia*, *cfr*. Raab 2018).

A la hora de estudiar el encabezamiento de los artículos, se ha establecido la distinción entre *lema* y *sublema,* pertinente en el análisis de la mayoría de los diccionarios desde *Autoridades* hasta el *DRAE* 1822[10]. En estas ediciones, el lema que encabeza el artículo, escrito en versales, se repite en versalita y párrafo aparte en cada nueva acepción, constituyendo sublemas. También son sublemas del lema principal las formas complejas —sean compuestos sintagmáticos, sean unidades fraseológicas—, igualmente consignadas en versalita. Esta configuración de las entradas, que se conserva de manera casi sistemática en cinco de las seis primeras ediciones del diccionario, se modifica en la séptima; en ella las acepciones y las formas complejas se agrupan en un solo párrafo y desaparecen los sublemas que preceden a las primeras. Esta disposición se mantiene hasta hoy en las ediciones en papel.

Desde 1832, pues, todas las acepciones de una voz aparecen bajo un solo encabezamiento. No se considera en ese momento relevante si para algunas acepciones cambia el étimo del lema o su categoría gramatical (*cantar* verbo y sustantivo, por ejemplo, tendrán una sola entrada); por lo tanto, no se plantea aún el problema de la homonimia o la polisemia en el diccionario. Los homónimos solo recibirán encabezamientos independientes a partir del *DRAE* 1884, justo cuando se reincorporan las etimologías al diccionario, aunque sin el superíndice característico en la lexicografía actual, que se añade en el *DRAE* 1970.

Entre los cambios metalexicográficos relevantes se encuentra asimismo la reordenación de las formas complejas —que tiene lugar en el *DRAE* 1817—, basada en criterios gramaticales que se han conservado hasta nuestros días y que evitan la subjetividad de ediciones anteriores (Buenafuentes 2018).

En el apartado del encabezamiento de los artículos, destaca la presencia de *lemas múltiples,* es decir, de lemas que incorporan varias formas. Estos ofrecen algunas variantes formales (*gradiolo ó gladiolo*) o morfológicas (*estajero ó estajista*) o, más frecuentemente, constituyen un inventario de los derivados

10 La única excepción es el *DRAE* 1791, en el que las acepciones se distribuyen, numeradas, en un solo párrafo.

apreciativos más comunes, como los conformados por los sufijos *-ico, -illo, -ito* y *-uelo* para el diminutivo (*ovejica, lla, ta*) o los terminados en *-azo, -ón* y *-ote* para el aumentativo (*animalón, te*). Se presenta, así, un catálogo de formas amplio, que puede ser punto de partida de futuras investigaciones. También puede serlo el hecho de que los diminutivos lexicalizados suelan presentarse en un lema independiente, y no como sublemas de las formas que conservan el valor apreciativo. El *DRAE* 1817 incrementa visiblemente los lemas múltiples con diminutivos, que no desaparecerán hasta 1884.

Las diferentes formas de los derivados apreciativos se presentan en el lema múltiple en un estricto orden alfabético, tal y como figuraban en la gramática de 1796 (Prat 2018). En cambio, el diccionario no es sistemático al consignar las variantes: en ocasiones aparecen como lema múltiple, pero lo más común es que se presenten en entradas independientes en el lugar que corresponde a cada una en el orden alfabético y, por lo general, enviando a la que se considera preferida. Dado su valor para orientar el uso, resulta de especial interés el análisis de los cambios en las preferencias de una edición a otra, así como la eliminación o adición de variantes.

3.3.3 Los cambios en la marcación

Los cambios en la marcación interesan por cuanto pueden constituirse en indicios de modificaciones de los criterios lingüísticos y lexicográficos que se verifican en el periodo estudiado. Hay marcas que caracterizan gramaticalmente el lema en sus diferentes acepciones y otras que informan de sus condiciones de uso.

Las primeras indican la categoría gramatical y alguna otra característica del lema o de una de sus acepciones, bien morfológica ([sustantivo] masculino, femenino, plural; [adjetivo] superlativo; etc.), bien sintáctica ([verbo] activo, pasivo; [conjunción] disyuntiva, distributiva; etc.). El inventario de estas marcas había quedado fijado en los diccionarios académicos del siglo XVIII, siendo la segunda edición del *Diccionario de autoridades* la que mayor aumento había aportado[11]. No hay novedades en las marcas gramaticales de la quinta edición, y las enmiendas registradas entre 1817 y 1852 no tienen relevancia teórica: unas veces se trata simplemente de adiciones o supresiones de letras en abreviaturas ya existentes; otras —lo que resulta más contraproducente desde el punto de vista de la técnica lexicográfica— se duplican abreviaturas (*d.* o *dim.* para diminutivo desde 1822;

11 De las dieciocho marcas del primer *Autoridades*, la segunda edición pasa a tener treinta y cuatro. En el *DRAE* 1803 se contabilizan cuarenta y cinco.

f. o *s. f.* para sustantivo femenino desde 1837; etc.). Se mantienen, además, las marcas con dos posibles equivalencias, como *p.* (plural o participio), *n.* (nombre o neutro), deficiencia que no se subsanará hasta el *DRAE* 1869.

Las muestras utilizadas en el cotejo entre las ediciones del diccionario no arrojan un número representativo de enmiendas en la marcación gramatical, si bien, como en tantos otros aspectos, se producen retoques, como la adición de marcas en acepciones que no las tenían. Resultan de especial interés los cambios en esta marcación cuando, en lugar de subsanar ausencias o deficiencias anteriores, pueden constituir el rastro de un cambio lingüístico o de un cambio en la orientación normativa de la Academia. Por ejemplo, en el *DRAE* 1852 *epigrama* pasa de ser ambiguo (marcado como *m. y f.*) a considerarse masculino (marcado como *m.*), aunque todavía se indica que es femenino en una nota de uso pospuesta a la definición («Se halla también usado como femenino»). Queda pendiente comprobar si la documentación que ofrecen los bancos de datos refrenda estas enmiendas.

Al lado de estas marcas gramaticales, la presencia de las de uso muestra el contacto del diccionario con la lengua y la sociedad que la habla. El *DRAE* 1817 destaca por la revisión de algunas marcas usadas en él. Dicha revisión resulta en simplificación, por ejemplo, en las cronológicas o de vigencia de uso y en las diafásicas[12]. Pero hay otros cambios, como la eliminación de alguna marca diatécnica[13] y la supresión de expresiones indicadoras del uso[14]. Desde entonces el número de abreviaturas experimenta cambios por aumento y supresión, e incluso por modificación.

Esta adición o supresión de marcas presenta diferencias en los ámbitos afectados por ellas: en las marcas de especialidad, el aumento y la selección dependen de las áreas que sean así consideradas[15], en lo que intervienen factores externos al diccionario, tales como el progreso en el conocimiento científico[16]; en el uso

12 Se eliminan las marcas *raro* y *poco usado* en beneficio de *anticuado*, y se simplifican *bajo, festivo, jocoso* y *vulgar* en *familiar*. La publicación posterior de otras ediciones hace que se incorporen otras marcas diafásicas (Azorín 2018). Así, en el *DRAE* 1822 se recupera la marca *vulgar* (Clavería 2016: 89, nota 71).

13 Se eliminan las marcas *Astrol.* y *Escol.*

14 *Otr. par.* se cambia por *En algunas partes.*

15 Las enmiendas de marcas de especialidad son poco significativas, aunque hay supresión en algún caso.

16 En la lista de abreviaturas de especialidad no se producen cambios importantes hasta 1822; en 1817 hay pocas modificaciones.

de marcas diafásicas y de estilo, su variedad resulta de la idea que se tiene en un determinado momento de la corrección lingüística[17].

Dicho esto, lema y acepciones se marcan, y en el paso de una edición a otra se producen cambios, consecuencia natural de la evolución de la lengua. En el *DRAE* 1822, por ejemplo, las marcas denotadoras de materias humanísticas se eliminan de lemas así caracterizados, y la supresión de la marca diatópica es ligeramente mayor que la adición, lo que ha de interpretarse como resultado de la generalización de un uso y de la constatación de esa generalización. A mediados del siglo XIX, en el inicio de lo que es un cambio de tendencia en la confección del diccionario, en el *DRAE* 1852, el deseo de rehabilitar arcaísmos lleva a quitar la marca a los que hasta entonces la tenían.

Se marca con abreviatura, pero también con explicación en la microestructura; y hay casos en que la definición informa del carácter restringido o especializado de un término (Clavería/Freixas 2015: 1311), lo que es poco sistemático, si se examina desde la perspectiva actual y se considera que la publicación de las distintas ediciones del diccionario comporta mejoras en el modo de presentar la información, con regularidad y homogeneidad[18].

Estas consideraciones sobre el uso de marcas, sobre su tipología y presencia en el diccionario muestran los cambios operados en ellas, pero han de servir, además, para señalar la conexión que las marcas establecen con otras informaciones contenidas en el diccionario, con el significado de las voces a las que acompañan, de manera particular, y para destacar su papel como suministradoras de una información necesaria para la adecuada comprensión y producción del léxico.

3.3.4 Cambios en las acepciones

Como otros cambios en el diccionario, los que afectan a las acepciones resultan de la adición, supresión y modificación. En la adición, se incorporan nuevos sentidos, preferentemente en voces familiares y metafóricas, y en las dialectales y técnicas. Ejemplo de ello son los siguientes:

17 Y de los cambios sociales, como muestra la marcación de las voces *bienestar* o *porvenir* (Azorín 2018).

18 Es observable este modo de marcación en el léxico dialectal en una proporción muy alta (Varela 2018).

Tabla 2: Adición de acepciones

Ámbito	Acepción
Lengua general	ELECTRIZAR. [2]. met. Exaltar, avivar el animo de alguno. *Animos vehementer accendere, inflammare* (*DRAE* 1817).
Dialectalismos	PAJARILLA. [2] s. f. *p. Ar.* Lo mismo que PALOMILLA por el insecto que destruye la cebada (*DRAE* 1817).
Tecnicismos	AOVADO, DA. adj. *Bot.* Se aplica á cualquiera parte de la planta, como la hoja, que siendo mas larga que ancha remata por la base en un segmento de círculo, y por la punta en otro mas estrecho. *Ovatus* (*DRAE* 1817S).

Es la consecuencia natural de lo expuesto a lo largo de estas páginas acerca de la peculiaridad del *DRAE* 1817, innovador en tantas cosas. A partir de ahí, las ediciones posteriores tratan de completar la tarea iniciada entonces, lo que explica que el *DRAE* 1822 destaque por la adición de acepciones y no de lemas, y lo mismo puede decirse del *DRAE* 1852: se incorporan acepciones y, si lo hacen voces con significados diferentes, se incorpora una acepción para cada una[19].

A su lado, la supresión de acepciones es el resultado no tanto de la eliminación de sentidos de una palabra cuanto de la enmienda, de la revisión que lleva a eliminar acepciones redundantes[20]: Freixas (2018) cita los casos de *valido* [5] y *valido* [6] del *DRAE* 1803, refundidos en una sola acepción, *valido* [6], en *DRAE* 1817[21]:

Tabla 3: Eliminación de acepciones redundantes

***DRAE* 1803**	***DRAE* 1817**
VALIDO. [5] s. m. El que tiene el primer lugar en la gracia de algun soberano, o es su primer ministro. Se usa tambien como sustantivo. VALIDO. [6] El que logra el primer favor, o estimación con qualquier particular.	VALIDO. [6] s. m. El que tiene el primer lugar en la gracia de algun soberano o en la estimación de cualquier particular.

19 A esa conclusión llegan Terrón/Torruella (2017) después del examen de una muestra de las dos primeras páginas de todas las letras del diccionario. Véase también Córdoba/Terrón (2017).

