Corpus históricos del español
Avances y tareas pendientes

Historical corpora of Spanish
Improvements and pending tasks

Virginia Bertolotti
Universidad de la República (Montevideo, Uruguay)
virginia.bertolotti@gmail.com
https://orcid.org/0000-0002-1845-1040

Concepción Company Company
Universidad Nacional Autónoma de México (Ciudad de México, México)
concepción.company@gmail.com
https://orcid.org/0000-0002-6351-715X

Recibido el 7/6/2024, aceptado el 17/6/2024, publicado el 18/10/2024

Cómo citar este artículo
Bertolotti, Virginia, Concepción Company Company 2024. Corpus históricos del español. Avances y tareas pendientes. *Studia linguistica romanica* 2024.12, 1-17. https://doi.org/10.25364/19.2024.12.1.

Resumen
Esta introducción analiza el papel que los corpus históricos electrónicos han tenido en el desarrollo de la lingüística histórica del español y en el enriquecimiento de la filología hispánica en las últimas décadas. A partir de allí y de una revisión de los corpus existentes, plantea las motivaciones para hacer este número temático sobre corpus electrónicos, establece, posteriormente, avances y retos en la construcción de corpus históricos y, finalmente, pone en contexto los artículos del número temático, con una breve presentación de cada uno de ellos.

Abstract
The introduction analyzes the role of electronic historical corpora in the development of Spanish historical linguistics and the enrichment of Hispanic philology over the past few decades. First, it reviews the electronic corpora currently available in Spanish and explains the motivation for this thematic issue. Second, it examines the main developments in Spanish corpus linguistics and discusses some of the challenges that need to be addressed to improve historical electronic corpora. Finally, the introduction presents and summarizes the articles included in this thematic issue.

Índice

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

1 Corpus electrónicos, lingüística y filología

[1] La creación de corpus electrónicos de lengua española, sincrónicos e históricos, a partir de la última década del siglo pasado ha modificado de manera profunda el modo de hacer lingüística en los últimos 30 años. No solo ha hecho posible disponer de muchos más datos para un fenómeno dado, sino también ha permitido una notable reducción del tiempo de búsqueda, selección de los ejemplos idóneos y correspondiente fichado, y ha fortalecido asimismo la posibilidad de hacer generalizaciones mucho más fuertes a partir de evidencias empíricas más robustas. Esta importante modificación de la hispanística ha impactado, de modo particular, en la lingüística funcional interesada en la variación de la lengua – muy especialmente la gramática y la lingüística históricas del español –, pero también la lingüística formal sincrónica se ha enriquecido de la existencia de corpus, porque las intuiciones y competencia lingüísticas del investigador – que constituían la base empírica fundamental de esos análisis – pueden ser respaldadas y confrontadas con ejemplos procedentes de una o más comunidades y no de un solo hablante-investigador.

[2] La irrupción de los corpus en la lingüística hispánica no debe ser simplemente concebida como una manera más cómoda de acceder a más datos, que sin duda lo es, como decimos arriba, aspecto este que suele ser pensado como el beneficio inmediato del acceso a corpus electrónicos, sino que – y esto es un hecho esencial – ha llevado a mejores planteamientos e interesantes matices en la teoría del cambio lingüístico. En efecto, la existencia de corpus ha permitido más y mejores generalizaciones y ha hecho posible articular mejor la historia externa con la historia interna de la lengua, ha habilitado una demarcación dialectal amplia, y también refinada si así se plantea la búsqueda y lo permite el corpus, dentro de los cambios que han tenido lugar en la lengua española, y ha llevado a una profunda reflexión metodológica sobre qué seleccionar, qué incluir y qué no incluir ante el espejismo de la documentación exhaustiva alentado por la informática.

[3] No hay que perder de vista en este 'nuevo' ambiente investigativo que los análisis con base en corpus electrónicos no se excluyen, en lo absoluto, con la investigación 'tradicional' basada en lectura de textos. Todo depende del fenómeno que se desea analizar y de las preguntas de investigación. La búsqueda en corpus será mucho más fructífera, y obligada ya hoy, si el investigador desea, por ejemplo, abordar el análisis de clases cerradas o que, de algún modo, son restringibles con las herramientas disponibles en un corpus dado; pensemos, sea por caso, en los adverbios en *-mente*. La búsqueda mediante lectura en texto será mucho más útil – y más rápida y casi la única posible – si el investigador desea analizar estructuras sintácticas que no son léxicamente acotadas, sea por caso, los complementos adjuntos o el orden de constituyentes de la oración. E incluso con corpus etiquetados gramaticalmente, es posible que un investigador no comparta los criterios de etiquetado que otro investigador hizo, y prefiera analizar y clasificar por sí

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

mismo los textos. No obstante, ambos modos de investigar son complementarios, porque, por ejemplo, una vez registrado un fenómeno dado mediante lectura de textos y establecidas las variables independientes sugeridas por esa lectura, el investigador puede completar y, sobre todo, enriquecer su análisis con búsquedas más puntuales y acotadas en un corpus.

