3

Corpus diacrónicos del español en las Américas

(Diachronic corpora on the Spanish of the Americas)

Virginia Bertolotti y Concepción Company Company

1. Introducción

Hasta hace relativamente poco, el español americano estaba infrarrepresentado en el conjunto de trabajos, sincrónicos y diacrónicos, de la lengua española. Desde los años 90 del siglo pasado, equipos de investigadores de diversos países empezaron a subsanar esta carencia, y rescataron documentación, escrita en español, proveniente de archivos históricos, con el objetivo de, por un lado, construir corpus que permitieran crear infraestructura para ahondar en la historia lingüística de los países americanos y avanzar en una reconstrucción sistemática de la historia del español en sus países, y, por otro, comprender mejor las dinámicas del cambio lingüístico con nuevas evidencias desde la lengua española. Este desarrollo de la lingüística basada en corpus diacrónicos americanos, publicados en su momento en papel, tiene un quiebre relevante a partir de la aparición del primer corpus informatizado abarcador de la diatopía y diacronía de todo el español de América: el *Corpus Diacrónico y Diatópico del Español de América* (CORDIAM), que conjunta, en una herramienta digital amigable y de libre acceso (www.cordiam.org), esos valiosos y fragmentados aportes, además de que contiene obras literarias y textos tomados de la prensa. El CORDIAM está conformado por tres subcorpus: CORDIAM-Documentos, CORDIAM-Literatura y CORDIAM-Prensa.

Este capítulo presenta, en primer lugar, el estado de la cuestión de los corpus diacrónicos del español y el grado de representación de testimonios americanos en ellos. En segundo lugar, se realizan algunas reflexiones teórico-disciplinares y metodológicas subyacentes al diseño del CORDIAM. En tercer lugar, se subrayan cinco de sus características: (i) la robustez y fiabilidad de sus datos, (ii) las prestaciones de su interfaz, (iii) el desarrollo de metadatos asociados a cada texto, (iv) el desarrollo de una tipología textual *ad hoc* y (v) su carácter colaborativo. Cerramos con algunas reflexiones sobre tareas pendientes y líneas a seguir.

Palabras clave: español de América; archivos; literatura; prensa; filología; lingüística histórica; CORDIAM

Until recently, American Spanish was underrepresented in the body of synchronic and diachronic Spanish language studies. In the 1990s, researchers from several countries started addressing this gap, drawing on archives to build a corpus that would provide input for

DOI: 10.4324/9780429329296-5

exploring Latin America's linguistic history and advance a systematic historical reconstruction of Spanish in those countries, while enhancing the understanding of linguistic change dynamics with new evidence from Spanish language sources. This linguistic development based on American diachronic corpuses, originally on paper, hit a major milestone with the first computerized corpus covering the full diatopic and diachronic range of American Spanish: the *Corpus Diacrónico y Diatópico del Español de América* (CORDIAM), a user-friendly open-access digital platform (www.cordiam.org) that gathers these valuable and fragmented contributions, along with press and literature, in three subcorpora: CORDIAM-Documents; CORDIAM- Literature; and CORDIAM- Press.

This chapter outlines the state of diachronic Spanish corpuses and the extent to which American texts are represented therein, addresses some underlying theoretical-disciplinary and methodological aspects of their design, and highlights five characteristics of the CORDIAM: (i) solid and reliable data, (ii) interface features, (iii) inclusion of metadata for each text, (iv) development of an *ad hoc* textual typology, and (v) collaborative nature. In conclusion, we discuss pending tasks and future lines of work.

Keywords: American Spanish; archives; press; literatura; philology; historical linguistics; CORDIAM

2. Conceptos fundamentales y estado de la cuestión

La lingüística histórica es, cosa sabida, una disciplina que está obligada a trabajar con corpus, "lingüística con corpus", pero no es, necesariamente, "lingüística de corpus" (Kabatek 2016, 3). La tendencia actual, cada vez mayor, es conjuntar lingüística histórica, con corpus, y lingüística de corpus, de manera que la primera se enriquece de la segunda y esta logra avances informáticos gracias, en buena parte, a nuevas evidencias y modos de acercarse al objeto de estudio, motivados, a su vez, por el ordenamiento de los textos con lógica informática. La lengua española, como muchas otras, refleja cabalmente este nuevo modo de acercarse a la lingüística histórica.

La lingüística histórica del español realizada con base en corpus históricos electrónicos tiene apenas tres décadas. Hasta donde tenemos noticia, existen nueve corpus electrónicos diacrónicos para el español, de acceso libre, que tienen mayor, menor o nula representación del español americano. Son los siguientes.[1]

CORDE. La ausencia de datos del español accesibles mediante corpus electrónicos diacrónicos fue, en efecto, un problema para los historiadores de la lengua hasta los años 90 del siglo pasado, en que aparece, a iniciativa de la Real Academia Española de la Lengua, el *Corpus Diacrónico del Español* (CORDE) (http://corpus.rae.es/cordenet.html), coordinado por Guillermo Rojo. Este corpus, que abarca nueve siglos, 1150–1975, cambió, sin duda, el modo de hacer lingüística histórica y de obtener generalizaciones y datos refinados. El CORDE, no obstante, su carácter de corpus pionero, tiene un bajísimo porcentaje de textos americanos, menos de 8%, considerando solamente el conjunto textual a partir del siglo XVI. Esta escasa presencia americana es explicable, en parte, porque en las fechas de su creación no existían suficientes ediciones críticas de documentos americanos y, en parte también, porque su objetivo era poner a disposición del estudioso obras para una diacronía de la lengua española, sin importar, posiblemente, la procedencia dialectal de los textos. En la actualidad, existen otros corpus digitales y accesibles a través de la red que se caracterizan, tal como el CORDE, por la poca robustez de los datos americanos —entendiendo por América lo que actualmente conocemos como América del Sur, Central, Caribe y del Norte— aunque, en términos generales, algo ha aumentado la presencia textual americana, sobre todo para el siglo XX.