20 Terrón/Torruella (2017) señalan que las supresiones de voces con marca de especialidad son raras.

21 Es el mismo número de acepción por el cambio de orden de las acepciones existentes en el *DRAE* 1803.

Esta revisión, que en este caso desemboca en supresión, lleva a otro de los cambios experimentados por las acepciones: la división de una en dos. Si los dos primeros cambios, adición y supresión, resultan, como sucede en los lemas, de factores externos al diccionario (neología de sentido, en el primero, y caída en desuso, en el segundo), el tercer cambio, la revisión, se produce como consecuencia de un mejor conocimiento de la realidad lingüística, lo que comporta una mejora del diccionario, y esto sucede en voces de todo tipo, como estas que señala Freixas (2018):

Tabla 4: División y ampliación de acepciones

Ámbito	***DRAE* 1803**	***DRAE* 1817**
Voces de la lengua general	ACAPARRARSE. [1] v. r. ant. Guarecerse debaxo de la capa de alguno, y metafóricamente ponerse baxo su proteccion y amparo.	ACAPARRARSE. [1] v. r. ant. Guarecerse debajo de la capa de alguno. ACAPARRARSE. [2] met. Ponerse bajo la proteccion de alguno.
Voces con marcación diafásica y estilística	ZUMBIDO. s. m. El ruido, o sonido continuado y bronco, que hacen las cosas en el ayre. *Susurrus, raucus stridor, bombus*.	ZUMBIDO. s. m. El ruido, susurro ó sonido continuado y agudo que hacen las cosas en el aire. En el estilo familiar se usa tambien para significar el golpe ó porrazo que se da á alguno. *Susurrus, raucus stridor, bombus: ictus.*
Voces del ámbito dialectal	FAJO. s. m. ant. Lo mismo que HAZ, Ó ATADO	FAJO. [1] s. m. *p. Ar.* Lo mismo que HAZ. FAJO. [2] s. m. ant. Lo mismo que HAZ, Ó ATADO. FAJOS. p. El conjunto de ropa y paños con que se visten los recién nacidos. *Infantiles fasciae.*
Voces del ámbito técnico	EFERVESCENCIA. [1] s. f. Hervor excesivo de la sangre, o de algun otro líquido. *Effervescentia*.	EFERVESCENCIA. [2] s. f. Especie de ebullición espumosa y con cierto ruido ocasionada por el desprendimiento de algun cuerpo que estando mezclado o convinado con otro, se desprende de él en estado de gas, como cuando fermenta el mosto ó se echa ceniza en algun ácido. *Effervescentia*.

El último ejemplo citado no solo resulta de la mejora del diccionario, sino del mejor conocimiento de la ciencia, en este caso, de este proceso químico (García Belmar/Bertomeu 2006: 66 y ss., *apud* Freixas 2018)[22].

En la definición, y en estos cambios en la definición, consecuencia de la relación de unas acepciones con otras, se observa una mejora en la confección del diccionario —la evolución del método lexicográfico, el desarrollo de la técnica lexicográfica—, lo que repercute en las acepciones, con la sistematización de la información. Un modo de conseguir esta regularización es a través de la remisión; por ello son muchos los casos en que se incorporan acepciones que son remisiones (Muñoz 2018), lo que permite relacionar variantes (gráficas, morfológicas y léxicas): se ha dicho más arriba (epígrafe 3.3.1) que una parte importante de los lemas incorporados tienen remisión; hay que decir ahora que la mayor parte de las remisiones se encuentra en lemas con una única acepción (Muñoz 2018).

3.3.5 Los cambios en la definición

En la quinta edición del *DRAE* se produce una intensa revisión de las definiciones de 1803, especialmente en las voces de las ciencias naturales. En las ediciones posteriores de la primera mitad del XIX siguen registrándose enmiendas, casi siempre tendentes a la reducción y simplificación de las entradas. El volumen de cambios decrece de manera drástica a partir de la edición de 1822, lo que es especialmente evidente en el *DRAE* 1832. En ambas ediciones, sexta y séptima, se percibe, en general, una acusada tendencia a la definición sinonímica.

La definición, como otros aspectos del artículo lexicográfico, cambia a veces por motivos meramente prácticos, como el intento de reducir el número de páginas del diccionario, lo que es evidente en la séptima edición (*DRAE* 1832), en cuyo prólogo se expresa el deseo de «disminuir el volumen» para hacerlo «más manejable» (*cfr.* Clavería/Paz, en prensa). Otras veces, el objetivo del cambio es presentar modelos de definición más uniformes y precisos —lo que implica mejoras en el método lexicográfico—, para despojarla de información innecesaria o prescindible, precisar los significados, aportar nuevos matices semánticos, etc. Hay, por último, cambios de tipo meramente estilístico. A partir del cotejo de las muestras seleccionadas se ha establecido una tipología de cambios, que pueden agruparse en tres categorías:

22 Véanse otros ejemplos, como el de la voz *elemento*, explicado por Garriga (2018), y a partir de ahí los cambios experimentados por las voces *agua, aire, tierra* y *fuego.*

a) Ampliación de la definición, que suele consistir en la adición de elementos que delimitan el valor semántico de las voces o en el desarrollo de definiciones enciclopédicas (aunque, por lo general, hay una tendencia a la concisión).
b) Reducción de la definición, que se lleva a cabo mediante el acortamiento de la información enciclopédica o de las explicaciones, la eliminación de redundancias (generalmente estructuras bimembres cuyos elementos son sinónimos), de información complementaria (etimológica, ortográfica o gramatical, por ejemplo) y de algunas muletillas que iniciaban la definición (*especie de…, cierta…, cualquiera…*), y que otras veces se simplifican (*cualidad de lo que es… > cualidad de…*).
c) Reescritura de la definición, que puede consistir en la corrección de erratas, en el cambio de un tiempo verbal, en modificaciones de la puntuación, en la sustitución de una palabra por otra, en la reescritura de la definición completa para hacerla más concisa y adecuarla en mayor medida al significado del lema, en cambios entre remisión y definición, o en el desdoblamiento de una acepción, evitando así definiciones excesivamente abarcadoras, como se ha visto en el epígrafe 3.3.4.

Los procesos de reescritura son de tal interés que bien pueden constituir el origen de investigaciones futuras. Una de ellas sería valorar en qué medida la sustitución de palabras en la definición responde a una actualización de la lengua o de los contenidos de la definición de acuerdo con los avances científicos y la cultura contemporánea a los académicos. Un ejemplo pueden ser las sustituciones léxicas que se producen en la entrada *bermellón* del *DRAE* 1832 (*cfr.* Clavería/Paz, en prensa):

Tabla 5: Sustituciones léxicas en la definición

***DRAE* 1822**	***DRAE* 1832**
BERMELLON. s. m. Fósil que se compone de azufre y azogue…	BERMELLON. m. Mineral que resulta de una combinación natural del mercurio con el azúfre…

Por otra parte, en el ámbito de la definición se han considerado también dos tipos de elementos que la completan: las notas de uso y los ejemplos. Las primeras ofrecen información paradigmática (por ejemplo, sobre sinónimos y variantes) e información sintagmática, de tipo sintáctico-semántico (sobre la construcción,

el régimen, lo que hoy conocemos como contorno, la combinatoria de las palabras, etc.). Por su parte, los ejemplos constituyen muestras prototípicas de lengua que sirven de apoyo a la definición lexicográfica o a la información sintáctico-semántica proporcionada por las notas de uso. Notas y ejemplos, que incrementan el valor codificador del diccionario académico, son una fuente importante para el conocimiento de la lengua de la época.

En el proyecto se ha iniciado una línea de investigación que analiza este tipo de componentes del artículo lexicográfico y la evolución que experimentan en las ediciones del diccionario[23]. Los materiales disponibles hasta ahora nos han permitido estudiar la naturaleza de las notas de uso y de los ejemplos, que se han clasificado de acuerdo con su contenido. Por otra parte, su evolución permitirá valorar la atención que el uso recibe en cada edición, que, en principio y en consonancia con otros aspectos, decrece tras 1817 y empieza a dar síntomas de revitalización en el *DRAE* 1843.

Por último, al interpretar el valor de los cambios, queda por discernir, como en las definiciones, si las enmiendas son de carácter práctico o técnico, para hacer las notas y ejemplos más breves o más claros, o si reflejan la evolución de los usos lingüísticos, como el abandono o el surgimiento de una variante, de una construcción, una combinación, etc., lo que exigirá indagar en el paralelismo entre estos cambios y el uso documentado.

3.4 Humanidades digitales

No puede concebirse actualmente un proyecto de investigación del área de humanidades al margen del adjetivo *digital*. La feliz conjunción entre humanidades y nuevas tecnologías está presente en la misma concepción del proyecto, afecta a su gestión y desarrollo, y alcanza de manera muy particular a la difusión y visibilidad de sus resultados.

3.4.1 Soporte informático

Como ya se ha señalado anteriormente, en el punto de partida de nuestras investigaciones se encuentra el aprovechamiento de los recursos digitales disponibles, en especial el *NTLLE*. El uso continuado de esta herramienta en nuestro proyecto ha conducido a una reflexión profunda sobre sus posibilidades (Clavería/Freixas 2018b); igualmente se ha alcanzado un conocimiento detallado de sus entresijos

23 Los primeros resultados de esta investigación fueron presentados en el XXI Congreso de la Asociación Alemana de Hispanistas, celebrado en Múnich en 2017.

y puntos débiles. La experiencia y los resultados obtenidos se han incorporado a nuestro repositorio de datos (epígrafe 3.4.2).

También en la fase preliminar de nuestro proyecto se ha desarrollado un sencillo programa informático con el fin de facilitar la comparación de la macroestructura entre las distintas ediciones del diccionario. Se trata del programa *COMPADRAE* que, como ya se ha indicado, posibilita una primera comparación de los listados de formas obtenidos a través del *NTLLE*. Posteriormente se procede, a través del cotejo, a la identificación de las variantes formales, las erratas, las incorporaciones y, un aspecto muy importante en algunas ediciones estudiadas (en especial, el *DRAE* 1822 y el *DRAE* 1832), las supresiones de lemas y formas.

3.4.2 Página web y repositorio de datos

Uno de los puntos fuertes de las Humanidades digitales se encuentra en los nuevos soportes que proporcionan para el almacenamiento y difusión de los resultados obtenidos. La página web del proyecto (<http://draesxix.wixsite.com/draesxix>) se configura como un portal desde el que se intenta difundir y dar visibilidad al proyecto y los logros alcanzados. La página alberga aquellos contenidos cuya difusión más apropiada corresponde a los medios digitales. Se obtiene de esta forma una enorme facilidad de acceso a unos materiales que están en continua actualización y ampliación. Destacan en este terreno el apartado dedicado a la bibliografía y el repositorio de formas de los diccionarios estudiados con las informaciones obtenidas en el proyecto que nos encontramos construyendo.

La bibliografía reúne los estudios publicados sobre el diccionario académico a lo largo de su historia. El repositorio de datos almacena las formas que constituyen la macroestructura de los diccionarios con informaciones acerca de sus variantes formales y su cronología en las obras lexicográficas. Se consigue con ello compartir información y de ese modo transferir a la sociedad el conocimiento generado gracias a proyectos de investigación financiados con fondos públicos. El soporte en base de datos facilita la recuperación del material con distintas posibilidades de consulta, con lo que se incrementa su utilidad.

4 Conclusión

Lo expuesto hasta aquí son, como se puede observar, las características de un proyecto de investigación, el método empleado en su ejecución y el modo como los miembros del equipo queremos dar a conocer los resultados. Que

las ediciones del *Diccionario de la lengua española* de la Real Academia Española merecen un estudio detenido, por la relevancia de la obra en el panorama lexicográfico del español y por la autoridad de la institución que la ha confeccionado, es algo que hoy ningún investigador en lexicografía pone en duda. Y esto es así porque no solo es un deseo, sino una necesidad conocer los cambios de cada edición y las razones que los han producido. Este proyecto, por medio de la comparación de ediciones, del cotejo de lemas y del examen de entradas, quiere trazar la historia interna del diccionario, una historia que completará la de la lexicografía académica, hasta ahora parcial, por el análisis de diccionarios concretos o aspectos concretos de los diccionarios, o por atender al desarrollo y evolución de la propia corporación. Desentrañar el diccionario, parcelar el contenido para examinarlo en las distintas ediciones, sea el aumento y el tipo de léxico al que pertenece ese aumento, la definición, la marcación, los elementos que permiten la conexión entre el diccionario y otras obras lexicográficas, entre otros muchos mencionados a lo largo de estas páginas, es la tarea en la que estamos empeñados; y lo estamos para poder así ofrecer el análisis de los resultados en trabajos de investigación, de los que ya hay muestras relevantes que recoge la bibliografía, y en el repositorio de datos que estará en nuestra página web, junto con otro tipo de información complementaria, a disposición de toda la comunidad científica.

Referencias bibliográficas

Alvar Ezquerra, Manuel (2014–): *Biblioteca virtual de la filología española* <https://www.bvfe.es/> [último acceso: 15/10/2017].

Azorín, Dolores (2015): «El neologismo en la tradición académica (1780–2014)», *Estudios de lexicografía* 2, 58–70.

Azorín, Dolores (2018): «El léxico con marcación estilística», en Gloria Clavería y Margarita Freixas (coords.), 427–458.