[4] En esta línea de diálogo entre innovación y tradición, el desarrollo de los corpus históricos ha enriquecido también la filología tradicional, aspecto este que parece casi paradójico, porque, a primera vista al menos, la informática y la filología parecieran no dialogar y no necesitarse recíprocamente. En efecto, aquella se asocia con el presente y el futuro mientras que esta suele ser vinculada con las humanidades clásicas. Sin embargo, la diversidad y sofisticación de los corpus históricos informatizados va de la mano de un creciente interés por el rescate de fuentes y por un acusado refinamiento ecdótico en el modo de editarlas. Una contribución filológica importante a la sofisticación de los corpus es la búsqueda y edición de fuentes cada vez más diversas textualmente, más allá de la literatura y de los documentos jurídico-administrativos extraídos de archivo, que fueron, como se sabe, las clases de textos que constituyeron la base de la investigación inicial en lingüística histórica del español. Los corpus históricos se abren crecientemente a la prensa, a los textos de especialidad, a las traducciones, a diarios personales, a muy diversos géneros literarios menores y también a testamentos, a epistolarios y a todo tipo de documentos entre particulares, desde notas íntimas hasta notas con inventarios de bienes de casa con indicaciones marginales coloquiales de cómo administrarlos, entre otros diversos géneros textuales. Seguramente, este enriquecimiento textual y el correlativo refinamiento informático de los corpus constituyen la razón de que seamos cada vez más exigentes con los instrumentos para la investigación. Sin duda, el enorme desarrollo en informática – y el salto dado en los últimos años por la inteligencia artificial en particular – nos da la idea de que todo es posible en términos técnicos. Los artículos que integran este número temático son muestra de esta complementación entre informática y filología y de cómo aquella y esta ponen cada vez el estándar más alto en la investigación.

2 Motivaciones para hacer este número temático

[5] Este número temático *Corpus históricos electrónicos de la lengua española. Retos y direcciones futuras* es una contribución a la mejora de los corpus existentes, ya que tiene como objetivo, por un lado, reflexionar sobre los avances informáticos y filológicos de estos y, por otro, (re)pensar los retos que los corpus históricos del español presentan a la fecha. Es resultado de un coloquio homónimo, realizado en la Universität Graz (Austria), entre el 22 y el 25 de febrero de 2023, en el marco del *XXIII Congreso de la Asociación Alemana de hispanistas: Nuevos ambientes de la hispanística: digitalización – reinscripciones – interfaces*. Convocamos a este coloquio, o sección como se le llama en los congresos de la Asociación Alemana de hispanistas, para poner en diálogo a los creadores-gesto-

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

res de corpus históricos y a los usuarios de estos, ya que unos y otros han modificado en las últimas décadas el panorama de la lingüística hispánica diacrónica. La finalidad última de este número temático, en suma, es contribuir, en alguna medida, a realizar mejoras tanto en la investigación en lingüística histórica del español basada en corpus electrónicos como en las herramientas de búsqueda que estos poseen.

[6] Nuestras preguntas de partida en la sección del *XXIII Congreso de la Asociación Alemana de hispanistas* fueron tres. Primero, ¿estamos a tiempo de marcar un derrotero de dialogicidad y vinculación informática entre los corpus ya existentes? Si se lograra, es posible que la comunidad de investigadores pudiera sacar mejor provecho de los esfuerzos ya realizados. Segundo, ¿en qué apuestas de prestaciones vale la pena invertir? Tercero, ¿cuál debe ser el peso en el diseño y selección de datos para que otras disciplinas pudieran hacer uso de los corpus históricos?

[7] Nuestro diagnóstico previo al trabajo de una semana en Graz era que no existían corpus históricos electrónicos con las siguientes características informáticas: a) lematización completa, y correcta, de cualquier etapa y dialecto del español antiguo; b) acoplamiento de facsímil y concordancia para una búsqueda, esto es, que el corpus sitúe la búsqueda en cuestión de forma simultánea en el facsímil y en la transcripción de este; c) posibilidad de interacción entre corpus distintos desde el corpus de base en que se esté trabajando, sin necesidad de salirse de uno y entrar a otro, y, de forma colateral, d) herramientas de transcripción automática en cuanto a fidelidad filológica y en cuanto a economía de tiempo en el resultado. El listado de estos aspectos fue resultado de un diálogo previo entre nosotras respecto a una valoración general de los corpus históricos existentes, como usuarias de ellos y gestoras de uno de ellos, diálogo y valoración que nos llevó a plantear la sección en cuestión.

[8] Si bien hemos confirmado este diagnóstico en lo general, se comienzan a ver también claros rumbos para el avance. Uno, desde la lingüística, señalábamos que debíamos avanzar en las arquitecturas deseables para lograr mejores y más robustas generalizaciones. Por ejemplo, uno, cómo afinar, y enriquecer si fuera el caso, el tipo de metadatos deseables para una mejor integración de historia interna e historia externa en la explicación de los cambios. La razón subyacente es que la lengua es una actividad transversal a la vida cotidiana de cualquier individuo, una actividad que supone siempre una interacción de gramática genéticamente adquirida y de condicionamientos motivados socioculturalmente, por lo cual la causación externa está, así sea de manera latente, en la codificación gramatical. Dos, cómo avanzar y profundizar en la reflexión sobre la clase de datos textuales (literarios, no literarios, prensa científica, documentos entre particulares, etc.) que alimentan los corpus y cómo permitir que estos capturen de manera automatizada el necesario, y ya presente, enriquecimiento textual. Tres, cómo avanzar en el trabajo transdisciplinario para la construcción de corpus históricos, lo cual implica diálo-

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

gos entre equipos de trabajo desde múltiples perspectivas. También nos percatamos en el coloquio de que, aunque los avances son importantes e interesantes, no son suficientes y hay que seguir teniendo en la mira el hecho de que los corpus históricos son perfectibles.