CREA. En la medida en que incluye ya más de una generación anterior de hablantes, el *Corpus de Referencia del Español Actual* (CREA), que abarca un lapso muy breve, el último cuarto del siglo XX y apenas los primeros años del XXI en versalitas (http://corpus.rae.es/creanet.html), puede ser considerado histórico. Somos conscientes de que esta caracterización puede ser cuestionable; no obstante, entendemos que una buena parte de los textos del CREA son ya históricos, en tanto que cumplen con un criterio usual para considerarlos como tales, a saber, tener una antigüedad mayor a 30 años. El CREA también es iniciativa de la Real Academia Española y está coordinado, asimismo, por Guillermo Rojo. Este corpus contiene textos escritos e incorpora transcripciones de textos orales, tanto de alta inmediatez comunicativa como de la distancia comunicativa (Oesterreicher 1996; López Serena 2007). En él, la representación de textos de origen americano es mucho mayor, un 50% del total de los textos incluidos.

CNDHE. El *Corpus del Nuevo Diccionario Histórico del Español* (CNDHE), también de la Real Academia Española (http://web.frl.es/CNDHE/) (con la colaboración de la Asociación de Academias de la Lengua Española) es la base para el futuro y esperado diccionario académico *Diccionario histórico de la lengua española*. Contiene 62 millones de palabras, 38 de España y 24 de América, 38% de textos americanos, por lo tanto. A este corpus nuclear se le ha incorporado el CORDE y el CREA, a través de la interfaz del *Corpus del Español del Siglo xxi* (CORPES~~-xxi~~) (http://web.frl.es/CORPES/).

CHARTA. La comunidad académica cuenta también con el *Corpus Hispánico y Americano en la Red: Textos Antiguos* (CHARTA), radicado en la Universidad de Alcalá de Henares (www.corpuscharta.es/), coordinado por Pedro Sánchez-Prieto Borja y alimentado con la contribución de investigadores de diversas instituciones. Se trata de un corpus de 2.000 documentos que reúne textos archivísticos en español de los siglos XII al XIX, que tiene una gran calidad filológica ya que contiene el facsímil, la transcripción paleográfica estrecha y la edición crítica para cada documento y cuenta con una interfaz de búsqueda. En él, la presencia del español de América es inferior a 7%.

CODEA. Con similar arquitectura a la del CHARTA, el *Corpus de Documentos Españoles Anteriores a 1700* (CODEA + *2015*) (http://corpuscodea.es/), confeccionado por investigadores de varias universidades y coordinado también por Pedro Sánchez-Prieto Borja, contiene poco más de 2.500 documentos de archivo y textos literarios, que incluyen algunos géneros literarios menores. De altísima fidelidad filológica y ecdótica, carece de textos americanos.

CE. El *Corpus del Español* (CE) también conocido como *Corpus Davies*, ya que Mark Davies lo construyó y lo dirige~~,~~ (www.corpusdelespanol.org/), abarca de los siglos XIII al XXI y se distingue por estar constituido por un universo de palabras gigante, casi dos billones, lo cual lo hace el mayor corpus del español con textos históricos. El subcorpus histórico del *Corpus Davies*, hasta el siglo XIX inclusive, cuenta con 80 millones de palabras, con algunos documentos de archivo, y está constituido, mayoritariamente, por documentos de prensa, siglos XIX y XX, y por obras de literatura a partir del siglo XIII. Su foco no es la calidad filológica y ecdótica, sino poner a disposición inmediata millones de formas para cualquier periodo de la lengua española con variedad dialectal y textual. La representación de textos americanos en el CE, en el periodo XVI—XIX, es de 17%.

POSTSCRIPTUM. El corpus *Post Scriptum. Arquivo Digital de Escrita Cotidiana em Portugal e Espanha na Época Moderna* (POSTSCRIPTUM), alojado en la Universidade de Lisboa~~,~~ (http://ps.clul.ul.pt/)~~,~~ bajo la dirección de Rita Marquilhas, aunque no es estrictamente de lengua española, tiene el interés y especificidad de contener solo cartas, algunas de gran inmediatez comunicativa, lo cual otorga un interés especial a este corpus en cuanto a la posible

reconstrucción de la oralidad de estados pretéritos de la lengua, porque, como es sabido, la carta es el único tipo de texto que se atreve a escribir quien realmente no sabe escribir (Company 2001). Estas "manos inhábiles" (Marquilhas 2000), para emplear la terminología acuñada por la coordinadora de este corpus, están muy bien representadas en POSTSCRIPTUM, corpus que contiene algunos textos escritos por portugueses y por españoles en América.

BIBLIAS. Del mayor interés para estudios históricos, el *Corpus de Biblias. Biblia Medieval*, de la Universitat de les Iles Balears y del Centro de Lenguas de San Millán de la Cogolla (www.bibliamedieval.es/), coordinado por Andrés Enrique Arias, es un corpus refinadísimo en ecdótica y en motor de búsqueda, ya que permite el acoplamiento en línea de cualquier búsqueda en todas las traducciones de la Biblia al español y en otras tradiciones textuales bíblicas en la historia del español. Hasta el momento, como el nombre del corpus indica, no contiene testimonios bíblicos americanos.

Algunos de los corpus mencionados arriba cuentan, como ya señalamos, con textos americanos. Sin embargo, basta consultar obras de referencia colectivas de envergadura para percatarse o bien de la falta de datos americanos en ellas o bien de la escasa representatividad de los existentes. Tal es el caso de la *Gramática descriptiva de la lengua española* (Bosque y Demonte dirs. 1999), de la *Sintaxis histórica de la lengua española* (Company dir. 2006b, 2009, 2014) o de la *Enciclopedia de lingüística hispánica* (-Gutiérrez-Rexach ed. 2016).