Azorín, Dolores/José Manuel Blecua/M.ª Ángeles Blanco/Cristina Buenafuentes/Gloria Clavería/Sheila Huertas/Margarita Freixas/Enrique Jiménez Ríos/Carolina Julià/Laura Muñoz/Ana Paz/Marta Prat/Matthias Raab/Joan Torruella/Sonia Varela (2017): «Historia interna del *Diccionario de la lengua castellana* de la Real Academia Española en el siglo XIX (1817–1852)», en Ignacio Sariego *et al.* (eds.), *El diccionario en la encrucijada: de la sintaxis y la cultura al desafío digital*. Santander: Escuela Universitaria de Turismo de Altamira-Asociación Española de Lexicografía Hispánica, 151–172. Disponible en <http://lexicografia.incacav.es/Libro-VII-Congreso-Lexicografi%CC%81a.pdf>

Battaner, Paz (1996): «Terminología y diccionarios», en *Jornada panllatina de terminologia. Perspectives i camps d'aplicació*. Barcelona: IULA, 93–117.

Blanco, M.ª Ángeles (2018a): «El contexto del diccionario: los códigos vigentes en 1817», en Gloria Clavería y Margarita Freixas (coords.), 57–63.

Blanco, M.ª Ángeles (2018b): «Las voces de la ortografía y la gramática», en Gloria Clavería y Margarita Freixas (coords.), 371–398.

Blanco, M.ª Ángeles (2018c): «Un proceso significativo de revisión: la definición de las letras», en Gloria Clavería y Margarita Freixas (coords.), 175–202.

Blecua, José Manuel (2018): «Presentación», en Gloria Clavería y Margarita Freixas (coords.), 9–11.

Buenafuentes, Cristina (2018): «Los criterios de lematización: las formas complejas», en Gloria Clavería y Margarita Freixas (coords.), 115–136.

Carriazo, José Ramón (2018): «Las voces de la navegación: náutica, ingeniería naval y áreas afines», en Gloria Clavería y Margarita Freixas (coords.), 339–355.

Clavería, Gloria (2016): *De vacunar a dictaminar: la lexicografía académica decimonónica y el neologismo*. Madrid/Fráncfort: Iberoamericana/Vervuert.

Clavería, Gloria (2017): «Historia interna del *Diccionario de la lengua castellana* de la Real Academia Española (1822–1852): investigaciones en curso y resultados», en *Seminario de investigación: Historia interna del «Diccionario de la lengua castellana» de la Real Academia Española (1822–1852)*. Barcelona: UAB. Disponible en <http:// http://draesxix.wixsite.com/draesxix/presentaciones-del-proyecto> [último acceso: 15/10/2017].

Clavería, Gloria (2018): «La quinta edición del *Diccionario de la Lengua Castellana* (1817) de la Real Academia Española al microscopio», en Gloria Clavería y Margarita Freixas (coords.), 15–55.

Clavería, Gloria/Margarita Freixas (2015): «La quinta edición del *Diccionario de la lengua castellana* de la Real Academia Española (1817): el aumento de voces», en José M.ª García Martín (ed.), *Actas del IX Congreso Internacional de Historia de la Lengua Española*, vol. II. Madrid/Fráncfort: Iberoamericana/Vervuert, 1309–1326.

Clavería, Gloria/Margarita Freixas (coords.) (2018a): *El diccionario de la Academia en el siglo XIX: la quinta edición (1817) al microscopio*. Madrid: Arco/Libros.

Clavería, Gloria/Margarita Freixas (2018b): «El *Nuevo tesoro lexicográfico de la lengua española*: un museo lexicográfico como base de datos», *Cuadernos del Instituto Historia de la Lengua* 11, 117–138.

Clavería, Gloria/Margarita Freixas/Joan Torruella (en prensa): «Historia interna del *Diccionario de la lengua castellana* de la Real Academia Española en el siglo XIX (1817–1852): el léxico especializado», en Cecilio Garriga y M. Betulia Pedraza (eds.), *Estudios de lengua y ciencia*. A Coruña: Anexos de la *Revista de Lexicografía*.

Clavería, Gloria/Ana Paz (en prensa): «El discurso científico en la definición lexicográfica académica (*DRAE* 1817–1852)», en Xosé A. Álvarez *et al.* (eds.), *Nuevas perspectivas en la diacronía de las lenguas de especialidad*. Alcalá de Henares: Servicio de Publicaciones de la Universidad de Alcalá.

Córdoba, Jessica/Natalia Terrón (2017): «Cambios en la microestructura: *DRAE* 1852 *vs. DRAE* 1843», en *Seminario de investigación: Historia interna del «Diccionario de la lengua castellana» de la Real Academia Española (1822–1852)*. Barcelona: UAB <https://draesxix.wixsite.com/draesxix/presentaciones-del-proyecto> [último acceso: 15/10/2017].

Esteve, Abraham (1982): *Estudios de teoría ortográfica del español*. Murcia: Universidad de Murcia.

Freixas, Margarita (2018): «La definición y la descripción», en Gloria Clavería y Margarita Freixas (coords.), 139–173.

Freixas, Margarita/Gloria Clavería (2018): «Los criterios de lematización: los lemas simples y los lemas múltiples», en Gloria Clavería y Margarita Freixas (coords.), 93–113.

García Belmar, Antonio/José Ramón Bertomeu (2006): *La revolución química. Entre la historia y la memoria*. Valencia: Universitat de València.

Garriga, Cecilio (2018): «Las voces de la química», en Gloria Clavería y Margarita Freixas (coords.), 313–337.

Gómez Asencio, José J. (2011): *Los principios de las gramáticas académicas (1771–1962)*. Berna: Peter Lang.

Gómez de Enterría, Josefa (2018): «Las voces de las ciencias naturales y áreas afines», en Gloria Clavería y Margarita Freixas (coords.), 275–311.

GRAE 1796 = Real Academia Española (1796): *Gramática de la lengua castellana*, cuarta edición, corregida y aumentada. Madrid: Viuda de don J. Ibarra.

«Historia interna del *Diccionario de la lengua castellana* de la Real Academia Española en el siglo XIX (1817–1852)». <https://draesxix.wixsite.com/draesxix> [último acceso: 15/10/2017].

Hoyos, José Carlos (2018): «Las voces de la economía política», en Gloria Clavería y Margarita Freixas (coords.), 357–370.

Jiménez Ríos, Enrique (2018): «El léxico con marcación diacrónica», en Gloria Clavería y Margarita Freixas (coords.), 399–426.

Jiménez Ríos, Enrique/Gloria Clavería (2018): «Las correspondencias latinas», en Gloria Clavería y Margarita Freixas (coords.), 495–527.

Julià, Carolina (2018): «Un ejemplo de marcación diatópica: la incorporación de meridionalismos», en Gloria Clavería y Margarita Freixas (coords.), 481–494.

Lázaro Carreter, Fernando (1972 [1980]): «El primer diccionario de la Academia», en *Estudios de lingüística*. Barcelona: Crítica, 83–148.

Martínez Alcalde, M.ª José (2010): *La fijación ortográfica del español: norma y argumento historiográfico*. Berna: Peter Lang.

Muñoz, Laura (2018): «La sinonimia y las remisiones», en Gloria Clavería y Margarita Freixas (coords.), 203–221.

NTLLE = Real Academia Española (2001): *Nuevo tesoro lexicográfico de la lengua española*. Madrid: Espasa Calpe, 2 DVD. Disponible en <www.rae.es>.

Paz, Ana (2018): «La información lexicográfica de las formas complejas», en Gloria Clavería y Margarita Freixas (coords.), 223–252.

Prat, Marta (2018): «Los criterios de lematización: los lemas múltiples y la sufijación apreciativa», en Gloria Clavería y Margarita Freixas (coords.), 103–113.

Raab, Matthias (2018): «La lexicografía no académica y la quinta edición del *DRAE*», en Gloria Clavería y Margarita Freixas (coords.), 543–570.

Real Academia Española (1815): *Ortografía de la lengua castellana*, octava edición notablemente reformada y corregida. Madrid: Imprenta Real.

Real Academia Española (1844): *Prontuario de ortografía de la lengua castellana,* dispuesto de Real Orden para el uso de las escuelas públicas por la Real Academia Española con arreglo al sistema adoptado en la novena edición de su diccionario. Madrid: Imprenta Nacional.

Sarmiento, Ramón (2001): *La norma ortográfica de la Real Academia Española (1741): aportación al estudio del español moderno*. Madrid: Ediciones de Cultura Hispánica.

Terrón, Natalia (2018): «La regularización ortográfica», en Gloria Clavería y Margarita Freixas (coords.), 67–91.

Terrón, Natalia/Joan Torruella (2017): «*DRAE* 1822: cambios en la microestructura», en *Seminario de investigación: Historia interna del «Diccionario de la lengua castellana» de la Real Academia Española (1822–1852)*. Barcelona: UAB. <https://draesxix.wixsite.com/draesxix/presentaciones-del-proyecto> [último acceso: 15/10/2017].

Torruella, Joan/Sheila Huertas (2018): «Las voces de especialidad: caracterización general», en Gloria Clavería y Margarita Freixas (coords.), 253–273.

Varela, Sonia (2018): «El léxico con marcación diatópica», en Gloria Clavería y Margarita Freixas (coords.), 459–480.

José Ignacio Pérez Pascual

Las publicaciones periódicas y el estudio del léxico de la «Edad de Plata»[1]

Resumen: Una de las etapas más desatendidas por los historiadores de la lengua es el medio siglo conocido como la «Edad de Plata» (1885–1936), pues solo se aprecian cambios en el plano léxico. Hemos querido colaborar a su mejor conocimiento estudiando el vocabulario de la época; recurrimos especialmente a las publicaciones periódicas, que nos permiten analizar numerosas voces con mayor precisión que otras herramientas digitales. En esta ocasión nos servimos, a modo de ejemplo, de palabras incluidas hace muy poco en los diccionarios de referencia (*armañac*, *moriles*, *pisco*, *quianti* o *albariño*) y comprobamos, al tiempo, la utilidad de este tipo de textos en el caso de otras documentadas mucho más tempranamente (*filipichín*).

Palabras clave: Lexicografía, Diccionario histórico, «Edad de Plata», Publicaciones periódicas

Abstract: The period known as the «Silver Age» (1885–1936) is one of the least explored ages by historians of the Spanish language due to the fact that changes are observed just at the level of lexicon. We want to contribute to this knowledge by studying the vocabulary of those days; we use especially periodical publications, which allow us to analyze a great number of words with more precision than with other digital tools. On this occasion, we examine, as an example, some words included very recently in relevant dictionaries (*armañac*, *moriles*, *pisco*, *quianti* o *albariño*) and, at the same time, we verify the value of this type of texts in case of other words documented much earlier (*filipichín*).

Keywords: Lexicography, Historical Dictionary, «Silver Age», Periodical Publications

1 Este trabajo se integra en los proyectos «*Diccionario del español de la "Edad de Plata" (continuación)*» (ref. FFI2011-23085, MINECO) y «*Documenta Philologa.* Los archivos como fuente de información para la historia de la Filología española: el *Centro de Estudios Históricos*» (ref. FFI2015-65939-P, MINECO-FEDER), y se ha beneficiado también de una «Axuda para a consolidación e estruturación de unidades de investigación competitivas do Sistema Universitario de Galicia» de la Xunta de Galicia concedida al *Grupo de Investigación Hispania* de la Universidad de A Coruña como «Grupo con Potencial de Crecimiento» (ref. GPC2015/028).

1 Introducción

Entre las tareas que tiene pendientes la Filología hispánica figura disponer en un plazo razonable del *diccionario histórico* del que todavía carece nuestra lengua[2]. No es necesario explicar en este momento que el diccionario histórico no es un lujo, sino un instrumento de trabajo imprescindible para multitud de tareas filológicas. Como señalaba Manuel Seco, para conocer la historia de nuestro léxico:

> El primer instrumento para ello sería el diccionario histórico: ese inventario total del léxico del que disponen casi todas las lenguas cultas de nuestro entorno, y del que imperdonablemente carece la nuestra (Seco 2007: 10)[3].

De hecho, ante la falta de un diccionario histórico, solemos acudir al *Diccionario crítico etimológico castellano e hispánico* de J. Corominas y J. A. Pascual (en adelante *DECH*), aunque esta obra dista mucho de poder sustituirlo, por más que el filólogo catalán se refiriese a la orientación profundamente histórica de su obra (*cf. DCEC*: ix y *DECH*: xiii); los filólogos suelan considerar, como Gregorio Salvador, que aunque no disponemos de un diccionario histórico

> sí tenemos, en cambio, un *Diccionario etimológico*, el de Corominas, que suple hasta donde puede esa carencia. […] la suma de las historias de las palabras de la lengua no es la historia de la lengua. Pero la verdad es que esa suma de historias, una suma hecha homogéneamente y de la misma mano, no existe para ninguna lengua del mundo si no es para el español. Gracias a Corominas, ese gigante (Salvador 1983: 147–148).

