Blas Arroyo, José Luis; Puertas Ribés, Elia (2024): INCOM: Un corpus de inmediatez comunicativa para el estudio sociolingüístico del español en su historia. *Cultura, Lenguaje y Representación,* Vol. XXXIII, 7–29
ISSN 1697-7750 · E-ISSN 2340-4981
https://doi.org/10.6035/clr.7200

INCOM: Un corpus de inmediatez comunicativa para el estudio sociolingüístico del español en su historia[1]

INCOM: A corpus of communicative immediacy for the sociolinguistic study of Spanish through history

JOSÉ LUIS BLAS ARROYO
UNIVERSITAT JAUME I
https://orcid.org/0000-0002-6700-0068

ELIA PUERTAS RIBÉS
UNIVERSITAT JAUME I
https://orcid.org/0000-0002-5653-6854

Artículo recibido el / *Article received*: 2023-02-26
Artículo aceptado el / *Article accepted*: 2023-05-11

RESUMEN: En el marco del debate acerca de la necesaria presencia de géneros discursivos cercanos al polo de la inmediatez comunicativa, como mejor estrategia para aproximarnos a la oralidad de tiempos pretéritos, en el presente artículo se describen los principios hermenéuticos y metodológicos de un corpus compilado a lo largo de la última década por el grupo de investigación *Sociolingüística*, de la Universitat Jaume I. Integrado por tradiciones escritas de impronta oral, mayoritariamente correspondencia privada, a la que se añaden en menor proporción algunos géneros autobiográficos (diarios, memorias de servicios, crónicas de soldados), el corpus supera ya los catorce millones de palabras, escritas por cerca de siete mil españoles de diferente extracción social y dialectal entre finales del siglo XV y la primera mitad del XX. A partir de las limitaciones que para el análisis sociolingüístico presentan otros corpus, en el artículo se revisan los principales fundamentos que guían la arquitectura de este, entre las que sobresale la necesidad de contar

[1] El presente estudio forma parte del Proyecto de investigación "Componentes socioestilísticos, idiolectales y discursivos en la variación y el cambio lingüístico en español: contribuciones desde la sociolingüística histórica (2022-2026)", financiado por el Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación y por fondos FEDER Una manera de hacer Europa (Ref. PID2021-122597NB-I00).

DOI: https://doi.org/10.6035/clr.7200

con una selección suficientemente amplia y representativa de los diferentes periodos estudiados, haciendo posible así la investigación sobre variables (morfosintácticas, léxicas, discursivas) que, por su propia naturaleza, poseen escasa recurrencia en el discurso. Tipológicamente, se configura, pues, como un corpus *específico*, por la finalidad de su aprovechamiento, eminentemente sociolingüística; y *secundario*, dada la explotación que en él se hace de textos editados previamente, aunque seleccionados de acuerdo con parámetros rigurosos.

Palabras clave: sociolingüística histórica, lingüística de corpus, corpus de inmediatez comunicativa, tradiciones discursivas.

ABSTRACT: In the context of the debate on the necessary presence of discursive genres close to the pole of communicative immediacy as the best strategy for approaching the vernacular language of the past, this article describes the hermeneutic and methodological principles of a corpus built up over the last decade by the *Sociolinguistic* research group (Universitat Jaume I). Composed of written traditions with an oral imprint, mostly private correspondence and a smaller representation of several autobiographical genres (diaries, service memoirs, soldiers' chronicles), the corpus now exceeds fourteen million words, written by almost seven thousand Spaniards from different social and dialectal backgrounds between the end of the fifteenth century and the first half of the twentieth century. Given the limitations of other corpora for historical sociolinguistic analyses, the article outlines the main principles that guide the architecture of the corpus, including the need to have a sufficiently broad representation of the different periods studied, making it possible to study variables (morphosyntactic, lexical, discursive) that by their nature have little recurrence in discourse. In the end, it is conceived as a *specific* corpus, due to its eminently sociolinguistic purpose; and *secondary* corpus, due to the use of previously published texts, albeit selected according to strict parameters.

Key words: historical sociolinguistics, corpus linguistics, corpus of immediacy text, discursive traditions.