CORDIAM. El panorama anterior de ausencia de documentación americana en calidad y cantidad suficientes~~,~~ como para poder completar la historia de la lengua española en un capítulo muy amplio en el tiempo, cinco siglos, y muy extenso geográficamente —los actuales 19 países de Hispanoamérica, más algunos otros países americanos en que antiguamente se habló y escribió en español— motivó a las autoras de este capítulo a construir, a partir de 2012, el *Corpus Diacrónico y Diatópico del Español de América* (CORDIAM) (www.cordiam.org). Este corpus, radicado en la Academia Mexicana de la Lengua (AML), se abre a la comunidad académica en noviembre de 2015, en ocasión del *XV Congreso de la Asociación de Academias de la Lengua Española* (ASALE); cuatro años después, en noviembre del 2019, en el *XVI Congreso* de la ASALE, esta asociación lo acoge por unanimidad, y es hoy un programa AML-ASALE.

El CORDIAM se creó apoyado en una idea muy simple: poner a disposición, con una interfaz y diseño amables, un conjunto de corpus ya existentes de documentos tomados de archivos, transcriptos con una alta fidelidad filológica. Estos eran poco accesibles a los investigadores por estar impresos en papel y por tener una baja circulación, ya que estaban publicados mayoritariamente por editoriales universitarias. Se distingue claramente de los anteriores por su alta especificidad, a la vez que los complementa. Se trata del único corpus electrónico histórico que solo contiene textos americanos, escritos en América y en su gran mayoría por hispanohablantes nativos americanos. Abarca, como ya dijimos, los 19 países de la actual Hispanoamérica más otros cinco países americanos – Belice, EE.UU., Guyana, Jamaica y Trinidad y Tobago –, en donde se habló y escribió en español, ya que formaban parte del imperio español. Temporalmente, se extiende de 1494 a 1905, es decir, tiene una profundidad histórica de 400 años.

El CORDIAM, respecto de su fecha de apertura en 2015, ha crecido sustancialmente, tanto cuantitativa como cualitativamente. Cuantitativamente, se han incorporado ya casi todos los textos históricos americanos transcriptos con fines lingüísticos disponibles, varios miles de documentos, y a ellos se han sumado varios cientos de documentos transcriptos específicamente con la finalidad de ser incorporados al CORDIAM. Cualitativamente, el CORDIAM ha crecido incorporando dos nuevos ámbitos textuales, obras de literatura y textos de prensa, los cuales suman, a la fecha, otros varios miles de unidades textuales. La incorporación de

literatura y de prensa se ha realizado bajo el mismo requisito de mantener calidad filológica y ecdótica, es decir, los textos tienen un apego casi total al original y la manipulación realizada es la mínima necesaria para garantizar su procesabilidad informática.

Solo se suben al CORDIAM textos que han sido paleografiados o revisados por especialistas en lingüística histórica o filología, ya que un imperativo absoluto de este corpus es incorporar materiales editados con una estricta óptica ecdótica. En el caso de los documentos tomados de archivos, no incorporamos textos transcriptos por historiadores, que suelen concentrase en el contenido y no en la forma, a no ser que hayamos corroborado la fidelidad de la transcripción al original. En el caso de los textos de prensa, los convertimos a procesador de texto y cotejamos la conversión con los originales. En cuanto a los textos literarios, cuando el equipo del CORDIAM considera que no existe una edición crítica fidedigna de una determinada obra, opta por la edición príncipe, tal es el caso de los *Comentarios reales* y la *Historia general del Perú* del Inca Garcilaso de la Vega. En el caso de varias ediciones críticas, se opta por la más conservadora.

El proceso de selección y edición de los textos es, asimismo, controlado y coordinado por especialistas. En el caso de la prensa, Magdalena Coll (Universidad de la República y Academia Nacional de Letras, Uruguay) es la coordinadora de este subcorpus. En el caso de las obras literarias, son seleccionadas y editadas bajo la coordinación de Jorge Gutiérrez Reyna (Universidad Nacional Autónoma de México). La supervisión de la sistematización de documentos de archivo corresponde a las autoras de este capítulo. El desarrollo del motor de búsqueda del CORDIAM y sus sucesivas remodelaciones y actualizaciones están a cargo de los especialistas en procesamiento del lenguaje natural Alexander Gelbukh y Grigori Sidorov (Instituto Politécnico Nacional, México).

El CORDIAM al día de hoy (20-10-2021) –20) cuenta con 15669907 unidades textuales, que suman casi doce millones de palabras (11.749.864 9.644.566), repartidas en las siguientes cantidades textuales: CORDIAM-Documentos, 5.703 4.969; CORDIAM-Literatura, 3.173 2.436; CORDIAM-Prensa, 6.973 5.525, como muestra la Tabla 3.1, que incluye también el número de palabras por subcorpus.

El concepto de "unidad textual" varía acorde con el subcorpus y no está asociado a la cantidad de palabras. En *Documentos*, es aquella que el investigador consideró un texto independiente para integrar la edición crítica en cuestión, desde una carta o una nota de diez líneas, hasta un expediente jurídico completo, 70 u 80 folios. En *Literatura*, la unidad textual es cada una de las divisiones contenidas en el conjunto de la obra literaria de la edición crítica o príncipe subida al CORDIAM, un capítulo, un poema, un sermón, etc. En *Prensa*, la unidad textual es un texto con autonomía comunicativa y reconocible gráficamente dentro del periódico del cual está tomada, sea cual sea su extensión en palabras; dicho de otro modo, una "noticia", "artículo" o "publicidad" recortados electrónicamente.

El listado de los corpus de documentos tomados de archivo actualmente procesados e incorporados al CORDIAM puede ser consultado en www.cordiam.org/doc/documentos-referencias.html.

Tabla 3.1 Universos de palabras y textos de CORDIAM.

	Documentos	*Literatura*	*Prensa*	*Total*
Número de textos	5.703 4.969	3.173 2.436	6.973 5.525	15.669 12.907
Número de palabras	4.543.606 4.196.259	4.517.856 3.244.912	2.688.402 2.203.395	9.644.566

Ejemplos de textos especialmente transcriptos para el CORDIAM –que se suman a los textos ya relevados con anterioridad, oportunamente mencionados– son los *Textos para la historia del español. El Salvador de los siglos XVII a XIX*, transcriptos por José Luis Ramírez Luengo, zona hasta el momento casi inexplorada desde la historia de la lengua española, o el *Diario de un soldado*, documento de un soldado argentino del siglo XIX, transcripto y publicado con anterioridad, pero cotejado con su autógrafo y sistematizado para su incorporación al CORDIAM por Gabriela Resnik, o el *Epistolario* del mexicano Guillermo Prieto, siglo XIX, transcripto especialmente para el *CORDIAM* por Miguel Pastrán.