Y, en efecto, no hay duda de que dentro de los estudios de lexicografía histórica supuso un gran avance la publicación del *DECH*, pero, con todo, no es menos cierto que esta segunda edición ampliada del *DCEC* no pudo superar del todo dos defectos habituales en la mayor parte de los trabajos de corte histórico sobre nuestra lengua: haberse apoyado excesivamente en testimonios literarios y contar con una información descompensada con respecto a los diferentes períodos cronológicos[4]. Así pues, a pesar del carácter ingente de la obra del filólogo

---

2 Obviamente no podemos comparar nuestra situación con la inglesa, pues la primera edición del *Oxford English Dictionary* culminó en 1928, pero tampoco con la francesa o la catalana. Para la historia y problemas del diccionario histórico español, puede consultarse Alvar Ezquerra (1976: 30–39), Lapesa (1992: 13–86), Porto Dapena (2000) y Seco (2003: 109–156 y 163–182).

3 Añade que en un segundo término podremos acudir a la tradición lexicográfica, pero que los diccionarios generales «informan, en teoría, sobre el léxico de su propia época» y «su ayuda es bastante dudosa en el aspecto cronológico» (Seco 2007: 10).

4 En líneas generales, podemos afirmar que han sido cientos los trabajos que se han aproximado al léxico castellano medieval —especialmente al del período alfonsí— y

catalán, sería deseable contar con un verdadero diccionario histórico, y no seguir utilizando el *DECH* como su sucedáneo.

Hace ya quince años que se hizo pública la intención de la Real Academia de afrontar la redacción del *Nuevo diccionario histórico del español* (en adelante *NDHE*), bajo la dirección de José Antonio Pascual, proyecto concebido como un diccionario relacional y que, por tanto, ya no se sometía a la tiranía de la redacción siguiendo el orden alfabético, como se había hecho en anteriores intentos[5].

Es nuestra intención, en la medida de nuestras limitadas fuerzas, contribuir a remediar la situación existente; para ello, con absoluto pragmatismo, hemos elegido centrar nuestra atención en una parcela concreta de la historia de nuestra lengua, ese período situado entre dos siglos que convencionalmente llamaremos «Edad de Plata», recurriendo a una denominación aceptada en los estudios sobre la cultura española del tiempo, y cuyos límites se situarían, según los diferentes estudiosos, entre 1898 y 1936 o entre 1902 y 1939; por nuestra parte, hemos preferido ampliar un tanto el período, hasta abarcar medio siglo, desde 1885, fecha del fallecimiento de Alfonso XII, hasta 1936, fecha suficientemente significativa[6]. Nos ha parecido que el período así limitado poseía unas características propias, pues una serie de acontecimientos históricos de gran trascendencia, acompañados de profundas transformaciones políticas, económicas y sociales, y hasta la existencia de toda una serie de movimientos culturales —entre los que ocupan un destacado lugar los literarios—, avalan esta segmentación y nos animan a analizar un período que, *grosso modo*, viene a coincidir con el reinado de Alfonso XIII (incluyendo su minoría de edad) y la II República.

---

en menor grado al de los Siglos de Oro, aunque no haya sido sino para establecer concordancias o vocabularios de obras literarias concretas. Tampoco resultan escasos, sobre todo en los últimos años, los trabajos acerca de las más recientes manifestaciones léxicas del español, bien sea para estudiar ciertos lenguajes sectoriales o para analizar los más recientes neologismos.

5 Las contribuciones científicas de algunos colaboradores del proyecto ilustran la realidad del diccionario, que ya ha puesto a nuestra disposición un millar de voces en su página web (http://www.rae.es/recursos/diccionarios/nuevo-diccionario-historico); *cf*., entre otras aportaciones, Campos Souto (2007, 2015 y 2016), Campos Souto y Pascual (2012a y b), Pascual (2015), Pascual y Campos (2014), Pinillos Laffon (2015), Salas Quesada y Torres Morcillo (2011 y 2015).

6 Fechas que, además, coinciden casi perfectamente con las de impresión de dos importantes ediciones del *Diccionario* de la Real Academia Española (en adelante *DRAE*, con precisión de la fecha de la edición utilizada): 1884 y 1936; para la consulta de las distintas ediciones del *DRAE* (hasta la de 1992) y de otras obras lexicográficas, cuyas informaciones resultan también muy valiosas para conocer el léxico del período, hemos

2 «Nuevas realidades, nuevas ideas, nuevas palabras»

No ha sido este medio siglo una etapa atractiva para los historiadores de la lengua, y resulta hasta cierto punto lógico que hayan preferido dedicar su atención a períodos anteriores, en los que es posible analizar toda una serie de transformaciones en los planos fonológico, morfológico y sintáctico; cierto es, en efecto, que los últimos siglos de nuestra historia, desde finales del XVII, contemplan casi exclusivamente cambios en el plano léxico. Así pues, la «Edad de Plata» ha sido desatendida por los estudios de carácter diacrónico, posiblemente porque su cercanía hace que demos por supuesto lo innecesario de toda investigación histórica; y, sin embargo, este prejuicio no debería impedirnos estudiar la gran evolución producida en el plano léxico durante esta etapa[7]. Sin embargo, hemos de lamentar la ausencia de visiones de conjunto acerca de la lengua de la época y, si comprobamos la relación de obras utilizadas por Corominas y Pascual en el *DECH*, podemos observar que se han servido fundamentalmente, para la etapa que nos interesa, de las diversas ediciones del diccionario de la Academia y de algunos trabajos de filólogos del momento. Además, si tomamos en consideración los escasos estudios que se han acercado al léxico del período, solemos encontrarnos con acercamientos de carácter literario de los que —tangencialmente— podemos extraer algún dato valioso acerca del componente léxico de nuestro idioma[8]; de hecho, el estudio más completo al que podemos acudir es «La lengua» de Rafael Lapesa (1994), un breve trabajo recogido en la *Historia de España Ramón Menéndez Pidal*, especialmente las páginas centradas en «Nuevas realidades, nuevas ideas, nuevas palabras», donde proporciona escasas referencias bibliográficas[9].

Con la intención de contar, para la tarea que afrontamos, con un corpus que tuviese una cierta representatividad[10] y que, al tiempo, no resultase inmanejable, seleccionamos para nuestro proyecto:

---

disfrutado de la ventaja de contar con el *Nuevo Tesoro Lexicográfico de la Lengua Española* (en adelante *NTLLE*), accesible en la página de la Real Academia.

7 Desde una perspectiva geográfica, ha sido nuestra intención centrarnos en el territorio español, aunque en ocasiones debamos contrastar nuestros datos con los que vamos recogiendo de territorios americanos.

8 Como resulta lógico, son los escritores de primera fila quienes han recibido una mayor atención por parte de los investigadores (Galdós, los grandes autores del 98…).

9 A los títulos allí citados podemos añadir algunos otros como la recopilación de trabajos del propio Lapesa (1996), o las diversas aportaciones de Abad (1985), Senabre (1999) y Álvarez de Miranda (2004).

10 Ha de cuidarse la selección de textos, para que estén bien representados los distintos períodos, tipos de textos, autores, zonas geográficas de este lado del Atlántico, etc.,

a) una serie de obras literarias representativas, fundamentalmente narrativas y teatrales[11];
b) un cierto número de textos «técnicos» (enciclopedias, libros de texto, estudios técnico-científicos de la época…), en su mayor parte ignorados por los lingüistas;
c) una selección de la prensa de la época (diarios y revistas), en la que encontramos valiosas precisiones acerca del momento de introducción en el idioma de determinadas voces, su significado exacto en ese tiempo, la forma que adopta, etc.[12]

En lo que respecta al tipo de léxico examinado, sin dejar de registrar todo tipo de voces, decidimos dar preferencia en un primer acercamiento al que se relaciona con la vida cotidiana, pues las denominaciones de la indumentaria, la cosmética, los utensilios del hogar, los deportes, el ocio, las costumbres sociales… componen un retrato del léxico y, a la vez, de la realidad del momento. El estudio de cada nueva palabra es también el estudio del cambio histórico-social que conlleva. Con ello pretendimos obtener una información que habrá de tener, en el futuro, infinidad de aplicaciones:

a) En el campo de la precisión cronológica. Mediante los datos allegados (principalmente a través de publicaciones periódicas) podremos fechar con mayor exactitud el momento en que penetran en nuestro idioma numerosas voces que diccionarios etimológicos o históricos apenas catalogan como «modernas» o «contemporáneas». No carece de interés comprobar, por ejemplo,

---

sin que se produzcan desequilibrios. Hemos podido comprobar, por ejemplo, que la selección en el *CORDE* de un texto como *El visitador* de José Milla y Vidaurre provoca que una palabra como *barbero* vea incrementado el número de testimonios en varios cientos, mientras que son apenas unas decenas los que podemos encontrar en la totalidad de las restantes obras seleccionadas para el XIX.

11 Las calas realizadas en algunas novelas de corte realista han confirmado nuestra hipótesis previa de que en sus descripciones de la España del momento habríamos de encontrar valiosas informaciones acerca del léxico de la vida cotidiana.

12 Como ha advertido Manuel Seco, el lenguaje de la prensa «es uno de los frentes más activos en la renovación del léxico. Ya lo advirtió Unamuno hace más de cien años, cuando solo se disponía del periódico. "No creo –decía– que haya institución más a propósito que la de la prensa para recoger el idioma vivo, el que [...] se está formando día a día, en labor incesante, junto al lenguaje hecho ya y consagrado". […] En nuestros días sigue siendo muy cierto que las páginas del periódico, y hoy los espacios de los medios orales, constituyen un ancho escaparate de todas las nuevas tendencias que se mueven en el léxico» (Seco 2007: 16).

cómo los autores decimonónicos suelen utilizar la voz *transparencias*[13] o *cortinillas* y no *visillos*, usual hoy en día y que el *DRAE* recoge sólo desde 1884. Tampoco resultará difícil matizar opiniones como la de un investigador habitualmente tan preciso como Rafael Lapesa, quien, sin embargo, afirma que la voz *telégrafo* se registra en el *DRAE* desde 1884, cuando ya está en la edición de 1803 (y se documenta ampliamente en el *CORDE*): naturalmente, el antiguo *telégrafo*, con sus torres situadas cada poca distancia, se ve sustituido por la moderna —en aquel momento— telegrafía, pero el término ya existía.

b) En el dominio de la etimología. Sin duda posee una innegable capacidad poética la explicación que ofrecen Corominas y Pascual, aceptando una sugerencia de Spitzer, acerca del origen del producto de bollería conocido como *magdalena*, documentado en español desde el diccionario de Gaspar y Roig, en 1855, y por la lexicografía académica desde el *DRAE* 1869: «quizá llamado así porque se emplea para mojar y entonces gotea 'llorando como una Magdalena'» (*DECH*: s. v. *magdalena*). La lectura de Pérez Galdós nos proporciona una explicación etimológica más razonable, aunque sin duda de menor belleza; en realidad, se trata de uno de esos casos, bien conocidos en los estudios de semántica, en que el creador del producto ha visto su nombre convertido en sustantivo común para designar a una de sus especialidades: Madeleine Paumier, doncella de Commercy, quien recurrió a una receta familiar para agasajar en un banquete —ante la ausencia del repostero oficial— al Duque Stanislas, suegro de Luis XV. Esta explicación (que acepta el *DRAE* 2001), figura ya en *Le grand dictionnaire de la cuisine* de Alejandro Dumas (1873), de donde seguramente pasó la información al autor de origen canario.

c) En lo que se refiere al significado. La información del momento puede precisarnos el sentido exacto que cada vocablo tenía en la época (así, por ejemplo, los ya citados *visillos* comenzaron siendo una cortina que se utilizaba en las ventanas de los carruajes). Podremos registrar, además, voces que no figuran en la tradición lexicográfica académica y que, sin embargo, han tenido en su momento una cierta difusión; por no alejarnos del mencionado Dumas, cabe recordar que si hoy el término *montecristo*, con un buen número de ocurrencias en el *CREA*, designa un excelente tipo de cigarros (es bien conocida la razón por la que lleva ese nombre), esta misma forma léxica designaba a fines del XIX a un tipo de abrigo masculino, como lo prueba una decena de apariciones del término en textos de finales de nuestro período (a uno y otro lado del Atlántico, incluso en escritores del calibre de Emilia Pardo Bazán o

13 *Cf*. Corominas y Pascual (*DECH*, *s. v*. *ver* y *parecer*), donde figuran estas voces sin ninguna indicación acerca de su significado ni precisión cronológica alguna.