3 Avances y retos en la construcción y empleo de corpus históricos electrónicos

[9] El diálogo fructífero mantenido en el coloquio de Graz, que se refleja parcialmente en este número temático, nos confirma, en síntesis, el potencial de los corpus electrónicos y de las herramientas informáticas cada vez más sofisticadas para el análisis en lingüística histórica. Este potencial se puede resumir en la posibilidad de hacer generalizaciones fuertes dada la base empírica robusta, en la posibilidad de lograr explicitud total, en la posibilidad de hacer un tratamiento cuantitativo sofisticado, así como en la posibilidad de realizar una gestión independiente de los datos por parte del investigador, en aquellos corpus con haces de metadatos bien diseñados. Todo lo cual son ventajas evidentes y ganancias enormes para la solidez y agilidad de los procesos de investigación. Así también, constituyen avances, y es parte del potencial de que hablamos, el planteamiento de nuevos problemas y de nuevas hipótesis a partir del examen de los datos que emergen de estos grandes conjuntos documentales, que se conoce por su expresión en inglés como *corpus-driven*.

[10] Junto con estas virtudes, se empiezan a sopesar, asimismo, las consecuencias indeseadas de la existencia de los corpus, que fue un hecho notable en el diálogo que mantuvimos en el coloquio. Se ha señalado que la presión cuantitativista oscurece y empobrece el análisis cualitativo, y esto es cierto si se trabaja con corpus con acceso muy restringido a los datos y restringido y nulo a los fragmentos textuales que los arropan, ya que no se puede completar el análisis con contextos mayores o, incluso, con la lectura del documento completo.

[11] Los artículos que aquí se presentan, en su mirada analítica sobre los corpus históricos y su empleo, son una muestra clara de cómo los avances logrados – o mejor, gracias a los avances logrados – hacen posible, como comunidad científica, establecer nuevos retos. Veamos algunos de ellos, siete al menos, surgidos en el coloquio y hechos explícitos en varios de los artículos de este número temático.

[12] Reto 1. Comparabilidad de los datos en un corpus y entre corpus. En el plano ecdótico, se señala la incomodidad en la comparabilidad de los datos como consecuencia de que los criterios de transcripción no son homogéneos, en el caso de corpus colaborativos, como lo son la mayoría, en tanto que han sido creados a partir de la autorización de documentos y textos transcritos por distintos especialistas. Como es bien sabido, la transliteración de los documentos manuscritos provenientes de archivos, e incluso de aquellos que provienen de la prensa histórica, suponen muchas horas de trabajo por folio facsimilar, lo cual lleva a considerar

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

que es inadecuada una retranscripción masiva homogénea de la decena y media de corpus históricos del español operativos en la actualidad (véase § 5). Además lo anterior implicaría la tarea no menor de, primero, de que los diversos equipos gestores de corpus se pongan de acuerdo en la pertinencia de volver a las transcripciones; segundo, ponerse de acuerdo en los criterios de homogeneización paleográfica y ecdótica; tercero, entrenar a muchos colaboradores en la misma mirada paleográfica y ecdótica; cuarto, realizarlo en tiempos similares, etc. Un camino más factible para compensar, al menos parcialmente, este hecho constatable es hacer explícitos y accesibles los criterios de transcripción, mientras esperamos que los desarrollos de la inteligencia artificial nos ayuden en esta compleja tarea homogeneizadora.

[13] Reto 2. Aprovechamiento de la inteligencia artificial. El reconocimiento y la transcripción automáticos es un anhelo que hasta hace poco era impensable. Seguramente, la combinación de dos aspectos de la inteligencia artificial, el procesamiento del lenguaje natural y el procesamiento de imágenes, nos permitan en un futuro, quizá no lejano, avanzar sustancialmente en la transcripción de documentos, que es un proceso que demanda muchísimo tiempo filológico. El salto comenzará por los textos impresos, que presentan una menor variación gráfica – aunque no pequeña, por cierto –, para su reconocimiento automático, pero son los textos escritos por escribientes poco profesionales de la escritura, con *usus scribendi* bastante diferentes entre sí, los más buscados para alimentar los corpus históricos, y esos textos manuscritos siguen siendo el reto mayor.

[14] Reto 3. Hablantes y escribientes infrarrepresentados. Se señala también la existencia de grupos humanos infrarrepresentados en los corpus y se plantea la necesidad de focalizar el rescate documental en estos grupos, como está haciendo un conjunto importante de investigadoras e investigadores, por ejemplo, al poner el foco en la selección de cartas y otros documentos escritos por mujeres. Los documentos escritos en español por indígenas o por personas africanas no han tenido, hasta el momento, las mismas posibilidades de ser incorporados en los corpus históricos del español; hasta donde sabemos, no han sido objetivo explícito de ninguno de los grupos de investigación que está trabajando en la gestión de los corpus. En este plano, cabe señalar que la presencia de metadatos sociolingüísticos claros permite, para los corpus que cuentan con ellos, realizar recortes y hacer nuevos agrupamientos y subcorpus, incluso de estos grupos infrarrepresentados. Se señalan también grupos textuales infrarrepresentados, por ejemplo, el de las traducciones, por lo general científicas, que permitiría, junto con los textos en español escritos por indígenas y africanos, arrojar nueva luz a la gramática del español y a los muchos ángulos del contacto y del préstamo lingüístico. En esta línea de diálogo, se ha reclamado contar con *corpus de contacto*, entendido este de una manera amplia, lo cual es, sin duda, una necesidad: la historia de toda lengua es también la historia de sus contactos. Sin embargo, estos corpus suman a los monolingües un conjunto de desafíos mayores, ya que, para su explotación automática,

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

deberían transliterarse y traducirse, su presentación digital posiblemente requiera emparejamiento de lengua fuente y lengua meta, además de que requerirían, entre otras cosas, estar etiquetados, anotación que habría que realizar manualmente y con toda seguridad sería muy costosa, para que tuviera el carácter de exhaustividad, cuestión que retomamos más abajo. En relación con la conveniencia de contar con cualquier otro metadato que pudiera ser seleccionado para realizar las agrupaciones requeridas por un investigador, se confirmó la idea de que cuanto más y mejores metadatos tenga cada documento, mayor será la posibilidad de empleo del corpus en cuestión, ya que el usuario podría construir el o los subcorpus que considere necesarios.