El listado de periódicos de los cuales se tomaron unidades textuales, ya incluidos en el CORDIAM, se encuentra en www.cordiam.org/doc/prensa-referencias.html. La selección, hasta el momento, se ha realizado en las hemerotecas digitalizadas hispanoamericanas, y el equipo de especialistas del *CORDIAM* ha hecho un trabajo de recorte y cotejo que comienza con los primeros textos de prensa publicados en el continente americano, siglo XVIII, tales como la *Gaceta de México*, la *Gaceta de Santafé de Bogotá* o la *Gaceta de Lima*.

Las obras literarias incorporadas están consignadas en www.cordiam.org/doc/literatura-referencias.html. Entre ellas y también allí consignadas, el CORDIAM cuenta con ediciones críticas autorizadas por sus autores, como es el caso de la edición de Rosario Gala de la *Crónica* de Huamán Poma de Ayala, o textos hasta ahora inéditos como los *Villancicos* de Sor Juana Inés de la Cruz, editados y autorizados para CORDIAM por Jorge Gutiérrez Reyna.

Basta una mirada al listado de los corpus de documentos, de los textos de prensa y de las obras literarias incorporados al CORDIAM para constatar el carácter colaborativo de este corpus construido a partir de corpus preexistentes en el caso de CORDIAM-Documentos, y a partir de trabajos previos de digitalización de prensa. Además de ello, otras dos cuestiones merecen ser destacadas. Por una parte, merece la pena mencionar la riqueza de los materiales de investigación que eran poco accesibles a la comunidad académica por estar impresos, por ser de poca circulación o por ser manuscritos o impresos antiguos inaccesibles, en tanto que no habían sido editados hasta su subida al CORDIAM. Por otra parte, hasta la existencia del CORDIAM estos textos estaban, como es lógico, inconexos, de manera que investigar con ellos suponía un enorme esfuerzo adicional. Los datos hasta hace poco dispersos dificultaban fuertemente tener una visión general y poner en relación las continuidades, así como las discontinuidades entre las variedades del español europeo y las del español americano. En síntesis, a través del CORDIAM, cual si tuviéramos un *dron*, empleando una expresión coloquial, accedemos a la visión general, pero podemos acercarnos mucho a parcelas menores y podemos acceder así a datos para resolver problemas investigativos y preguntas de investigación antes poco abordables o poco factibles de ser realizadas. De igual manera, mediante el CORDIAM, el investigador tiene ahora acceso inmediato no sólo a nuevos espacios textuales de la lengua, la literatura y la prensa, para los que antes se requería ir directamente a hemerotecas, físicas o virtuales, o a bibliotecas. En el caso de la literatura, a través de este corpus se puede acceder a ediciones príncipe o a ediciones críticas, existentes de forma individual, claro está. La reunión de esta variedad textual permite comparar similitudes y diferencias entre, por ejemplo, escritura creativa y no creativa, literatura y documentos, respectivamente, o entre textos con diversos grados de inmediatez o distancia comunicativa, documentos, prensa y literatura, o permite poner en relación cualesquiera ángulos de investigación, diacrónica, diatópica y textual, que el usuario desee o necesite establecer. En las próximas secciones ahondaremos sobre las características de los tres subcorpus.

El CORDIAM, como muchos corpus electrónicos, es una obra inacabada y sigue creciendo. Actualmente, se realizan dos grandes subidas anuales de unidades textuales, sistematizadas bajo estrictos criterios filológicos, con tres agendas en paralelo: en CORDIAM-Documentos,

seguimos invitando a potenciales colaboradores y continuamos transcribiendo materiales de archivo, bien de países poco representados a la fecha en el corpus, ~~o~~ bien de periodos menos representados, tal es el caso del siglo XIX, una centuria bastante desatendida en la historia de la lengua española en América, que suele centrar toda su atención, en cuanto a documentos de archivo se refiere, en los siglos coloniales. En CORDIAM-Literatura, la agenda es subir el canon literario de los siglos XVI al XIX para cada país, en la medida en que tengan creación literaria, acorde con los criterios de especialistas de literatura de los diferentes países hispanoamericanos, en diálogo con el coordinador de este subcorpus. Finalmente, en CORDIAM-Prensa, la agenda es subir todos los periódicos del siglo XVIII —no existe prensa en lengua española previa a este siglo para aquellos países que tuvieron prensa en ese siglo, sea cual sea la periodicidad hemerográfica, y subir el siglo XIX de todos los países, pero fundamentalmente aquellos periódicos que se mantuvieron en lapsos considerables.

Para cada uno de los subcorpus, contamos con un *manual del editor*, que orienta en cada uno de los pasos de la adecuación filológica e informática de los textos. Realizada esta adecuación, imprescindible para su posterior inclusión en un proceso que sería oneroso explicar aquí, el documento, el texto de prensa o el texto literario en cuestión es revisado por las directoras y por los coordinadores y, finalmente, es enviado para su incorporación a la base de datos, componente central del CORDIAM, como veremos más adelante.

3. Consideraciones metodológicas

Como surge del estado de la cuestión, no hay intentos de informatización generalizadores americanos previos al CORDIAM. En ocasión de su construcción, además de la idea simple de juntar lo disperso, se tomaron algunas decisiones fundadas en concepciones de la investigación y del cambio lingüístico que se reflejan en la arquitectura y funcionamiento del corpus. En todo momento~~,~~ se buscó, además, que pudiera ser también empleado por usuarios ajenos a las disciplinas lingüísticas, en tanto que la lengua es el soporte que atraviesa la vida diaria de todo ser humano~~,~~ y, por ello, es también el soporte de investigación de otras disciplinas, tales como, entre otras, la sociología, la historia, la antropología, etc. El CORDIAM es, como todo corpus, una herramienta de servidumbre transdisciplinaria, aunque fue construido pensando, en primer lugar, en generar la infraestructura para poder realizar la historia del español en América y, en segundo lugar, para integrar las variedades dialectales del español de este continente en una sólida historia general del español, sin parcelaciones dialectales, cuando ellas no sean pertinentes. Presentamos a continuación los cinco conceptos fundamentales que guiaron la arquitectura del CORDIAM.