Ramón del Valle Inclán); sin embargo, a pesar de ello, sólo lo recoge una obra lexicográfica (Toro y Gómez, 1901) y no figura en ninguna edición del *DRAE* (para las denominaciones de la indumentaria en la época, véase el excelente volumen de Cotelo 2014).

d) Nuestra labor puede facilitar incluso el estudio de las disputas lingüísticas en la España del XIX, pues no pocas páginas de la prensa de la época se centran en justificar o rechazar determinadas palabras apoyándose para ello en las más variadas argumentaciones: el deseable casticismo del idioma, el antigalicismo, etc.[14]. Un minucioso seguimiento de la prensa favorece, además, el estudio del proceso histórico de asimilación a nuestra lengua de términos foráneos como, por ejemplo, la designación del actual *coche-cama* como *sleeping-car* (*ABC* de 1901), que aparece únicamente en tres diccionarios (Zerolo 1895; Alemany 1917 y Rodríguez Navas 1918).

3 Neologismos y préstamos en las publicaciones periódicas

Nuestro examen del léxico del período nos ha convencido de que los cambios históricos, sociales y tecnológicos que se suceden durante esa etapa dan lugar a un continuo proceso de innovación en el léxico, en el que abundantes préstamos y numerosos neologismos se introducen en nuestra lengua; en algunos casos se trata de creaciones pasajeras, pero otras muchas palabras enraizarán en la lengua. Se trata de un nuevo y abundante caudal léxico cuya difusión conviene rastrear en la prensa periódica, cuyo interés para el estudio diacrónico del léxico, por encima de los corpus que habitualmente se consultan, pretendo recalcar en estas páginas[15]. Para ello hemos seleccionado un reducido número de voces de la vida cotidiana[16].

Para comenzar con esta muestra, he seleccionado un par de palabras que se han incorporado hace muy poco al diccionario de la Academia; se trata de *armañac* y *moriles*. En el primero de estos ejemplos la entrada en el diccionario es tan

14 En otro lugar (Pérez Pascual en prensa) aludo a la información periodística acerca de las decisiones académicas sobre la inclusión de algunos arabismos contemporáneos en el *DRAE*.

15 He podido leer un sugerente trabajo muy reciente de Octavio de Toledo y Huerta quien recomienda, frente a «los corpus textuales más al uso», «fondear en bancos de datos menos explorados hasta la fecha, e incluso […] a emprender una azarosa singladura por el ancho mar sin fronteras de internet» (2016: 24).

16 He optado por prescindir de voces técnicas y científicas, de las que me he ocupado en alguna otra ocasión, acudiendo también en esos casos a la información que proporcionan las publicaciones periódicas; *cf*. Pérez Pascual (2012, 2013 y 2014).

reciente que se ha incorporado al repertorio académico ya en este milenio, en la edición de 2001[17]:

> *armañac.* (Del fr. armagnac, y este de Armagnac, región del suroeste de Francia). 1. m. Aguardiente de uva, muy parecido al coñac (*DRAE* 2001 y *DLE* 2014).

¿Qué información proporcionan sobre este término los corpus que utilizamos habitualmente cuando tratamos de documentar alguna voz en la historia del español? Hemos de comenzar, como es lógico, por el *CORDE* y nos encontraremos con apenas dos testimonios en un único texto (Pío Font Quer, *Plantas Medicinales. El Dioscórides Renovado*, 1962) en los que se mantiene la grafía francesa:

> Añadidle el armagnac ya quemado y echad el caldo, poquito a poco…
> En una pequeña cacerola, calentad e inflamad el armagnac. Dejadlo quemar hasta que se apague por sí solo.

Naturalmente, hay más ejemplos en el *CREA*, tanto manteniendo la grafía propia de la lengua de procedencia del vocablo:

> habrá ayudado a terminar esa robusta botella de armagnac que testimonia mi glorioso retorno (del argentino Marcos Ricardo Barnatán, *Con la frente marchita*, 1989).

> con dos botellas de armagnac. En el estuche leo y releo la dedicatoria (del español José Ramón de la Morena, *Los silencios de El Larguero*, 1995).

como adoptando una grafía hispanizada, a partir de 1994, en España:

> sufren una grave crisis en el sector del armañac (*La Vanguardia*, 21/05/1994),

y en América:

> pueden usar el nombre de cognac, usarán armañac u otro nombre (José María de Romaña, *Las bebidas en el Perú*, 1995).

El *CDH*, por su parte, anticipa ligeramente la documentación de la forma más antigua, la que respeta la grafía propia del francés, pues la documenta en 1956 en la Península («les manda por Navidades un barrilito de Armagnac», Rafael Sánchez Mazas, *La vida nueva de Pedrito de Andía*) y en 1962 en territorio americano, en México («jamón de bayona, estofados de res rociados de Armagnac, cuellos de oca rellenos de paté de puerco», Carlos Fuentes, *La muerte de Artemio Cruz*); además, el *CDH* recoge también muestras más tempranas que las del *CORDE* de la escritura españolizante:

---

17 No me demoraré en el examen de los cambios que se producen entre las diversas ediciones académicas que menciono, pues no es el objeto de este trabajo.

> Se puede incluir un *Armañac*. Es parecido al coñac pero con personalidad propia (José Antonio de Urbina, *El arte de invitar. Su protocolo*, 1989).

Con todo, podemos obtener sin gran esfuerzo datos mucho más tempranos consultando diversas publicaciones periódicas, buena parte de las cuales son de fácil acceso gracias a la Hemeroteca de la Biblioteca Nacional[18], donde aparecen ejemplos que adelantan casi un siglo nuestra documentación. Dejaremos de lado, claro está, aquellos textos en los que se alude al «aguardiente de Armagnac»[19], y nos centraremos en los casos en que el topónimo de la región gascona ya se ha lexicalizado para designar a esta bebida[20], con ejemplos muy claros desde 1875, fundamentalmente en folletines publicados en algunos periódicos:

> sobre todo si hay provisión de este excelente armagnac, porque es Armagnac lo que acabo de beber, ¿no es cierto? (Fortune de Boisgobey, «El as de oros», folletín de *El Imparcial*, 06/10/1875: 4).

> un armagnac de tres años que os daré a bajo precio. —Veamos el armagnac» (Javier de Montepin, «El médico de las locas», folletín de *La Unión*,16/08/1880: 3).

> es Armagnac de 1820 (*Ilustración artística*, 15/05/1893: 12).

Y si los ejemplos citados hasta ahora empleaban la grafía afrancesada, la opción castiza se rastrea desde 1890:

> saboreaba una copa de armañac (Pedro Sales, «Un drama financiero», folletín de *La Época*, 05/11/1890: 4).

> tras un sorbo de Armañac viejo (Enrique Gaspar, «Las personas decentes», folletín de *La Época*, 14/12/1899: 4).

Tampoco es muy antigua la presencia de *moriles* en el diccionario académico, pues se incorpora a él en 1970 y en el *Diccionario Manual* de la Academia (en adelante *DRAEM*, con precisión de la edición empleada), y sufre mínimos cambios en su definición durante estas décadas, pues la inicial:

---

18 Si no indicamos otra cosa, nos servimos de publicaciones madrileñas; debe tenerse en cuenta el predominio de publicaciones españolas en esta valiosa hemeroteca.

19 Expresión que ya figura en *El Correo mercantil de España y sus Indias* de uno de mayo de 1797 y que podemos registrar en muchísimas publicaciones: *Diario de avisos de Madrid* (25/03/1826: 2), *El Heraldo* (03/09/1842: 4), *El Lloyd español* (31/10/1861: 2).

20 Dejamos de lado algunos casos en el que no estamos seguros de que armagnac ya estuviese lexicalizado («*Líquidos espirituosos*. —Sin variación cotizándose los del Languedoc á 95; Bethime, 85; Armagnac, 90, aguardiente de caña de 50á 60», *El Lloyd español*, 04/11/1862: 3).

> Vino de fina calidad que se cría y elabora en el término municipal de Moriles, pueblo de la provincia de Córdoba, España (*DRAE* 1970 y *DRAEM* 1983).

se abrevia en 1992:

> Vino fino que se cría y elabora en el término municipal de Moriles (*DRAE* 1992),

aunque la referencia a la provincia cordobesa se recupera en el 2001:

> Vino fino que se cría y elabora en el término municipal de Moriles, en la provincia de Córdoba, España (*DRAE* 2001 y *DLE* 2014).

La documentación del *CORDE* arranca apenas en 1972, en el conocidísimo recetario de Simone Ortega («la manzanilla, los moriles y montilla, el whisky, la ginebra», *1080 Recetas de cocina*), y, ese mismo año, Alonso Zamora Vicente utiliza este término en dos ocasiones en su obra narrativa:

> y luego otra cerveza, y más vino, y otro vino, y moriles, y coñac, y chinchón un par de veces […] se come unas gambas bien mojadas en moriles, y mira con descaro a una vecina (Alonso Zamora Vicente, *A traque barraque*).

En esta ocasión, el *CDH* se remonta algo más de un cuarto de siglo con respecto a los datos del *CORDE*, al documentar la palabra en un texto publicitario de 1945:

> a través del Moriles… todo es más bello, Regalo de un botellín de Moriles a todo concursante cuyo número coincida en sus dos últimas cifras con el premiado (1945, *ABC*, 19/12/1945: 12).

Pero mucho más antiguos son algunos de los ejemplos que podemos recoger a través de las hemerotecas, un siglo anteriores a los que nos ofrecía el *CORDE*:

> Vino de Montilla de 30 a 32 reales; Moriles de 30 a 32; Córdoba de 40 a 48 (*La Correspondencia de España*, 20/01/1861: 4).

> Los más exquisitos de la Francia, pueden competir con nuestros Malvasías, Pedros Jiménez, Moscateles, Canarias, Montillas, Moriles, Nava, Aloques, Málaga […] (Lorenzo de Merlo, «Los vinos en España», *La Ilustración española y americana*, 22/09/1875: 9).

> Moriles superior, 15 años, 3,25 pts. botella (*El Imparcial*, 24/12/1894: 4).

Es muy interesante, y no solo desde una perspectiva lingüística, el detalle del menú consumido en una cena pantagruélica que el cabildo catedralicio ofrece al obispo recién nombrado en 1898, un año clave de nuestra historia en que el país no estaba precisamente para fiestas. El menú publicado por los periódicos fue este:

Sopa:	Consomé a la royale.
Relevé:	Filete a la princesa.
Frito:	Variado.

Entradas: Pescada a la Tártara, vol-au-vent de ave, Foie-gras al aspic, ponche a la romana.
Legumbres: Menestra con jamón.
Asado: Pavipollos con cressons.
Ensalada: del tiempo.
Dulces: Budin de gabinete, queso helado a la vainilla, postres y entremeses.
Vinos: Moriles, Oro López Diéguez, Rioja, Marqués del Riscal.
Champagne: Gladiateur Carte d'Or.
Licores: Benedictine, chartreuse, anisete María Brizar, Curaçao d'Hollande.
Café, cigarros habanos.
Este festín baltasárico fue costeado por el cabildo Catedral, por esos que no hacen nada y ganan de 3 a 3500 pesetas.
¡Cuántos esa noche se acostarían sin cenar en Córdoba!

(*Las Dominicales del libre pensamiento,* 18/08/1898: 4).

La denominación de alguna otra bebida alcohólica tiene una presencia bastante más longeva en el diccionario académico; así sucede, por ejemplo, con *pisco,* que se registra ya en otros productos lexicográficos antes de acceder al repertorio académico. El testimonio más antiguo nos lo ofrece el diccionario de Alemany:

Pisco. *Amer*. En Chile, aguardiente muy bueno de uva, que se fabrica en Pisco y otros lugares del norte. // *Amer*. En Chile, tinaja pequeña de barro en que los productores venden este aguardiente. […] (Alemany 1917).

Pero es el de Pagès, de 1925 («Chile y Perú. Aguardiente superior fabricado en Pisco, lugar peruano. // Perú. Botija en que se exporta este aguardiente»), de donde parecen proceder directamente las definiciones que aparecen en la primera edición del *Diccionario manual* de la Academia:

pisco. m. Chile y Perú. Aguardiente superior fabricado en Pisco, lugar peruano. // Perú. Botija en que se exporta este aguardiente (*DRAEM* 1927).