[15] Reto 4. Diversificación textual. Otra aspiración manifestada en las discusiones de la sección es la necesidad de enriquecer los corpus en cuanto a variedad textual y hacer tipologías textuales adecuadas. Se señaló que, por sesgos de selección, algunos géneros – como los textos científicos y textos técnicos – no se suelen incorporar en los corpus históricos. Una explicación parcial a este escollo está relacionada, seguramente, con el problema de por qué hay siglos sobrerrepresentados y siglos infrarrepresentados en los corpus existentes. El desarrollo de la ciencia y de la técnica y su exposición textual es tardío en el mundo hispánico y coincide con aquellos siglos menos preferidos por los investigadores, por ejemplo, el siglo XVIII. A esto se suma el hecho de que estos textos – tal como sucede actualmente con la escritura académica – son altamente modelizados y durante algunas décadas los moldes (literarios, administrativos, jurídicos, etc.) eran evitados en la selección documental para integrar corpus, ya que los investigadores navegaban con el norte puesto en los textos que se aproximaban, en la medida en que puede hacerlo un texto escrito, a la oralidad, lo que se ha llamado *oralidad concepcional*.

[16] Reto 5. Lematización. Un reclamo constante de los usuarios es contar con corpus bien lematizados. La lematización, esto es, el proceso de emparejar una palabra del corpus con su forma canónica o su lema; por ejemplo, todas las formas del verbo *cantar* (*canto*, *cantaré*, *cantaba*, *cantábamos*, *cantarías*, *cantarían*, entre otras) deben estar informáticamente relacionadas con el lema *cantar*, de manera tal que una búsqueda de este verbo como lema muestre todas las apariciones de *cantar* en cualquier tiempo, modo, número o persona que aparezcan en el corpus, y con cualquier grafía en el caso de corpus históricos – *cantaua*, *cantava*, *cantaba*, *he cantado*, *e cantado*, etc. Este procedimiento, que tiene una tradición establecida y una muy buena resolución desde hace ya tiempo para los corpus informatizados de español actual, se enfrenta al problema de la ortografía para corpus históricos, como acabamos de indicar, o más bien de la ausencia de *orto*grafía, porque este concepto es bastante tardío, como es sabido. Es un hecho que ningún buen corpus histórico dejaría de respetar la ortografía original, que no es la ortografía canónica actual. Es claro que buena parte del contenido de los corpus históricos no es automáticamente lematizable, sino que requiere procesos complementarios, manuales muchas veces, de elevado costo y tiempo.

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

[17] Reto 6. Etiquetación gramatical. Otra demanda de los usuarios es la codificación, esto es, el proceso de agregar información al texto a través de etiquetas para su tratamiento informático de variada índole: categorial, morfológica, sintáctica, pragmática. Un texto codificado (etiquetado, anotado) tiene, indudablemente, grandes ventajas a los efectos de la precisión de las búsquedas y del empleo más inmediato de los datos obtenidos. Sin embargo, es necesario hacer notar que el trabajo y la responsabilidad de la codificación no son los mismos en todos los casos. Parece razonable pensar que diferentes anotadores harían un trabajo similar en el etiquetamiento de las categorías gramaticales básicas (verbo, nombre, preposición…), pero una anotación sintáctica y una anotación pragmática suponen tomas de decisión mucho más complejas por parte del anotador – por ejemplo, ¿adjetivo o sustantivo? en los numerosos casos de adjetivos que, aunque potencialmente adjetivos, suelen aparecer sustantivados en el uso real; ¿información conocida en texto o información conocida recuperable situacionalmente?, ¿tópico o foco? –, anotaciones que podrían ser fácilmente puestas en duda por otros investigadores, hecho que ya planteábamos al inicio de esta introducción. Si bien hemos avanzado mucho en el conocimiento de la sintaxis histórica y de la semántica y pragmática históricas, estamos muy lejos de tener un conocimiento tan asentado que permitiera poder anotar corpus con comodidad y que quienes los emplean para la investigación, sobre todo en el caso de los corpus históricos, que son los que aquí nos competen, tomen la anotación como un insumo transparente en su investigación, en la medida en que todo acto de etiquetar es, en última instancia, un acto de interpretar. En caso de que se hiciera, sería obligado redactar documentos claros y concisos, a la vez que muy explícitos, acerca de los criterios de etiquetación y, por supuesto, hacerlos disponibles.

[18] Reto 7. Mapeo de fenómenos. Como se mostró en la sección en Graz, los avances de la informática permiten interfaces de geolocalización de los documentos incluidos. El siguiente paso, en este sentido, sería mapear las densidades por fenómenos identificados, lo cual supondría un etiquetado de un elevado nivel de sofisticación, con el que ningún corpus cuenta actualmente y cuya probabilidad ponemos en duda por lo dicho en el reto anterior.

[19] En resumen, el diálogo mantenido, parcialmente reflejado en los textos de este número temático, nos permitió saber dónde estamos, saber cuáles aspectos más precisos son de urgencia inmediata y mediata en la construcción de corpus, así como saber en qué medida se pueden integrar nuevas herramientas a los viejos campos de estudio, y hasta dónde tal integración es deseable. Agradecemos, pues, a quienes desde diferentes experiencias con corpus y desde diferentes experiencias generacionales se sumaron a la discusión durante tres días en la maravillosa ciudad de Graz. Agradecemos también a Martin Hummel, a Katharina Gerhalter y a Stefan Schneider la motivación para realizar el coloquio en Graz y el diálogo para este número temático; sin su trabajo en el congreso y en el número temático, no existiría este último.