3.1. Concepción de historia de la lengua

Por una parte, el CORDIAM fue creado para poder realizar una historia de la lengua construida en evidencias robustas y no basada en datos espigados, como se había hecho tradicionalmente con bastante frecuencia. Por citar una obra clásica de referencia obligada, véase Lapesa (1980). Por otra parte, para poder vincular la historia interna con la externa, cuando el objeto de análisis lo requiere, decidimos elaborar plantillas de metadatos con información histórica, geográfica, étnica del autor del documento o universo de palabras del texto en cuestión, entre otras, para cada uno de los tres subcorpus contenidos en el CORDIAM, Documentos, Literatura y Prensa. Cada unidad textual de estos tres subcorpus lleva asociada una plantilla de metadatos con la información *ad hoc*. Expondremos más adelante la conformación de estas tres plantillas. En suma, el CORDIAM fue creado con la idea de poder hacer historia de la lengua

surgida de datos, de poder alcanzar niveles explicativos —además del nivel descriptivo— y de poder vincular, con datos fidedignos, las relaciones entre historia interna e historia externa.

2. *Integración entre Filología + Lingüística Histórica*. El CORDIAM fue creado integrando la mirada clásica de la filología, en lo que hace al cuidado de las ediciones en cualquiera de los tres subcorpus, con la propia de la lingüística histórica, centrada en el cambio lingüístico, como un concepto no solo ligado a la variación diatópica, diastrática y diafásica, sino también ligado a la variación concepcional, y ligado, asimismo, a la idea de que el cambio es la suma de continuidades más discontinuidades (Bybee 2010, cap. 2; Company 2016).

3.2. *Discursividad y textualidad como condicionantes del cambio lingüístico*

Como viene siendo señalado desde hace décadas en el ámbito de la lingüística histórica renovada, el CORDIAM fue creado bajo el requisito de que había que incorporar textos diversificados, porque esa diversificación es manifestación de diferentes *loci* culturales, de modo que el corpus debía ofrecer una clasificación textual de los tipos representados en él, esto es, una tipología textual coherente, operativa y no atomizada (Bertolotti y Compay 2018 y referencias allí citadas). La tipología textual que está en la base de la construcción y funcionamiento del CORDIAM se sustenta en una concepción de que el tipo textual y la tradición discursiva en que aquel se inserta pueden, y suelen, ser condicionantes de modos distintos de manifestarse la gramática, las construcciones o el léxico en cualquier etapa de la lengua, esto es, modos distintos de elaboración, y son, casi siempre, condicionantes de dinámicas distintas del cambio lingüístico.

Aun en la aceptación de que nunca tendremos acceso a todas las manifestaciones culturales de un estado de lengua dado, la construcción del CORDIAM fue realizada con la idea de recoger las máximas posibles conservadas, y bajo esta óptica, estructuramos la clasificación textual de los tres subcorpus ya mencionados, Documentos, Literatura y Prensa, cada uno de los cuales ofrece una tipología textual propia, regida, en esencia, por los diferentes grados de elaboración lingüística del material y por las diferentes formas de circulación del texto en cuestión (volveremos sobre esto en §4).

3.3. *Importancia de la frecuencia de uso como síntoma de la estructuración de la gramática*

El CORDIAM fue construido, desde su concepción inicial, sobre la idea de que concentrados frecuenciales diferentes de un mismo fenómeno o construcción, sean por siglo sean por país, pueden ser un reflejo de distintas puestas de relieve o distintas elecciones gramaticales, a la vez que la frecuencia es un valioso indicador de continuidades y discontinuidades en los procesos diacrónicos (Company 2006a). Por ello, el CORDIAM es el único corpus electrónico histórico, hasta donde tenemos noticia, que proporciona una *disponibilidad inmediata en la misma pantalla de concordancias de los siguientes datos cuantitativos*: el número de concordancias acompañado del universo de palabras en el que se encuentran esas concordancias, así como también el universo de unidades textuales en que ellas se documentan. Además, el CORDIAM muestra el universo de palabras y de textos en los que se hizo una búsqueda, aunque no contengan el término o construcción buscados. A continuación, mostramos la información cuantitativa que proporciona el CORDIAM como resultado de la búsqueda *peste:*

> *Encontrados 176 158 casos en 138 126 (de 15669 12 907) documentos que contienen 311579 275 740 (de 117498649 644 566) palabras*.

Otros corpus no proporcionan información cuantitativa de este tipo (aunque pueda obtenerse por vías no automáticas), y otros, tal es el caso del CORDE, proporcionan el número de ejemplos en el número de documentos, pero, en la medida en que un documento puede ser una carta de donación de un folio y otro puede ser una obra literaria de cientos de páginas, esta información, carente del correlato del universo de palabras, desorienta mucho al investigador.

3.4. Respeto hermenéutico y responsabilidad social-académica

El CORDIAM fue construido sobre la idea de que las bases y la infraestructura para la investigación deben ser de acceso abierto, esto es, sin obligación de suscripción, registro o pago, y deben ser totalmente explícitas. Por ello, da acceso al corpus y a los textos que constituyen la base de datos en su totalidad y no sólo al contexto de la concordancia. Ofrecer la posibilidad, además, de bajar la unidad textual al dispositivo del usuario y permite ordenar automáticamente las concordancias seleccionadas con sus datos para ser analizados y manipulados con diversos *software* (hojas de cálculo o bases de datos). Estas prestaciones de acceso total y reempleo o manipulación suponen una diferencia cualitativa informática muy importante respecto de todos los corpus electrónicos mencionados anteriormente, ya que aquellos, aun siendo de acceso libre en la red, solo ofrecen concordancias en un contexto restringido del *locus* textual de la búsqueda y algunos corpus sólo dan acceso a la concordancia. Por ejemplo, por citar un par de casos: BIBLIAS solo proporciona las concordancias, pero este es el objetivo de este corpus, a saber, poner en paralelo diferentes testimonios bíblicos para una búsqueda. CHARTA permite ver el documento completo, pero este no puede ser bajado al dispositivo del investigador.

