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.

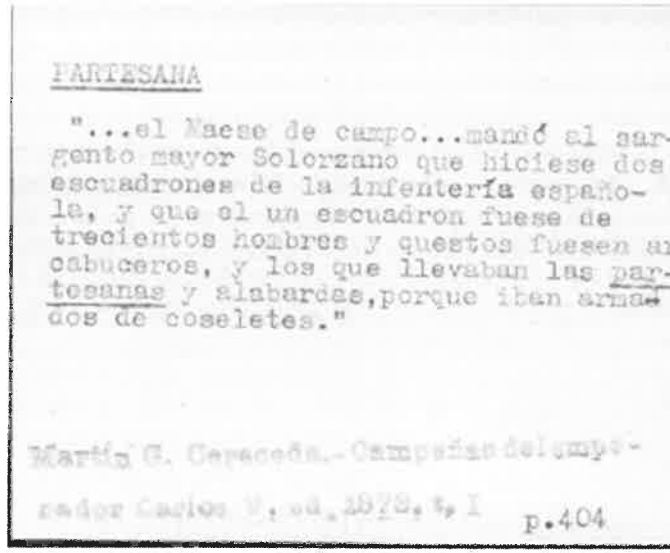
El protocolo establecido en el *NDHE* determina que, cuando se localiza un testimonio de interés en el *FG*, se debe intentar consultar directamente la edición citada en la cédula o, incluso, el manuscrito original[7]. Así sucede en el caso

---

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. Cepeda.- 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 lexi-:ográficos restringidos, por lo que abundan cédulas en que se recogen datos pro-:edentes de diccionarios dialectales o vocabularios de lenguas de especialidad. Este hecho explica que la primera documentación de *güiro*, como sinónimo de *:abeza*, 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 :uentes lexicográficas restringidas), el *FG* continúa siendo un filón inagotable, ɔues en él se incluyen repertorios de tanta relevancia como el *Vocabulario mate-nático-etimológico* de F. Picatoste (1862), el *Diccionario militar* de J. Fernán-dez Mancheño (1822), el *Vocabulario de mexicanismos* de J. García Icazbalceta (1894), el *Vocabulario andaluz* (1933) de A. Alcalá Venceslada, los *Hondureñis-mos* (1895) de A. Membreño o el *Vocabulario cubano* (1859) de J. García Arbo-leya, 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 ver-sión electrónica del *FG*, precisamente el propósito que perseguía el proyecto inacabado que se emprendió en 1993. Posteriormente, en el año 2006, la Aca-demia 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 per-mitiría 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	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	106.99	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	87.76	6.08
luz	sustantivo	36	5.97	99.39	6.0
norma	sustantivo	35	8.51	149.39	5.91
cosa	sustantivo	35	3.56	47.68	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	8.06	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

Periodo	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

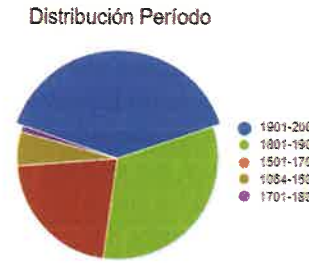


Gráfico 1: Distribución de ocurrencias por período en la extensión diacrónica del *CDH*

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.

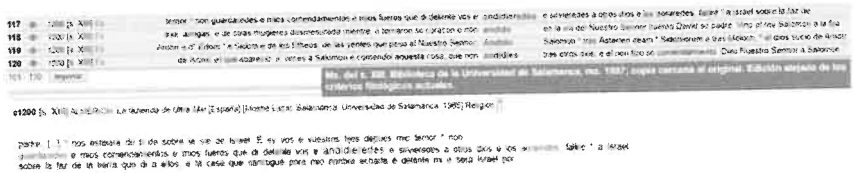


Ilustración 3: Muestra de la breve caracterización filológica de los textos medievales del *CDH* nuclear

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).

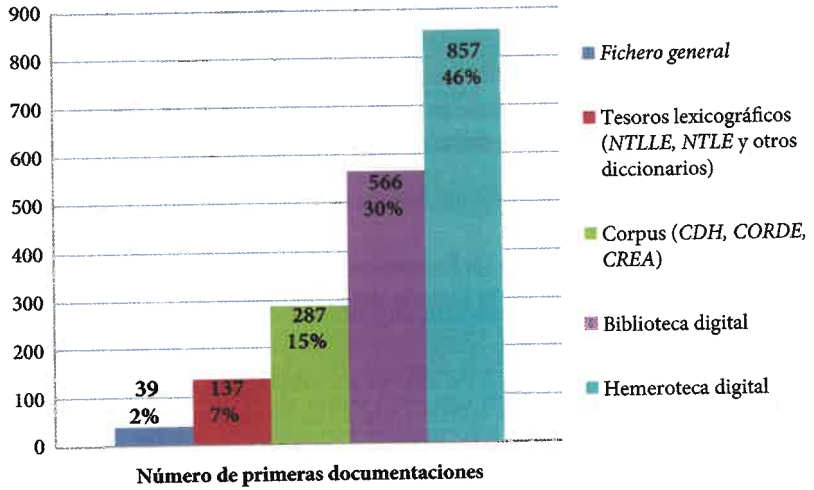


Gráfico 2: Procedencia de los primeros testimonios de las acepciones léxicas posteriores a 1700

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.

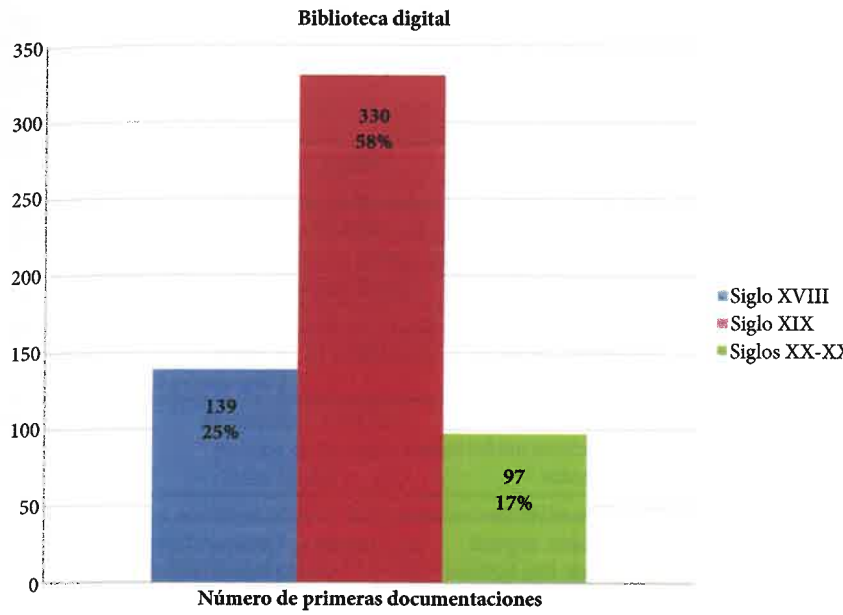


Gráfico 3: Distribución, por siglos, de los primeros testimonios con fuente Biblioteca digital

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.

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