La definición ha sufrido todo tipo de cambios en estos pocos años, pues del diccionario manual se transfiere al *DRAE* 1936, sin la mención a Perú en la segunda acepción; algo más tarde, en el suplemento al *DRAE* 1970, se eliminan las referencias geográficas también en la primera acepción. En el *DRAE* 1984 se introducen algunas modificaciones en la primera acepción («*pisco*. m. Aguardiente fabricado originalmente en Pisco, lugar peruano. ‖ Botija en que se exporta este aguardiente»), cambios que también se integran en el *DRAEM* 1985 y 1989, aunque en ambas ediciones reaparece la marca Perú en la segunda acepción. En el *DRAE* 1992 se introduce la marca *desusado* en la segunda acepción, que se

elimina en el *DRAE* 2001, al tiempo que se modifica la única que queda y se recurre a la explicación entre paréntesis para exponer el origen de esta voz:

> *pisco*[1]. (De Pisco, ciudad peruana en el departamento de Ica). 1. m. Aguardiente de uva (*DRAE* 2001, *DLE* 2014).

En este caso el *CORDE* ofrece testimonios de finales del XIX y, como podríamos sospechar, aparece primeramente en textos procedentes de la costa americana del Pacífico:

> ¿Qué toma usted? le dijo a uno que entraba a su casa por la primera vez; coñac, italia, pisco? (Juan Montalvo, *Las catilinarias*, 1880–1882, Ecuador).
>
> lo demás, hasta los alfombrados y pipas de ron y de pisco, fue devorado por las llamas (Benjamín Vicuña Mackenna, 1881, *La campaña de Lima, Chile*).
>
> el teniente sacó de la bolsa del pellón una botella de pisco (Clorinda Matto de Turner, *Aves sin nido*, 1889, Perú).

Se extiende a otras áreas americanas ya en el siglo XX («el genio más travieso y mejor catador de pisco», Pastor Servando Obligado, *Tradiciones argentinas*, 1903, Argentina), antes de aparecer en España, ya en 1906, en *La vuelta al mundo en la Numancia* de Pérez Galdós, obra en la que se utiliza varias veces: «volvió Mendaro de la tienda con una botella de pisco y dos de vino del país», «mi compadre Amador, aficionadillo al pisco»; en una de esas apariciones se nos ofrece también una definición del término, todavía muy poco conocido en ese momento en nuestro país:

> Este aguardiente blanco que llamamos pisco, es de vino… cosa buena (Benito Pérez Galdós, *La vuelta al mundo en la Numancia*, 1906).

El *CDH* no ofrece en esta ocasión documentación anterior a la del *CORDE* de esta acepción de la voz[21], pero en cambio las revistas contenidas en la Hemeroteca ofrecen, junto a algún ejemplo anterior no completamente claro[22], el testimonio de una publicación madrileña de 1849:

---

21 Aunque sí del topónimo: «de Lima se traen cada año muy gran cantidad de botijas de vino de Pisco» (Juan Requejo Salcedo, *Relación histórica y geográfica de la provincia de Panamá*, 1640, Panamá).

22 «Mira, Claudio, en lugar de pensar tristemente, bebamos un poco de ese bueno de Pisco, y dejemos correr los sucesos, que el tiempo dirá el resultado» (Adela y Matilde, folletín, por el coronel D.R.S., *El Boletín del ejército*, 09/11/1843: 3).

> hacía una visita de inspección a los 50 o 60 indios que trabajaban todavía por su cuenta, y á quienes pagaba en mercancías, en wisky, en *pisco* (aguardiente del país), de que aquellos desgraciados hacían un espantoso consumo (*El Heraldo*, 10/03/1849: 2).

También en esta ocasión, como puede apreciarse, la palabra va acompañada de una aclaración necesaria para los lectores del periódico. Añadamos solo un ejemplo más, muy informativo:

> se refrescaba, como decía, con una especie de agua fuerte llamada *pisco*, del nombre de un pueblo famoso por el dicho producto, y aunque es detestable, allí piden *pisco* como pudieran pedir coñac, jamaica o kirschen wasser (*Semanario pintoresco español*, 29/04/1855: 2).

No carece de interés tampoco acercarnos a otra bebida cuya denominación entra en los repertorios académicos, como en el caso de *pisco*, a partir de Pagès-1925, cuyo lema se transcribe literalmente en el *Diccionario manual* de 1927 y en el usual de 1936:

> *quianti*. (En it. Chianti, n. p.).1. m. Vino común, pero muy estimado, que se elabora en la Toscana (*DRAEM* 1927, *DRAE* 1936).

La entrada se mantiene sin cambios hasta que en 1992 se produce una supresión en la definición y una ligera modificación en la información etimológica, permaneciendo de este modo hasta la última edición:

> *quianti*. (Del it. Chianti, n. p.).1. m. Vino común que se elabora en Toscana (*DRAE* 1992, *DRAE* 2001, *DLE* 2014).

Si acudimos al *CORDE*, no encontramos ningún testimonio de la grafía castellanizada que ya proponía Pagès, y sí algunos ejemplos escritos a la italiana en una obra de Blasco Ibáñez:

> todo lo que quieras, y un frasco de chianti o dos: cuantos puedas beber […] ¿Conque había bebido chianti? ¡Ah, egoísta! ¡Con tanto que le gustaba! […] ¡Vaya por el chianti! […] Destaparían un frasco de chianti para recordar su vida de Roma (Vicente Blasco Ibáñez, *La maja desnuda*, 1906–1919).

Por esas fechas se documenta en otros escritores españoles («¿Quiere? ¡Vino italiano: chianti! La sirvió», Felipe Trigo, *Jarrapellejos*, 1914) y, algo más tarde, hispanoamericanos («no faltó quien prometiera una botella de vino chianti y esto ya era 'sustancia'» (Miguel Ángel Asturias, *Al Congreso de la Prensa Latina (I)*, 1925).

En esta ocasión el *CDH* no aporta ningún testimonio grafiado *quianti*, y con la forma *chianti* solo hay algunos ejemplos más, entre ellos nuevamente uno de Blasco Ibáñez, ligeramente anterior a los del *CORDE*:

> en aquel país tan calurosamente descrito por Leonora, desde los macarrones del almuerzo y el Chianti en empajada y ventruda redoma, hasta el castellano defectuoso y musical de los dueños del hotel (Vicente Blasco Ibáñez, *Entre naranjos*, 1900).

Pero si no habíamos encontrado ni en el *CORDE* ni en el *CDH* ejemplos de *quianti*, sí podemos obtenerlos gracias a la consulta de la hemeroteca, aunque nunca en gran número:

> No recomendamos a nuestros lectores que coman queso de Gorgonsola, ni que beban vino de Quianti, ni que escuchen *La Traviata* (*Gedeón*, 28/08/1910: 4).

> Los tiempos se ponen feos…/Al Quianti, sigue el Burdeos;/y al *Fascio*, la fracesilla…/La *izquierda* en todo se mete…/(Y es que cambió la *omelette*…)/(Quiero decir: la tortilla). (*Muchas gracias*, 21/06/1924: 3).

O en una curiosa *Noticia* procedente del californiano Monterrey:

> *La carne es flaca. Entran a robar y el vino les hace olvidar su profesión*
> El champaña, el burdeos, el sauternes, el borgoña, el falerno, el quianti, el porto, el rin, el jerez, el coñac, el cointreau, el curaçao, el ron, la ginebra… empezaron a bailar una danza de tentación ante los ojos de los enmascarados que, por efecto de la ley seca, apenas si de tarde en tarde pueden permitirse el lujo de tomarse una copita de whisky de contrabando (*La Voz*, 26/04/1926: 1).

La voz de origen italiano sirve para un interludio humorístico en el diario madrileño *El Sol*:

> *Las cosas de mi Madrid*
>
> Un pollo esdrújulo, de los de mucho fijador en la coquera, le dice al camarero:
> —¿Tienen ustedes "quianti"?
> —El mozo, que es nuevo y le huele la pregunta a pitorreo, contesta rápido:
> —No señor, no tenemos *canti*. Solamente baile; pero luego, más tarde. (*El Sol*, 08/11/1929: 3).

Como puede apreciarse, son ejemplos posteriores a los de *chianti* que habíamos recogido en el *CORDE* y en *CDH*, pero también podemos aportar nuevos ejemplos de la ortografía italianizante muy anteriores, de los que espigamos apenas unos cuantos:

> Se bebieron 600 frascos de Chianti; 2210 botellas de vino de Champagne, y otra porción de vinos extranjeros (*La Regeneración*, 17/07/1871: 4).

> El Barolo, el Chianti, el Barberá y el Moscato corrieron en abundancia» (*El Día*, 06/09/1886: 2).

Entre ellos llama la atención su empleo, en sentido figurado, en un comentario acerca de una novela italiana:

> Bien hubiera yo querido que fuese de mi cosecha el vino simple y puro que conviene beber á los amantes de las letras, en lugar de los adulterados y sobrado alcohólicos licores que les sirven, y acaso me aventurase á vendimiar algún majuelillo de mi caletre para exprimir el zumo de sus vides en un librejo. ¿Pero quién tan sandio qué deje lo cierto por lo dudoso? ¿Y qué más dudoso que la novela propia y qué más cierto que las novelas de Farina?
>
> Aquella —siguiendo el símil enológico—, es uva que ha de dar más raspa y hollejo que vino, y el que diere desabrido ó agrio; mientras que considero la novela del italiano cual legítimo *Chianti* —ya con destreza elaborado y embotellado lindamente— el cual, ni tan ácido como el Burdeos, ni tan ardiente como el *Jerez*, ni tan pesado como el *Oporto*, ni tan estrepitoso y vano como el *Champagne*, es ligero, agradable y digestivo, y á la vez que conforta, lisonjea.
>
> Tales son, en efecto —trasladando al lenguaje natural el lenguaje figurado—, las cualidades propias de las obras de Farina («Un prólogo literario», de Luis Alfonso, reseña de la novela *Cabellos rubios* del italiano Salvatore Farina, *La Época*, 10/09/1886: 1).

Veremos, por último, otro término enológico que solo ha sido incorporado al diccionario académico en este milenio, en este caso una voz de procedencia gallega:

> *albariño*. (Del gall. *albariño*). 1. m. Vino blanco afrutado, originario de Galicia (*DRAE* 2001 y *DLE* 2014).

Este término, a pesar de no haber sido incluido en los repertorios académicos, figuraba en el lemario del segundo intento de redacción del *Diccionario histórico* por parte de la Academia, con dos acepciones, la segunda de las cuales es la que nos interesa en este momento:

> *albariño, ña.* (De *albar* + *-iño.)* m. y f. *Gal.* Variedad de vid de fruto blanco.
>
> 1852 PLANELLAS GIRALT, J. *Flora fanerogámica* 146: ~: Variedad de vid de fruto blanco. 1891 BUEN, ODÓN *Dicc. Hist. Nat.*: ~: Variedad de uva blanca cultivada en Galicia (Valladares). 1915 COTARELO VALLEDOR, A. *Castellano en Galicia* (1927) 103: Albariña (uva): s. f. Albillo; una clase de uvas. 1950 *Dicc. Enciclop. UTEHA* (1953). 1954 OTERO PEDRAYO, R. *Guía de Galicia* 62: Las antiguas castas de uva […], albariño, torrontes, moza fresca, castellano, se han modificado mucho por el injerto. 1966 PÉREZ, J. y ALSINA, R. *Dicc. vinos* 137: Se autoriza la producción de uvas: treixadura, jerez […], albilla, loureiro, a l b a r i ñ o , para los [vinos] blancos. 1968 *Dicc. Agric. Soroa* 46*b*.
>
> 2. Cierta clase de vino blanco peculiar de Galicia.
>
> 1960 CASTROVIEJO, J. M[a]*Galicia* 166: Cayendo ya septiembre se celebra anualmente la gran fiesta del blanco albariño. (*+ 7 del mismo autor.*) 1965 RÍAS BAJAS [*Folleto Dir. Gral. de Turismo*]: Platos y vinos típicos. […] Deben probarse en las Rías otros vinos

> de la región: los blancos «albariños» de Cambados, Ribadumia, Sotomayor, Arbo y el espumoso de El Rosal. 1966 PÉREZ, J. y ALSINA, R. *Dicc. vinos* 139: En Cambados […] se celebra anualmente […] la fiesta del Albariño. El albariño es uno de los mejores vinos del mundo. 1972 *Mundo Gallego* núm. 33,11: Sus vinos: el Albariño, el del Rosal y el del Condado tienen justa fama y son de una excepcional calidad y singular paladar (*DH*-1960–1996).

El *CORDE* no ofrece documentación de *albariño* (tampoco de la posible variante *alvariño*) y tanto el *CDH* como el *CREA* lo localizan por primera vez en 1992, en textos de Xabier Domingo, bastante más tarde, por tanto que los testimonios que recogía el fallido diccionario histórico:

> el vaso de albariño o de rosal, ese vino fino e inteligente, y frío, levemente turbio (Xabier Domingo, *El sabor de España*, 1992).