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

4 Los contenidos de este número temático

[20] Ocho artículos componen este número temático, agrupables en tres conjuntos. El primero, con tres textos, evalúa los corpus existentes, desde diferentes puntos de partida y con una mirada crítica (Rojo, Ramírez Luengo y Espinosa Elorza & Zieliński). Otro conjunto de artículos describe nuevos corpus y nuevas herramientas digitales en fases diversas de maduración (Del Rey Quesada & Carmona Yanes, Cruz Volio, Fajardo Aguirre & Corbella Díaz y Albers). Por fin, contamos con un trabajo de análisis lingüístico a partir de un corpus (Albitre Lamata).

[21] El primer artículo de este número temático, *El futuro de los corpus de referencia*, a cargo de Guillermo Rojo, va más allá del título. Nos ofrece el autor la mirada reposada de quien ha vivido el surgimiento de los corpus informatizados desde sus comienzos, los ha investigado y ha reflexionado[1]. Imparte el autor una clase sobre lingüística de corpus al caracterizar los dos puntos extremos en el *continuum* de los corpus, los corpus masivos y los corpus especializados, para luego situar y discutir las virtudes y los defectos de los corpus de referencia, sobre cuyo futuro se detiene. Describe los principales corpus masivos y señala la cantidad de datos como su evidente virtud, a la vez que la falta de codificación es su limitación evidente, a la que se suma, agregamos nosotras, la falta de criterios claros de selección textual que va de la mano de un bajo control filológico en la selección del material que se sube. En cuanto a los corpus especializados, necesariamente pequeños en comparación con los primeros, señala Rojo como virtud sus posibilidades de codificación cuidada y como defecto la relativa limitación para realizar generalizaciones fuertes a partir de los datos que de ellos se obtienen. Explica los criterios de diseño de los corpus de referencia y su propósito, además de su utilidad: poder asociar valores definidos en el diseño a los resultados de las búsquedas, así como la posibilidad de controlar frecuencias y de establecer subuniversos de datos comparables. Finalmente, sopesa el valor de los corpus de referencia para diferentes campos de las ciencias del lenguaje y señala la necesidad de ciencias auxiliares como la estadística para la interpretación de los datos, pero advierte del peligro que supone centrar la discusión lingüística solamente en la estadística. Cierra su trabajo señalando cuál será y cómo debería ser el futuro de los corpus de referencia, con un señalamiento sobre la incorporación del ámbito sonoro.

[22] El segundo artículo, a cargo de José Luis Ramírez Luengo, se centra en el análisis de un corpus de diseño, el *Corpus diacrónico y diatópico del español de América* (*CORDIAM*). El autor evalúa este corpus y hace una propuesta de historiografía de la filología centrada en documentación americana para luego enfocarse en dos problemas de la mayor importancia para los corpus: la calidad de los

1 Guillermo Rojo fue el gestor del *Corpus diacrónico y diatópico del español* (*CORDE*), corpus pionero en el ámbito de la lengua española y, sin duda, el más usado y citado – e incluso es fuente única en numerosas obras de sintaxis y de semántica históricas. Obras de referencia de gran envergadura, como la *Sintaxis histórica de la lengua española*, no habrían alcanzado el grado de generalizaciones y de ricas y matizadas evidencias sin la existencia del *CORDE*.

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

datos y la representatividad de los materiales. En la revisión del trabajo con datos lingüísticos americanos previo a la era de los corpus informatizados, el autor distingue un conjunto de problemas: la poca fidelidad de los textos transcriptos por historiadores o cientistas sociales (*avant la lettre*), el marcado grado de estereotipia de los textos literarios coloniales americanos, ya que son mayoritariamente costumbristas y más cuando se vuelven locales, o la falta de explicitud y disponibilidad automática de los textos transcriptos para realizar los estudios. Ramírez Luengo ubica el *CORDIAM* en la quinta etapa de la edición de textos americanos en la que se conjuntan la superación de los tres problemas señalados con las herramientas informáticas que permiten el procesamiento y el hacer públicos gran cantidad de datos, así como también permiten formas sofisticadas de explotación. Destaca su carácter colaborativo, su alcance americano, su amplio abanico temporal, la calidad de sus datos, muchos de ellos inaccesibles por medio alguno más allá de este corpus. Señala, sin embargo, que el origen diverso de los textos incluidos – editados algunos con criterios distintos, puesto que fueron autorizados por diferentes investigadores – puede hacer sombra filológica en algunas zonas del corpus. Se centra luego en el hecho de que el *CORDIAM* no es un corpus equilibrado temporalmente, ya que en él predominan los documentos del siglo XVI, y tampoco está equilibrado geográficamente, en la medida en que algunos países están sobrerrepresentados y otros subrepresentados, aunque ambas cuestiones estén claramente ligadas al acumulado o a la falta de acumulado de documentación disponible. Señala asimismo un segundo problema relacionado con la ausencia de representatividad dialectal al interior de los países. El ejercicio de Ramírez Luengo es un instrumento valioso ya que señala con claridad las zonas (geográficas e históricas) más desiertas de documentación. Indica asimismo el deseo de incorporar textos que sean producto de los múltiples contactos del español en América, lo cual implicaría generar un corpus con un diseño diferente al actual. En la misma línea, sugiere diseñar un nuevo subcorpus de textos propios de los discursos técnicos y científicos.