3.5. Arquitectura y prestaciones de CORDIAM

El CORDIAM tiene tres componentes fundamentales: (a) un conjunto de datos agrupados en tres subcorpus, ya descriptos (CORDIAM-Documentos, CORDIAM-Literatura y CORDIAM-Prensa); (b) un motor de búsqueda diseñado por Alexander Gelbukh y Grigori Sidorov (ajustado en sus últimas versiones por Alexander Gelbukh), concebido con adecuación epistemológica y metodológica con lo expresado en la sección anterior y (c) una interfaz amigable que vehiculiza las diferentes funciones y herramientas de búsqueda. En cuanto a las prestaciones, el CORDIAM ofrece búsquedas simples y complejas, como cualquier corpus, pero, a diferencia de la mayoría, las búsquedas complejas pueden ser de una cierta sofisticación sintáctica. El resultado de las búsquedas se ofrece a través de concordancias, como cualquier corpus, pero ofrece además, como dijimos, un contexto mayor y el acceso al documento completo en el cual aparece la concordancia, para permitir la interpretación de la concordancia en todos los casos en que el fenómeno bajo análisis requiere de informaciones contextuales, como podemos ver en la Figura 3.1. Es claro que un contexto mayor habilita una interpretación más refinada, imprescindible para una correcta hermenéutica de muchos fenómenos lingüísticos y obligada para la comprensión de la textualidad de los problemas y de la textualidad *per se*.

Posee, para las búsquedas avanzadas, un conjunto de variables que habilitan a limitar u orientar las búsquedas. Los tres subcorpus ya nombrados están claramente delimitados gráficamente en la interfaz. El CORDIAM ofrece, por defecto, una búsqueda en todo el corpus, esto es, en la suma de los tres subcorpus, pero el usuario puede hacer la selección de subcorpus y de subtipos textuales dentro de cada uno que le sea útil para una determinada investigación o una determinada búsqueda. El conjunto de metadatos buscables, asociado a cada unidad textual, permite al usuario la creación de su propio subcorpus, esto es, la

Figura 3.1 Concordancia con ventana lateral que muestra acceso a contexto mayor y ~~a pdf~~ del documento completo.

selección de un subconjunto de documentos o textos que respondan a una o más variables buscables. Esta cantidad de categorías intermedias entre la concordancia y el corpus principal constituye, como es sabido, una buena práctica de la lingüística de corpus (EAGLES 1996). Para garantizar el equilibrio en la comparación de subcorpus, el usuario del CORDIAM cuenta, además, con la posibilidad de establecer el universo de palabas aproximado con el que desea trabajar, por ejemplo, puede pedir un subcorpus de 15.000 palabras de documentos de prensa de México del siglo XVI y otro similar del siglo XIX que contengan los artículos *el* o *un*. Las variables funcionales para búsquedas son: (a) temporales (siglos, años o periodos y, en el caso de la prensa, día(s) específico(s)); (b) geográficas (países actuales); (c) autorales (sexo, datos étnicos de los autores y, en el caso de la prensa y de la literatura, el nombre de los autores); (d) para los textos de prensa, el nombre del periódico es una variable buscable y (e) para el caso de literatura, el nombre del autor es una variable, como es usual en los corpus electrónicos.

Para cada uno de los tres subcorpus, hemos construido una tipología textual que también puede ser empleada para realizar búsquedas. *Grosso modo*, y sin soslayar las muchas implicaciones teóricas que conlleva la creación de una tipología, las hemos conciliado con el trabajo inductivo y con decisiones operativas guiadas por la pregunta de *qué* y *cómo* buscaría un usuario en un corpus electrónico (para más detalles y para la discusión de los límites de la tipología, ver Bertolotti y Company 2018, 82–92). Se encuentran en el CORDIAM cuatro tipos textuales de documentos, cinco de textos literarios y tres de prensa. Consideramos como *tipo textual* la abstracción de un conjunto de clases o géneros, agrupables, al menos, por un rasgo externo común: la finalidad del texto y la recurrencia de ciertas regularidades internas como el tipo de secuencias (descriptivas, narrativas, argumentativas y dialógicas), los temas, el léxico o el grado de complejidad sintáctica. Con este concepto de *tipo*, establecimos los correspondientes a cada uno de los subcorpus, los cuales se pueden ver en la Tabla 3.2.

Veamos en (1), (2) y (3), respectivamente, un ejemplo del tipo *cronísticos* del subcorpus CORDIAM-Documentos, un ejemplo de tipo *prosa varia* del subcorpus CORDIAM-Literatura y un ejemplo de tipo *publicitarios y anuncios varios* del subcorpus CORDIAM-Prensa.

(1) Los/señores trayan sus collares de quentas de oro baxo y algunas/piedras verdes que llaman chalchuyo que preçiavan/ellos mucho en aquel tiempo sus mujeres se vestian de la/manera que el dia de oy andan bestidas con camisas que llaman/huipiles de algodón rricas labradas de muchas colores/y sus naguas rricas con sus çenefas y ansi las señoras/andan mas señaladas que las otras. [Año 1579, México, Documentos cronísticos, CORDIAM]

(2) Esta alma hijos mios, quando sale de este cuerpo no se acaba, ni muere, como se acaban las bestias, y animales, que en muriendo el caballo ó el perro, le echays en el muladar, y no ay mas quenta {p.2} con el, porque ya se acabó del todo, mas los hombres no somos assi, antes quando el alma sale deste cuerpo, va luego a otra vida. // [Año ca. 1621, Chile, Prosa varia, CORDIAM]

Tabla 3.2 Tipos textuales en CORDIAM.