Pero nuevamente la consulta de la hemeroteca ofrece testimonios más tempranos de esta palabra, comenzando por un texto del gallego Joaquín Pesqueira, publicado en la conocida revista bonaerense *Caras y caretas*, ya en 1924:

> sentados los tres en el banco de piedra, a la sombra dulce del castañar secular, mientras bebíamos unas tazas de rubio albariño, fueron contándome las últimas novedades acontecidas en la mansa quietud de la aldea (Joaquín Pesqueira, «De cómo volví una vez a mi pazo pairal», *Caras y caretas*, 07/06/1924: 33).

En periódicos de la península, contamos con un texto de Luis de Zulueta glosando la recepción en la Academia Española del poeta gallego Ramón Cabanillas:

> Recordaron quizá los versos del propio Ramón Cabanillas, cordial camarada, «romeiro de tantas romerías, festeiro de tantas festas», en los que canta, a la vez que el rojo sangriento del vino «espadeiro» o el matiz dorado del «albariño», la palidez del «loureiro» o el brillo del «caiño» encendido como los rubíes (Luis de Zulueta, «Cabanillas en la Academia: el poeta de Galicia», *El Sol*, 18/06/1929: 3).

También es un escritor destacado quien incluye otra mención de este vino; se trata del gallego Eduardo Blanco-Amor:

> Las botas, damajuanas y pellejuelos empiezan a circular, llenas de ese bravo vino «chirlán» de la orillamar, de albariño rubio o de sosegado y abacial ribero, según los gustos (Eduardo Blanco-Amor, «Escenas de pesca en la costa galaico-portuguesa», *España marítima*, 30/06/1929: 13).

Hay otros textos de esos años, pero solo añadiré un anuncio publicado en la revista madrileña *Galicia en Madrid* (01/06-31/07/1935, 41: 11):

Ilustración 1: *Galicia en Madrid* (01/06-31/07/1935, 41: 11)

Hasta ahora nos hemos ocupado de algunas denominaciones de bebidas alcohólicas que se han incorporado hace relativamente poco tiempo a los diccionarios de referencia en nuestro idioma, por lo que podríamos sospechar que el recurso a las hemerotecas podría ofrecernos información valiosa sobre la historia de estas voces; sin embargo, se trata de un recurso muy útil también en el caso de voces que ya se habían integrado tempranamente en la tradición lexicográfica española, pero de las que carecemos de suficiente documentación. Un buen ejemplo de ello es el de *filipichín*, palabra que ya aparece en la primera propuesta lexicográfica de la Academia, el *Diccionario de Autoridades*:

> FILIPICHIN. s. m. Especie de texido de lana, a modo de chamelotón, que tiene unas labores hechas con prensa. Latín. *Linea tela praelo floribus insignita.* Pragm. de Tass. año 1680. f. 5. Cada vara de *filipichínes* de colores a diez reales (*DA*),

y que se ha mantenido sin interrupción en el diccionario académico hasta nuestros días[23], sin portar ningún tipo de marca en la más reciente edición del *DRAE*:

> *filipichín*. De or. desc. 1. m. Tejido de lana estampado. 2. m. Lechuguino, afeminado (*DLE* 2014).

23 Figura igualmente en la denominada lexicografía extraacadémica: Núñez de Taboada, Salvá, Domínguez, Gaspar y Roig, Zerolo, Toro y Gómez, Pagès, Alemany y Bolufer, Rodríguez Navas…

Es cierto que *Autoridades* ya proporciona la referencia a una documentación de esta palabra en 1680, cuya exactitud hemos podido comprobar, pero las consultas del *CORDE* y del *CDH* coinciden en aportar como primer ejemplo documentado uno de Manuel Bretón de los Herreros:

> Colchas de *filipichin*,/Casacas de filoseda,/Volved al raído cofre/Y á la carcomida percha,/Y con vosotras se encierren/Hasta el dia de la feria/Tantos modernos pecados/Y tantas culpas añejas.

Ambos corpus ofrecen, además, una datación muy poco precisa, pues al haberse servido de la edición de las *Poesías* de este autor (Madrid, Imprenta Miguel Ginesta, 1884), optan por asignar como fecha para el poema un abanico tan amplio como 1828–1870. Pues bien, en este caso la consulta de las hemerotecas nos permite comenzar precisando esta información tan vaga: hemos podido localizar en 1839 el poema «La cuaresma» de Bretón de los Herreros, subtitulado «Romance joco-serio», en la revista *El Panorama, periódico literario que se publica todos los jueves* (1839: 168); además, se nos indica en una nota que el texto había sido «Leído en el Liceo Artístico y Literario en la sesión de competencia del jueves 7 de marzo de 1839».

El resto de los escasos ejemplos allegados por el *CORDE* y el *CDH* son varias décadas posteriores (el *CREA* no recoge ningún testimonio); contamos apenas a finales del XIX con un par de citas del peruano Ricardo Palma, respectivamente de 1874 («Y tanto que vueseñoría la ferió una basquiña de filipichín y un refajo redondo», *Tradiciones peruanas, segunda serie*) y 1883 («para la fiesta del Corpus, una caperuza de filipichín y una falda de angaripola», *Tradiciones peruanas, quinta serie*), y una del español Juan Valera («aunque el damasco era poco, y era más el filipichín que le remeda», *Juanita la Larga*, 1895); ya en el siglo XX, contamos con un único ejemplo procedente de Colombia («falda roja de filipichín, con lunares; camisa de escote», Tomás Carrasquilla, *La marquesa de Yolombó*, 1928).

Son notablemente más antiguos, por tanto, los datos que podemos obtener de la consulta de las hemerotecas, en las que encontramos algún ejemplo ya en 1758:

> Se desaparecieron en días passados dos Guardapieles azules, uno de filipichín, y otro de droguete, ambos alistados, un jubón de grana […] (*Diario noticioso, curioso, erudito y comercial público y económico*, 29/03/1758: 3).

Podemos sumar a este temprano testimonio algunos otros del mismo siglo («tiene de venta una colgadura para cama, de filipichín de exquisito gusto», *Diario curioso, erudito, económico y comercial*, 23/11/1786: 3; «se vende una docena de taburetes con su canapé forrados de filipichín amarillo», *Diario de Madrid*,

18/01/1790: 3; «sargas, para forros, filipichín para cortinas, y otras telas de poco valor», *El Correo mercantil de España y sus Indias*, 04/12/1794 : 6…), mientras otros todavía aparecen en el XIX («unas descoloridas y no nada nuevas cortinas de filipichín, que más que toldo de embarcación parecían cortinas de Sacristía de convento pobre suprimido», *Fr. Gerundio*, 24/07/1840: 6; «sobre unos cojines de filipichín rojo», *El Americano*, 03/02/1873: 9; «un cortinaje de filipichín amarillo», *El Pabellón nacional*, 18/11/1875: 1) sin que podamos allegar documentaciones posteriores de *filipichín* con esta acepción.

No cabe duda, pues, a la vista de lo que hemos podido comprobar en esta última ilustración, de que no solo en el caso del léxico más reciente, sino también en el de voces ya documentadas con anterioridad, acudir a las publicaciones periódicas nos permite refinar considerablemente el conocimiento que poseemos acerca de la historia de nuestro vocabulario.

Referencias bibliográficas

Abad, Francisco (1985): «El lenguaje del siglo XIX. Problemas que plantea su estudio», en *Serta gratulatoria in honorem Juan Régulo*. La Laguna: Universidad de La Laguna, 27–41.

Alemany, José (1917): *Diccionario de la lengua española*. Barcelona: Ramón Sopena.

Alvar Ezquerra, Manuel (1976): *Proyecto de lexicografía española*. Barcelona: Planeta.

Álvarez de Miranda, Pedro (2004): «El léxico del español desde el siglo XVIII hasta hoy», en Rafael Cano (coord.), *Historia de la lengua española*. Barcelona: Ariel, 1037–1064.

Campos Souto, Mar (2007): «Hacia una ordenación morfológica del *NDHE*: primer esbozo», *Verba* 34, 125–155.

Campos Souto, Mar/José Antonio Pascual (2012a): «*Dalle que dalle*: la Filología como intermediaria en el salto de la cantidad a la calidad», en Tomás Jiménez Juliá *et al.* (eds.), *Cum corde et in nova grammatica. Estudios ofrecidos a Guillermo Rojo*. Santiago de Compostela: Universidad de Santiago de Compostela, 183–192.

Campos Souto, Mar/José Antonio Pascual (2012b): «Lexicografía, filología e informática: una alianza imprescindible», en Dolores Corbella *et al.* (eds.), *Lexicografía hispánica del siglo XXI: Nuevos proyectos y perspectivas. Homenaje al profesor Cristóbal Corrales Zumbado*. Madrid: Arco/Libros, 151–170.

Campos Souto, Mar (2015): «El *NDHE* como muestra de la nueva lexicografía digital», *Estudios de Lexicografía* 3 (monográfico sobre el *NDHE*), 71–93.

Campos Souto, Mar (2016): «Léxico del futuro para la lengua del pasado», en Rosalía Cotelo García (coord.), *Entre dos coordenadas: la perspectiva diacrónica y diatópica en los estudios léxicos del español*. San Millán de la Cogolla: Cilengua, 33–72.

CDH = Instituto de Investigación Rafael Lapesa de la Real Academia Española (2013): *Corpus del Nuevo diccionario histórico (CDH)* [en línea]. <http://web.frl.es/CNDHE> [último acceso: 14/12/2016].

CORDE = Real Academia Española: Banco de datos (*CORDE*) [en línea]. *Corpus diacrónico del español*. <http://www.rae.es> [último acceso: 14/12/2016].

Cotelo García, Rosalía (2014): *Vocabulario de la indumentaria en la Edad de Plata*. La Coruña: Universidade de A Coruña.

CREA = Real Academia Española: Banco de datos (*CREA*) [en línea]. *Corpus de referencia del español actual*. <http://www.rae.es> [último acceso: 14/12/2016].

DCEC = Corominas, Juan (1955–1957): *Diccionario crítico etimológico de la lengua castellana*, 4 vols. Madrid: Gredos.

DECH = Corominas, Joan/José Antonio Pascual (1980–1991): *Diccionario crítico etimológico castellano e hispánico*, 6 vols. Madrid: Gredos.

DLE = Real Academia Española (2014): *Diccionario de la lengua española*. Madrid: Espasa. <http://dle.rae.es> [último acceso: 20/02/2017].

DRAE = Real Academia Española (2001): *Diccionario de la lengua española*. Madrid: Espasa. <http://lema.rae.es/drae2001> [último acceso 20/02/2017; si se trata de alguna de las ediciones anteriores se precisa en cada caso].

DRAEM = Real Academia Española (1989[5]): *Diccionario de la lengua española*. Madrid: Espasa-Calpe. [Si se trata de alguna de las ediciones anteriores se precisa en cada caso].

Gaspar y Roig (1853-1855): *Diccionario enciclopédico de la lengua española*. Madrid: Imprenta y Librería de Gaspar y Roig.

Lapesa, Rafael (1992): *Léxico e historia. II. Diccionarios*. Madrid: Istmo.

Lapesa, Rafael (1994): «La lengua», en Pedro Laín Entralgo *et al.* (eds.), *La Edad de Plata de la cultura española (1898–1936). Volumen II. Letras. Ciencia. Arte. Sociedad y Culturas*. Madrid: Espasa, 5–40.

Lapesa, Rafael (1996): *El español moderno y contemporáneo*. Barcelona: Crítica.

NDHE = Instituto de Investigación Rafael Lapesa de la Real Academia Española: *Nuevo diccionario histórico del español* [en línea]. <http://

www.rae.es/recursos/diccionarios/nuevo-diccionario-historico> [último acceso: 14/12/2016].

NTLLE = Real Academia Española: *Nuevo Tesoro Lexicográfico de la Lengua Española* [en línea]. <http://www.rae.es/recursos/diccionarios/diccionarios-anteriores-1726-1992/nuevo-tesoro-lexicografico> [último acceso: 14/12/2016].

Octavio de Toledo y Huerta, Álvaro S. (2016): «Sin *CORDE* pero con red: *algotras* fuentes de datos», *Revista Internacional de Lingüística Iberoamericana* XIV, 2, 19–48.

Pagès, Aniceto de (1925): *Gran diccionario de la lengua castellana*. Barcelona: Fomento Comercial del Libro, tomo IV.

Pascual, José Antonio (2015): «Introducción a una celebración lexicográfica: a propósito de la reciente publicación de un millar de palabras del *NDHE* en el portal de la RAE», *Estudios de Lexicografía* 3 (monográfico sobre el *NDHE*), 7–13.