[23] El tercer artículo, *Corpus electrónicos históricos y usuarios. Con atención especial al CORHEN*, de Rosa M. Espinosa Elorza y Andrzej Zieliński, se ocupa de delinear el camino hacia un mundo en donde los investigadores en lingüística histórica del español tendrían mucho más que datos disponibles: tendrían buena parte de los datos ya procesados. Se centran los autores en los corpus de especialidad, en particular, en los corpus históricos, y los analizan en diferentes planos que van desde lo gráfico-fonético hasta lo pragmático, pasando por lo morfológico, sintáctico y semántico, sin dejar de lado lo paleográfico y lo ecdótico. En cada uno de estos ángulos, ejemplifican 'defectos' de los actuales corpus y proponen soluciones, costosas muchas veces, aunque no por ello no deseables. Algunas apuntan a mejorar la calidad filológica, por ejemplo, correferir el pasaje de la concordancia o el documento en su totalidad con la imagen del original. Otra mejora comentada por los autores es facilitar el trabajo de los investigadores mediante

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

corpus etiquetados, ya que la anotación, en opinión de ellos, minimizaría el trabajo de expurgo o discriminación que deben hacer actualmente los investigadores al momento de evaluar las concordancias obtenidas.

[24] Los siguientes tres artículos presentan proyectos de corpus creados con objetivos diversos, con herramientas diferentes y que están en distintos estadios de avance. El texto *Los corpus digitales en el proyecto DiacOralEs*, de Santiago Del Rey Quesada y Elena Carmona Yanes, presenta un corpus de alto diseño cuyo foco amplio de interés es lograr la caracterización discursivo-tradicional del español en diferentes estados lingüísticos, tomando en cuenta, y en esto radica parte de su originalidad, la influencia de otras lenguas en estas configuraciones a partir de la traducción. Destacan los autores algunos aspectos novedosos del proyecto: el énfasis en períodos de lengua infrarrepresentados (el español moderno) y la presencia de las traducciones (y textos fuente) entendiendo la compulsa entre ellos como fuente de conocimiento lingüístico. Reflexionan con cuidado sobre las relaciones entre la lingüística de corpus y la traducción, mostrando un enfoque no aplicado de la primera para la segunda sino uno que apunta al mejor conocimiento del cambio lingüístico a partir de la objetivación de los procesos de traducción. Este artículo muestra, una vez más, que la propia creación y diseño de un corpus supone en sí mismo un proceso de investigación y de elaboración teóricas. Del Rey Quesada y Carmona Yanes valoran, finalmente, los pequeños corpus digitales que no apuntan a grandes datos, como es el caso del que proponen, el *Corpus del discurso dialógico en la historia de las lenguas romances* (*CorDisDial*), un conjunto de textos dialógicos producidos entre el siglo XV y el siglo XIX, cuyo proceso de creación describen pormenorizadamente. Finaliza el artículo con la descripción de la gestión de otro corpus, el *Corpus de textos periodísticos traducidos del grupo EHA* (*EHA-PRESTUS*), centrado en documentos de la prensa anterior a 1850, primera etapa del desarrollo del discurso periodístico en España y constituido por textos en francés, dado el peso del discurso elaborado en esa lengua en el protoperiodismo en lengua española.

[25] El artículo de Gabriela Cruz Volio, *El Corpus histórico del español de Costa Rica (COHIECOS)*, inicia con la argumentación de por qué es pertinente la creación de un corpus tan acotado geográficamente. La respuesta es que hay una muy escasa presencia de documentos de ese país en los corpus históricos ya existentes. Describe la autora las características y la conformación del *Corpus histórico del español de Costa Rica* (*COHIECOS*), que se compone de documentos jurídico-administrativos de distintos géneros (inventarios de diversa índole, informes, causas criminales, cartas de diversa naturaleza y propósito, testamentos, peticiones, demandas, causas criminales, acusaciones, denuncias, autos, quejas, relaciones), extraídos de fondos diversos (Mortuales coloniales, Complementario colonial, Cartago, Guatemala). Señala, por una parte, los criterios de unidad textual que integra e integrará este corpus: documentos o partes de documentos que inician con la mención del lugar y de la fecha de emisión del texto y finalizan con las

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

firmas del escribiente y de los testigos, y, por otra, los criterios geográficos, cronológicos y tipológicos: zona alta-zona baja para los geográficos; 1700-1739, 1740-1779 y 1780-1821 para los cronológicos, y una clasificación tipológica textual ya considerada y probada en otros corpus históricos. La autora aporta datos históricos y dialectales que fundamentan la selección de los dos primeros criterios. Señala que los documentos son transcritos paleográficamente y editados críticamente y que alcanzan en la actualidad 43000 palabras. Todos los documentos cuentan con metadatos (transcriptor, tipo-género textual, archivo de procedencia, síntesis de contenidos, fecha, localización, geolocalización y escriptor). Da cuenta a continuación de los criterios de transcripción y de la forma en que está empleando la herramienta *Transkribus*, una inteligencia artificial diseñada para la digitalización, el reconocimiento y la transcripción de textos, sobre todo a los efectos de la alineación. Evalúa positivamente el uso de esta herramienta que no ha requerido entrenamiento adicional al programa, aunque sí algún ajuste manual. Asimismo, explica otras funcionalidades de la herramienta en relación con el etiquetado y la exportación. La autora señala las dificultades de trabajar con un doble objetivo: el del empleo de *Transkribus* y los establecidos en el corpus *CHARTA*, y se detiene en los problemas que genera a la herramienta *Transkribus*, un material con elevada heterogeneidad gráfica, como es el recopilado por ella. Finalmente, da cuenta de la forma de anotación y etiquetado a través de TEI y de su expresión en el sistema *TEITOK*, así como de corpus antecedentes de esta opción, mostrando el potencial vínculo con otros corpus.