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

(3) UN MATRIMONIO\\\//En la Estancia del "Ombú" se precisa un matrimonio.//La mujer para cocinera y el marido para peon de la estancia. // Para tratar, en el mismo establecimiento, pero escusado es se presenten sin buenas recomendaciones. // [Año 1879, Uruguay, Documentos publicitarios y anuncios varios, CORDIAM]

Existen otros recursos para las búsquedas. El usuario puede orientar su búsqueda sustituyendo un carácter por un comodín, lo cual es muy útil en casos, por ejemplo, de variación gráfica. O puede sustituir dos o más caracteres, si solo quiere controlar el comienzo, el fin o el comienzo y fin de una palabra (puede buscar todas las palabras comenzadas por *bus**, todas las terminadas en **eda* o las comenzadas en *bus* y terminadas en *eda* con algo en medio como en la búsqueda *bus***eda*). Es posible considerar o no mayúsculas y acentos. Ya como prestaciones más sofisticadas, el CORDIAM habilita búsquedas contextuales consistentes en buscar una palabra, parte de palabra o una construcción y al mismo tiempo otra palabra o parte de palabra (o lema), estableciendo la distancia entre ellas, así como la posición de la segunda, antepuesta o pospuesta.

Finalmente, se cuenta entre las prestaciones del CORDIAM la posibilidad de visualización de los metadatos, bien como una ventana emergente al apoyar el cursor en la concordancia, bien como una ventana lateral a la derecha, al seleccionar *metadatos*, lo cual la contextualiza facilitando la decisión de conservar o eliminar la concordancia en cuestión. Esto es especialmente relevante en la medida en que con las concordancias seleccionadas (también con todas las arrojadas en la solicitud, por cierto) es posible hacer una base de datos en automático que consigna tanto las concordancias como algunos de sus metadatos. Podemos observar la ventana con metadatos en la Figura 3.2.

En cuanto a la clasificación de datos, forma, asimismo, parte de la arquitectura del CORDIAM una prestación para ordenar los datos de seis formas diferentes: la primera, y por defecto, es la ordenación cronológica como primer criterio, por país como segundo criterio y por tipo textual, como tercer criterio. Sin embargo, el usuario puede elegir cambiar cualquiera de estos tres órdenes y puede, además, solicitar un ordenamiento aleatorio (replicable o cada vez diferente), como se puede ver en la Figura 3.3 abajo. También se puede ver allí que el CORDIAM ofrece la posibilidad de escoger orden *Alfabético* para que las concordancias puedan, además, ser ordenadas de esta manera, herramienta muy útil para eliminar concordancias no deseadas. Por ejemplo, en una búsqueda de la terminación *-ais* como flexión verbal va a incluir en sus resultados ocurrencias de *país* y de *pais* (si se buscó sin tilde). Al ordenar alfabéticamente, todas las formas *país* quedarán juntas y facilitará al usuario marcarlas una tras otra para eliminarlas luego fácilmente.

4. Direcciones futuras y conclusiones

En este capítulo hemos descrito las propiedades, las características del CORDIAM, un corpus altamente especializado ya que solo contiene textos escritos en América y, en su gran mayoría,

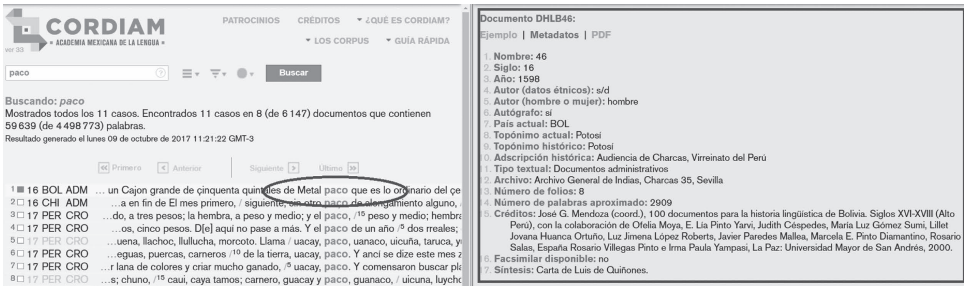

Figura 3.2 Metadatos desplegados en ventana lateral.

Figura 3.3 Formas alternativas de ordenamiento de las concordancias.

por hablantes/escribientes nativos de la totalidad de ese continente, con datos textualmente diversificados y de una profundidad histórica de 400 años.

Al momento de cierre de este artículo, y con menos de un lustro de abierto a la comunidad científica, *Google Scholar* arroja más de 5.000 citas, lo cual evidencia la pertinencia de su creación, que recoge un interés por incorporar a la investigación la lengua española en América.

Hemos mostrado los principios y los objetivos de la construcción de este corpus electrónico. Subrayados los aspectos en los cuales el CORDIAM vino a renovar la manera de hacer historia de la lengua en América, al proporcionar por primera vez un conjunto de datos cualitativa y cuantitativamente robusto, subrayadas las características de su interfaz, subrayado su talante colaborativo y el desarrollo de categorías de análisis de relevancia filológica y textual, cabe señalar algunos pasos para el camino futuro.

En cuanto a los aspectos filológicos y textuales del CORDIAM, seguiremos sumando documentos originarios de archivo, textos tomados de la prensa periódica americana y textos literarios, contando, como hemos contado hasta ahora, con la generosa y desinteresada colaboración de colegas investigadores de diversas partes el mundo, a la vez que seguiremos subiendo ediciones realizadas por el equipo de filólogos del CORDIAM.

Desde el punto de vista informático, estamos comenzando a avanzar en tres líneas. Por una parte, estamos mejorando la lematización a través del entrenamiento del programa en el reconocimiento de la variación gráfica; por otra parte, estamos iniciando la subida de facsímiles, y, en tercer lugar, estamos indagando maneras de lograr un acoplamiento de la

búsqueda en el texto informatizado con la zona que contiene esa búsqueda en el respectivo facsímil.

En cuanto al cobijo institucional, a partir del año 2020, contamos con el asesoramiento literario de varias de las academias de la lengua americanas, reunidas en el organismo *Asociación de Academias de la Lengua Española* (ASALE). De hecho, al día de hoy el *Cordiam* es, como ya dijimos, una obra AML-ASALE, esto es, Academia Mexicana de la Lengua y Asociación de Academias de la Lengua Española, si bien sigue siendo gestionada y respaldada por la AML.