Pascual, José Antonio/Mar Campos Souto (2014): «La morfología en el *NDHE*», en Bruno Camus Bergareche (ed.), *Morfología y diccionarios*. La Coruña: Universidad de A Coruña, 125–150.

Pérez Pascual, José Ignacio (2012): «El léxico de especialidad», en Luis Luque Toro, José Francisco Medina Montero y Rocío Luque (eds.), *Léxico español actual III*. Venecia: Università Ca'Foscari, 203–233.

Pérez Pascual, José Ignacio (2013): «El vocabulario médico en el XIX, entre la prensa y el diccionario», en Gloria Clavería *et al*. (eds.), *Historia, lengua y ciencia: una red de relaciones*. Fráncfort: Peter Lang, 199–216.

Pérez Pascual, José Ignacio (2014): «El décalage entre la sustitución de disciplinas y los cambios del léxico correspondiente a ellas en los diccionarios. A propósito de *albeitería* y *veterinaria*», *Cahiers de Lexicologie* 104, 197–218.

Pérez Pascual, José Ignacio (en prensa): «Nuevas herramientas y viejos saberes», en Pilar Garcés Gómez (ed.), *Perspectivas teóricas y metodológicas para la elaboración de un diccionario histórico*.

Pinillos Laffon, Octavio (2015): «Los materiales de redacción del *NDHE*», *Estudios de Lexicografía* 3 (monográfico sobre el *NDHE*), 95–116.

Porto Dapena, José-Álvaro (2000): «Diccionarios históricos y etimológicos del español», en Ignacio Ahumada (ed.), *Cinco siglos de lexicografía del español. IV Seminario de Lexicografía Hispánica*, Jaén: Universidad de Jaén, 103–125.

Real Academia Española (1960–1996): *Diccionario histórico de la lengua española*. Madrid: Real Academia Española. <http://web.frl.es/DH.html> [último acceso: 14/12/2016].

Rodríguez Navas, Manuel (1918): *Diccionario general y técnico hispanoamericano*. Madrid: Cultura Hispanoamericana.

Salas Quesada, Pilar/Abelardo Torres Morcillo (2011): «*ARDIDEs*: Aplicación de Redacción de un Diccionario Diacrónico del Español», *Revista de Lexicografía* XVII, 133–159.

Salas Quesada, Pilar/Abelardo Torres Morcillo (2015): «Aproximación a los fundamentos del *NDHE* a través de las herramientas informáticas usadas en su elaboración y presentación», *Estudios de Lexicografía* 3 (monográfico sobre el *NDHE*), 15–69.

Salvador, Gregorio (1983): *Semántica y lexicología del español*. Madrid: Paraninfo.

Seco, Manuel (2003²): *Estudios de lexicografía española*. Madrid: Gredos.

Seco, Manuel (2007): «El nuevo léxico», en Mar Campos Souto, Mª Montserrat Muriano Rodríguez y José Ignacio Pérez Pascual (coords.), *El nuevo léxico*. La Coruña: Universidad de A Coruña, 9–26.

Senabre, Ricardo (1999): *Capítulos de historia de la lengua literaria*. Cáceres: Universidad de Cáceres.

Toro y Gómez, Miguel (1901): *Nuevo diccionario enciclopédico ilustrado de la lengua castellana*. París-Madrid: Librería Armand Colin-Hernando y Cía.

Zerolo, Elías *et al.* (1895): *Diccionario enciclopédico de la lengua castellana*. París: Garnier Hermanos.

STUDIA ROMANICA ET LINGUISTICA

curant Daniel Jacob, Elmar Schafroth, Edeltraud Werner,
Araceli López Serena, André Thibault et Manuela Caterina Moroni

Band 1 Michael Metzeltin: Die Sprache der ältesten Fassungen des Libre de Amich e Amat. Untersuchungen zur kontrastiven Graphetik, Phonetik und Morphologie des Katalanischen und des Provenzalischen.

Band 2 Paul Miron: Aspekte der lexikalischen Kreativität im Rumänischen.

Band 3 Paul Miron: Der Wortschatz Dimitrie Cantemirs.

Band 4 Peter Wunderli: Valéry saussurien. Zur linguistischen Fragestellung bei Paul Valéry.

Band 5 Ekkehard Zöfgen: Strukturelle Sprachwissenschaft und Semantik. Sprach- und wissenschaftstheoretische Probleme strukturalistisch geprägter Bedeutungsforschung (dargestellt am Beispiel des Französischen).

Band 6 Marianne Wigger: Tempora in Chrétiens «Yvain». Eine textlinguistische Untersuchung.

Band 7 Christoph Strosetzki: Konversation. Ein Kapitel gesellschaftlicher und literarischer Pragmatik im Frankreich des 17. Jahrhunderts. Vergriffen.

Band 8 Maria Iliescu: Grundwortschatz Rumänisch (Deutsch-Englisch-Französisch).

Band 9 Hartmut Rentsch: Determinatoren für den Modusgebrauch im Neufranzösischen aus generativer Sicht.

Band 10 Alberto Zuluaga: Introducción al estudio de las expresiones fijas.

Band 11 Edeltraud Werner: Die Verbalperiphrase im Mittelfranzösischen.

Band 12 Wolfgang Rettig: Sprachliche Motivation. Zeichenrelationen von Lautform und Bedeutung am Beispiel französischer Lexikoneinheiten.

Band 13 Petra M.E. Braselmann: Konnotation - Verstehen - Stil. Operationalisierung sprachlicher Wirkungsmechanismen dargestellt an Lehnelementen im Werke Maurice Dekobras.

Band 14 Angela Karasch: Passiv und passivische Diathese im Französischen und Deutschen.

Band 15 Peter Wunderli/Wulf Müller (Hrsg.): Romania historica et Romania hodierna. Festschrift für Olaf Deutschmann zum 70. Geburtstag, 14. März 1982.

Band 16 Renate Tretzel: Glauben heißt nicht immer Wissen. Der Konjunktiv in abhängigen Subjekt- und Objektsätzen.

Band 17 Thomas Krefeld: Das französische Gerichtsurteil in linguistischer Sicht. Zwischen Fach- und Standessprache.

Band 18 Gudrun Krassin: Das Wortfeld der Fortbewegungsverben im modernen Französisch.

Band 19 Brigitte Nerlich: La pragmatique. Tradition ou révolution dans l'histoire de la linguistique française?

Band 20 Olaf Deutschmann: Ungeschriebene Dichtung in Spanien.

Band 21 Rudolf Windisch: Zum Sprachwandel. Von den Junggrammatikern zu Labov.

Band 22 Christoph Strosetzki: Konversation und Literatur. Zu Regeln der Rhetorik und Rezeption in Spanien und Frankreich.

Band 23 Gabriele Berardi: Studien zur Saussure-Rezeption in Italien.

Band 24 Peter Wunderli: Principes de diachronie. Contribution à l'exégèse du «Cours de linguistique générale» de Ferdinand de Saussure.

Band 25 Graciela E. Vázquez: Análisis de errores y aprendizaje de español / lengua extranjera. Análisis, explicación y terapia de errores transitorios y fosilizables en el proceso de aprendizaje de español como lengua extranjera en cursos universitarios para hablantes nativos de alemán.

Band 26 Andreas Gather: Formen referierter Rede. Eine Beschreibung kognitiver, grammatischer, pragmatischer und äußerungslinguistischer Aspekte.

Band 27 Anne-Marie Spanoghe: La syntaxe de l'appartenance inaliénable en français, en espagnol et en portugais.

Band 28 Kerstin Störl-Stroyny: Kausalität. Die Entwicklung des Ausdrucks von Kausalität im Spanischen.

Band 29 Ildikó Koch: Die Metataxe im deutsch-italienischen Sprachvergleich. Eine Studie der verbbedingten Abweichungen im Satzbau.

Band 30 Uta Schmitt: Diskurspragmatik und Syntax. Die Funktionale Satzperspektive in der französischen und deutschen Tagespresse unter Berücksichtigung einzelsprachlicher, pressetyp- und textklassenabhängiger Spezifika.

Band 31 Gabriele Kaps: Zweisprachigkeit im paraliturgischen Text des Mittelalters.

Band 32 Karin Ewert-Kling: *Left Detachment* und *Right Detachment* im gesprochenen Französischen und Spanischen. Eine formale und funktionale Analyse mit einem Ausblick auf Grammatikalisierungstendenzen.

Band 33 Andreas Dufter / Daniel Jacob: Syntaxe, structure informationnelle et organisation du discours dans les langues romanes.

Band 34 Maria Selig / Gerald Bernhard (Hrsg.): Sprachliche Dynamiken. Das Italienische in Geschichte und Gegenwart.

Band 35 Elmar Schafroth / Maria Selig (Hrsg.): *Testo e ritmi*. Zum Rhythmus in der italienischen Sprache.

Band 36 Valeriano Bellosta von Colbe / Marco García García (eds.): Aspectualidad – Transitividad – Referencialidad. Las lenguas románicas en contraste.

Band 37 Daniel Jacob / Katja Ploog (éds.): Autour de *que* - El entorno de *que*.

Band 38 Ursula Reutner / Elmar Schafroth (eds./cur./éds.): Political Correctness. Aspectos políticos, sociales, literarios y mediáticos de la censura lingüística. Aspetti politici, sociali, letterari e mediatici della censura linguistica. Aspects politiques, sociaux, littéraires et médiatiques de la censure linguistique.

Band 39 Sabine De Knop / Fabio Mollica / Julia Kuhn (Hrsg.): Konstruktionsgrammatik in den romanischen Sprachen.

Band 40 Ludwig Fesenmeier / Sabine Heinemann / Federico Vicario (Hrsg./a cura di): Sprachminderheiten: gestern, heute, morgen. Minoranze linguistiche: ieri, oggi, domani. 2014.

Band 41 Mathias Arden: Inszenierte und elaborierte Mündlichkeit bei TV Globo. Zur soziostilistischen Modellierung morphosyntaktischer Variablen des brasilianischen Portugiesisch. 2015.

Band 42 Elmar Schafroth / Maria Selig (a cura di/Hrsg.): La lingua italiana dal Risorgimento a oggi. Das Italienische nach 1861. Unità nazionale e storia linguistica. Nationale Einigung und italienische Sprachgeschichte. In collaborazione con/ In Zusammenarbeit mit Nora Wirtz. 2014.

Band 43 Romana Castro Zambrano: Diskursanalyse und mentale Prozesse. Sprachliche Strategien zur diskursiven Konstruktion nationaler Identität bei Hugo Chávez und Evo Morales. 2015.

Band 44 Anna-Maria De Cesare / Davide Garassino (eds.): Current Issues in Italian, Romance and Germanic Non-canonical Word Orders. Syntax – Information Structure – Discourse Organization. 2016.

Band 45 Martin Becker / Ludwig Fesenmeier (a cura di): Relazioni linguistiche. Strutture, rapporti, genealogie. 2016.

Band 46 Carlota de Benito Moreno / Álvaro S. Octavio de Toledo y Huerta (eds.): En torno a 'haber'. Construcciones, usos y variación desde el latín hasta la actualidad. 2016.

Band 47 Marta Fernández Alcaide / Elena Leal Abad / Álvaro S. Octavio de Toledo y Huerta (eds.): En la estela del Quijote. Cambio lingüístico, normas y tradiciones discursivas en el siglo XVII. 2016.

Band 48 Vivian Pereira-Koschorreck: Kontaktanzeigen kontrastiv. Französische und deutsche Kontaktanzeigen im diachronen und synchronen Vergleich. 2016.

Band 49 Ulrike Kolbinger: Indigene Schreiber im kolonialen Peru. Zur juristisch-administrativen Textproduktion im Jauja-Tal (16. und 17. Jahrhundert). 2017.

Band 50 Gabriela Cruz Volio: Actos de habla y modulación discursiva en español medieval. Representaciones de (des)cortesía verbal histórica. 2017.

Band 51 Daniela Pietrini: Sprache und Gesellschaft im Wandel. Eine diskursiv basierte Semantik der ‚Familie' im Gegenwartsfranzösischen am Beispiel der Presse. 2017.

Band 52 María Teresa Echenique Elizondo; Angela Schrott; Francisco Pedro Pla Colomer (eds.). Cómo se "hacen" las unidades fraseológicas: continuidad y renovación en la diacronía del espacio castellano. 2018.

Band 53 Dolores Corbella / Alejandro Fajardo / Jutta Langenbacher-Liebgott (eds.): Historia del léxico español y Humanidades digitales. 2018

www.peterlang.com