[26] El siguiente artículo, a cargo de Alejandro Fajardo Aguirre y Dolores Corbella Díaz, describe un recurso para la historia de los portuguesismos léxicos en el español: el *Observatorio de portuguesismos* (*OPORT*). Es una plataforma digital que, a partir de investigaciones centradas en la presencia del portugués en el léxico y la cultura del español, hace accesibles portuguesismos en la web, además de que permite la recuperación automática de material sobre la historia del patrimonio léxico común y permite su visualización cartográfica. Se centran los autores en una zona poco atendida a la fecha: los préstamos léxicos del portugués a español. Se trata, en rigor, de un corpus de fichas lexicográficas que cuentan con información etimológica, con información sobre su empleo, así como con información sobre su difusión geográfica, su clasificación ontológica y su documentación histórica. Con un decidido foco en los portuguesismos atlánticos, la revisión de fuentes metalingüísticas y fuentes textuales de diverso tipo arroja luz o modifica etimologías que hasta ahora eran consideradas seguras; el trabajo constituye un aporte a la determinación de la extensión geográfica e histórica de las voces tratadas. Este corpus lexicográfico cuenta con seis tipos de filtros (campo temático, localización geográfica, clase de palabra, fecha de documentación, origen inmediato y lengua de origen remoto) y seis niveles ontológicos (*general/abstracto*, *el individuo*, *vida humana*, *el mundo*, *cultura*, *economía y producción*, *ciencias humanas*, *ciencias experimentales* y *otros*), con varios subniveles internos.

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

[27] El penúltimo texto, de Marina Albers, *Tres corpus para el español del siglo XVIII. CHARTA, CORDIAM y un corpus jesuítico*, se ocupa de corpus ya instalados como herramientas en la comunidad científica y de un tercero, creación de la autora, que es explotado, por el momento, exclusivamente por ella. Este corpus está integrado por materiales inéditos producidos por jesuitas instalados en la región del Río de la Plata, escritos durante el siglo XVIII en la entonces llamada Provincia jesuítica del Paraguay, que comprendía en la época colonial tanto Paraguay como parte de los actuales territorios argentinos y uruguayos y la parte fronteriza de Brasil. Este corpus llena un vacío notorio en la documentación para estudios de historia de la lengua en el actual Paraguay, ya que es un país infrarrepresentado en los corpus. La autora explica la selección textual, en esencia documentos escritos por criollos, mayoritariamente epistolares, a partir de un material de más de 5600 documentos redactados por jesuitas en la Provincia jesuítica del Paraguay entre 1728 y 1765. Albers describe minuciosamente el proceso de informatización del corpus jesuítico y señala que el resultado no es automatizable para la lematización en la medida en que los materiales fueron transcriptos en paleografía estrecha y, por lo tanto, no están estandarizados ortográficamente. Expone la autora la necesidad de un trabajo manual de desambiguación en la etapa de etiquetado morfosintáctico, algo también esperado, y señala, como un déficit, la ausencia de etiquetado sintáctico en los dos corpus analizados por ella. Luego de una comparación entre los corpus *CHARTA* y *CORDIAM*, explica cómo y por qué confeccionó un subcorpus a partir de este último para alcanzar un grado de homogeneidad razonable con el fin de realizar las comparaciones dialectales que busca establecer. Realiza, asimismo, búsquedas comparadas que le permiten sopesar las virtudes y defectos de los tres corpus empleados en su investigación tanto en cuanto a las decisiones tomadas para su construcción como en cuanto a sus interfaces con el usuario.

[28] Cierra este número temático el trabajo de Paula Albitre Lamata *Pragmática histórica del español rioplatense. Actos directivos y estrategias de mitigación en correspondencia privada del siglo XIX*. Presenta la autora un estudio a partir de un corpus de 74 cartas privadas, dirigidas por remitentes mujeres a sus esposos, en las que analiza tanto las distintas construcciones lingüísticas en que se realizan actos de habla directivos como los valores de cortesía o de descortesía codificados mediante tales actos de habla, todo lo cual le permite poner en perspectiva histórica diferencias pragmáticas dialectales. El trabajo realizado por Albitre Lamata es totalmente manual, tanto en la creación del corpus, en el cual controla variables dialectales – español rioplatense –, temporales – siglo XIX – y discursivas – género epistolar –, así como control en la factura, circulación del texto y proximidad-distancia comunicativa entre interlocutores: son misivas de circulación privada, manuscritas por el remitente, que es siempre una mujer, y la relación entre emisora y destinatario es simétrica y de cercanía, pues entre ellos existe un vínculo afectivo-emocional de relación conyugal. Estas cartas fueron luego incor-

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

poradas, tras su adecuado tratamiento filológico e informático, a uno de los corpus históricos informatizados existentes, el *CORDIAM*.

5 Corpus históricos del español actualmente en funcionamiento

[29] Dado que de corpus históricos trata este número temático, nos ha parecido conveniente, y útil, hacer un listado y una brevísima descripción de los corpus históricos actuales disponibles en la red. Aparecen presentados en orden alfabético por su sigla o modo de empleo.

[30] *ADMyTE*, *Archivo digital de manuscritos y textos españoles*, es un corpus que abarca el español de la península ibérica en la Edad Media e incluye obras como el *Cantar de mio Cid*, el *Libro de Alexandre*, el *Libro de Apolonio* y «enciclopedias, diccionarios, gramáticas, novelas o *romans* (caballerescos y sentimentales), poemas narrativos en distintos metros (cuaderna vía, pareados o arte mayor), colecciones de cuentos y fábulas, cancioneros (individuales y colectivos), crónicas, biografías, tratados de religión y moralidad (escritos apologéticos, artes de bien morir, literatura ascética), traducciones bíblicas, tratados científicos de la más diversa naturaleza, manuales de medicina y veterinaria, fueros y ordenamientos legales, libros de viajes e itinerarios, tratados de música o traducciones de los clásicos grecolatinos, así como del árabe y del hebreo». Se puede consultar en el sitio web https://www.admyte.com y requiere de registro para acceder.