Por fin, y como surge del estado de la cuestión, hemos visto que existe un conjunto de corpus valiosos del español, inconexos entre sí. Sirvan estas últimas líneas como un llamamiento a buscar formas de colaboración y hacer de ese conjunto de corpus una herramienta más potente que las que tenemos actualmente, para emplear de manera inmediata datos de todos los corpus que se requieran, de modo que nos permitan avanzar en un mejor conocimiento de los datos históricos del español, en un mejor conocimiento de la teoría del cambio lingüístico, así como, en general, en nuestra capacidad de historiar la lengua española y de acercarnos a etapas pasadas de la cultura hispánica.

5. Lecturas adicionales

Dolores Corbella, D., Fajardo, A. y J. Langenbacher-Liebgott, eds. 2018. *Historia del léxico español y Humanidades digitales*. Berlín: Peter Lang.

Kabatek, J., ed. 2016. *Lingüística de corpus y lingüística histórica iberrorománica*. Berlín/Boston: Walter de Gruyter

Nota

1 Dos trabajos de informatización y disponibilización de corpus históricos del español están en curso: el *Corpus Histórico del Español del Reino de Granada*, a cargo de Miguel Calderón Campos y Mayte García Godoy de la Universidad de Granada y el *Corpus Histórico de Canarias*, a cargo de Dolores Corbella de la Universidad de La Laguna. Para un panorama de los corpus iberorrománicos existentes, ver el portal CORHIBER, bajo la coordinación de Joan Torruella y Johannes Kabatek, www.corhiber.org/.

Referencias citadas

Bertolotti, V. y C. Company Company. 2018. "El corpus para América: *Cordiam*". En *Historia del léxico español y humanidades digitales*, eds. D. Corbella, A. Fajardo y J. Langenbacher-Liebgott, 75–105. Berlín: Peter Lang.

Bosque, I. y V. Demonte, eds. 1999. *Gramática descriptiva de la lengua española*. Madrid: Espasa-Calpe.

Bybee, J. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

CHARTA. *Corpus Hispánico y Americano en la Red: Textos Antiguos*. www.charta.es.

CODEA. *Corpus de Documentos Españoles Anteriores a 1700*. http://corpuscodea.es/.

Company Company, C. 2001. "Para una historia del español americano. La edición crítica de documentos coloniales de interés lingüístico". En *Studia in honorem Germán Orduna*, eds. L. Funes y J. L. Moure, 207–105. Alcalá de Henares: Universidad de Alcalá.

Company Company, C. 2006a. "Gramaticalización y frecuencia de uso. Los paradójicos sintagmas con artículo + posesivo en el español medieval". *Revista de Historia de la Lengua Española* 1: 5–31.

Company Company, C. dir. 2006b. *Sintaxis histórica de la lengua española*. *Primera parte*: *La frase verbal*. México: Fondo de Cultura Económica-Universidad Nacional Autónoma de México.

Company Company, C. dir. 2009. *Sintaxis histórica de la lengua española*. *Segunda parte: La frase nominal*. México: Fondo de Cultura Económica-Universidad Nacional Autónoma de México.

Company Company, C. dir. 2014. *Sintaxis histórica de la lengua española*. *Tercera parte: Adverbios, preposiciones y conjunciones. Relaciones interoracionales*. México: Fondo de Cultura Económica-Universidad Nacional Autónoma de México.

Company Company, C. 2016. "Gramaticalización y cambio sintáctico". En *Enciclopedia de lingüística hispánica*, vol. 2, ed. J. Gutiérrez-Rexach, 515–526. London: Routledge.

CORDE. Real Academia Española. *Corpus diacrónico del español*. www.rae.es.

CORDIAM. Academia Mexicana de la Lengua. *Corpus Diacrónico y Diatópico del Español de América*. www.cordiam.org/.

CORHIBER. *Portal de Corpus Históricos Iberorrománicos*, eds. Torruella, Joan y Johannes Kabatek. www.corhiber.org/.

CORPES. Real Academia Española. *Corpus del español del siglo XXI*. http://web.frl.es/CORPES/view/inicioExterno.view.

CORPUS DAVIES. *Corpus del español*. www.corpusdelespanol.org/.

CORPUS DE BIBLIAS. *Biblia Medieval*. www.bibliamedieval.es/index.php.

CREA. Real Academia Española. *Corpus de referencia del español actual*. www.rae.es.

EAGLES. 1996. *Preliminary Recommendations on Corpus Typology*. www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Gutiérrez-Rexach, J., ed. 2016. *Enciclopedia de lingüística hispánica*. London: Routledge.

Kabatek, J. 2016. "Un nuevo capítulo en la lingüística iberorrománica: el trabajo crítico con los corpus". En *Lingüística de corpus y lingüística histórica iberorrománica*, ed. J. Kabatek, 1–17. Berlin/Boston: de Gruyter.

Lapesa, R. 1980. *Historia de la lengua española*. Madrid: Gredos.

López Serena, A. 2007. "La importancia de la cadena variacional en la superación de la concepción de la modalidad coloquial como registro heterogéneo". *Revista Española de Lingüística* 37 (1): 371–398.

Marquilhas, R. 2000. *A faculdade das letras. Leitura e escrita no Portugal no século XVII*. Lisboa: Imprensa Nacional-Casa da Moeda.

Nuevo Diccionario Histórico Del Español. http://web.frl.es/DH/org/login/Inicio.view;jsessionid=94C51B5D2832DB6146CF9DD225DE26B1.

Oesterreicher, W. 1996. "Lo hablado en lo escrito. Reflexiones metodológicas y aproximación a una tipología". En *El español hablado y la cultura oral en España e Hispanoamérica*, eds. T. Kotschi, W. Oesterreicher, y K. Zimmermann, 317–340. Frankfurt/Madrid: Vervuert/Iberoamericana.

POSTSCRIPTUM Arquivo Digital de Escrita Cotidiana em Portugal e Espanha na Época Moderna. http://ps.clul.ul.pt/.