[31] *Biblia medieval* recoge distintas versiones medievales en castellano de la Biblia, y permite su consulta en paralelo. Se aloja en http://bibliamedieval.es y es de acceso libre.

[32] *CDH*, *Corpus del Nuevo diccionario histórico del español*, abarca textos literarios, ensayísticos, periodísticos y documentos del período 1064-2005 procedentes de España, Filipinas y América. Es de acceso libre: http://web.frl.es/CNDHE.

[33] *CHARTA*, *Corpus hispánico y americano en la red*. Textos antiguos, contiene «piezas oficiales de la cancillería, la administración civil, la justicia, la Inquisición y contratos de compraventa; cartas particulares, billetes y notas sueltas» de los siglos XII al XIX, tanto de Europa como de América y Asia. Es de acceso abierto, mediante el vínculo http://www.corpuscharta.es.

[34] *CHEM*, *Corpus histórico del español de México*, se enfoca en «diversos géneros textuales» producidos en México. Para ingresar, se debe crear un usuario en http://www.corpus.unam.mx:8080/unificado/index.jsp?c=chem.

[35] *CODEA+ 2022*, *Corpus de documentos españoles anteriores a 1900*, recoge documentos de archivo «desde la Cancillería a las notas de manos inhábiles», producidos en español peninsular desde sus orígenes hasta 1900. Es de acceso libre, a partir del link https://www.corpuscodea.es.

[36] *COHIECOS*, *Corpus histórico del español de Costa Rica*, incluye documentos de la época colonial, siglo XVIII hasta 1821. Son documentos jurídico-administrativos: textos de declaraciones (interrogatorios, informaciones, peticio-

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

nes, instrucciones, declaraciones de testigos, etc.) y testamentos e inventarios. Se accede a través de https://teitok.ucr.ac.cr, con registro previo.

[37] *CORDE*, *Corpus diacrónico del español*, aloja textos narrativos, líricos, dramáticos, científico-técnicos, históricos, jurídicos, religiosos y periodísticos, entre otros, de «todos los lugares donde se habló español» y abarcando desde los inicios de esta lengua hasta 1974. Su acceso es libre en el siguiente vínculo: https://corpus.rae.es/cordenet.html.

[38] *CORDEREGRA*, *Corpus diacrónico del español del Reino de Granada*, abarca el período 1492-1833 en el Reino de Granada. Actualmente se encuentra disponible en internet en su nueva etapa *Oralia diacrónica del español* (*ODE*) (ver abajo).

[39] *CORDIAM*, *Corpus diacrónico y diatópico del español de América*, reúne tres conjuntos documentales: *Cordiam-Documentos*, *Cordiam-Literatura* y *Cordiam-Prensa*, que recogen textos en español americano de 1494 a 1905, escritos en América y, en su gran mayoría, por americanos hispanohablantes nativos; en el caso del subcorpus de prensa, el período abarca los siglos XVIII y XIX, porque, como es sabido, no había prensa, entendida en su sentido actual, en periodos previos. Su acceso es libre en https://www.cordiam.org.

[40] *COREECOM*, *Corpus electrónico del español colonial mexicano*, aloja documentos del siglo XVI al XIX producidos en España y en el Virreinato de la Nueva España. Contiene cartas personales, cartas de relación, informaciones, relaciones de cargas, testamentos, cartas de autodefensas, cartas de acusación, notas y recibos, y se puede consultar de forma libre en https://www.iifilologicas.unam.mx/coreecom.

[41] *CORHEN*, *Corpus histórico del español norteño*, contiene documentos privados de la Edad Media producidos en variedades castellanas norteñas. Se aloja en https://corhen.es y es de acceso libre.

[42] *CorLexIn*, *Corpus léxico de inventarios*, recoge «inventarios, tasaciones, cartas de arras y, en general, las relaciones de bienes conservadas en los registros notariales» del Siglo de Oro en España. Las consultas son libres en el siguiente vínculo https://apps2.rae.es/CORLEXIN.html.

[43] *Corpus del español*, en su subcorpus *Genre/Historical*, contiene archivos del período 1200-1900 en España y Latinoamérica: textos orales, de ficción, de periodismo y académicos. Se accede de forma libre en https://www.corpusdelespanol.org/hist-gen.

[44] *Léxico hispanoamericano*, aloja materiales escritos de América producidos entre 1493 y 1993. Se trata de «documentos comerciales, contratos, testamentos, cartas, relaciones, actas de cabildo, y copiosos informes relativos al Estado, a la Iglesia, y a la Inquisición. Crónicas, obras literarias y didácticas, correspondencia particular». Se encuentra en el link https://textred.spanport.wisc.edu/lexico_hispanoamericano/index.html y se puede consultar sin restricciones.

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815

[45] *ODE*, *Oralia diacrónica del español*, es un corpus que contiene inventarios de bienes, declaraciones de testigos en juicios penales y certificaciones de barberos y cirujanos producidos entre 1492 y 1833 en España. Se puede acceder de forma libre en http://corpora.ugr.es/ode.

[46] *Post Scriptum* recoge cartas privadas de la Edad Moderna, y algunas pocas de finales del siglo XV, producidas en España en su mayoría y algunas pocas en América y Filipinas. Su acceso es libre a través del siguiente link: http://teitok.clul.ul.pt/postscriptum/index.php.

DOI: https://doi.org/10.25364/19.2024.12.1 ISSN: 2663-9815