1. INTRODUCCIÓN

Como disciplina académica enfrentada al análisis de los factores que condicionan la variación y el cambio lingüístico, la lingüística histórica se ha visto enormemente favorecida por el desarrollo de las nuevas herramientas digitales que ha venido proporcionando de un tiempo a esta parte la lingüística de corpus. El método tradicionalmente empleado para la reconstrucción diacrónica de la lengua, que consistía en revisar y recontar manualmente los datos extraídos de diferentes textos, se ha visto sustituido en la actualidad por el uso cada vez más frecuente de corpus informatizados, que permiten compilar automáticamente –y en un tiempo incomparablemente corto– un número mucho mayor de testimonios. Estas herramientas computacionales se ponen a disposición del investigador como recurso indispensable no solo para agilizar y automatizar el proceso de obtención de los datos, sino también para perfeccionar su interpretación cualitativa y cuantitativa (Enrique Arias, 2012: 86).

Los corpus diacrónicos han crecido significativamente desde las postrimerías del siglo pasado, cuando vieron la luz los primeros bancos de datos. Para lenguas como el inglés cabe recordar, por ejemplo, hitos fundacionales como el *Helsinki Corpus of English Texts*, en el que se reunían textos que abarcaban casi un milenio (siglos VIII al XVIII). Por fortuna, los estudiosos del español empezaron a contar también por las mismas fechas con algunos repositorios de enormes dimensiones, como el *Corpus Diacrónico del Español* (CORDE), desarrollado por la Real Academia Española, con doscientos cincuenta millones de registros desde los orígenes del español hasta 1974; o el *Corpus del Español* (CdE), compilado bajo la dirección del profesor Mark Davies (Davies, 2002), que reunía una cifra aproximada de cien millones de palabras entre los siglos XIII y XX. En los últimos años, a estos se han sumado otros proyectos más específicos y de menores dimensiones, que compensan, sin embargo, con algunas mejoras notables tanto en aspectos lingüísticos y filológicos como computacionales (CODEA +2022, *Biblia Medieval*, *Post Scriptum*, CORDIAM, *Oralia diacrónica*, CorLexIn, etc.).

La proliferación de estos corpus ha propiciado en paralelo una reflexión crítica acerca de las ventajas, pero también de las limitaciones que presentan estas nuevas herramientas. Y es que, pese a haber abierto sin duda horizontes nuevos y líneas de investigación muy prometedoras en el estudio diacrónico del español –insospechadas en buena medida hace apenas unos años–, los corpus históricos no están exentos de problemas, que lastran algunas de sus posibilidades hermenéuticas. Estos son de diferente naturaleza, y entre ellos cabe citar algunas lagunas sobre las que se ha insistido recientemente, como los problemas en la datación de los documentos –especialmente en el periodo medieval (Rodríguez Molina y Octavio de Toledo y Huerta, 2017)–, las dificultades a la hora de recuperar la información contextual deseable (Enrique Arias, 2012; Díaz Bravo, 2018; de Benito, 2019), la ausencia de sistemas de lematización y etiquetado en la mayoría (Enrique Arias, 2009; García Salido y Vázquez Rozas, 2012; Díaz Bravo, 2018), el desequilibrio en la representatividad de las tipologías textuales (Kabatek, 2013; Rodríguez Puente, 2018) o las dimensiones variables de los corpus, que dificultan los análisis comparativos (Caravedo, 1999; Clavería Nadal, 2012; Rodríguez Puente, 2018), entre otras.

Otro de los temas recurrentes en este debate es la necesaria incorporación de géneros y tradiciones discursivas que, en la medida de lo posible, nos acerquen a la oralidad de tiempos pretéritos (Rodríguez Puente, 2018; Calderón y Vaamonde, 2020). Afortunadamente, esta laguna endémica de la lingüística histórica ha empezado a ser corregida en los últimos años de manera entusiasta por algunos investigadores dentro y fuera de nuestras fronteras. Así, son conocidos algunos corpus diacrónicos que han ofrecido datos muy relevantes acerca de la evolución del inglés y su conexión con ciertas variables extralingüísticas, como el *Corpus of Early English Correspondence*, el *Corpus of English Dialogues 1560-1760* o el *Old Bailey Corpus*. Lo mismo sucede en la tradición hispánica con la incorporación de diferentes tradiciones discursivas que, aun trasladadas al medio escrito, se aproximan a una concepción oralizante y cotidiana de la lengua, cercana, pues, al polo de la inmediatez comunicativa (Oesterreicher, 2004). Entre estas destacan, por ejemplo, las crónicas de soldados (Di Tullio y Resnik, 2019), las quejas (Octavio de Toledo y Huerta y Pons Rodríguez, 2017), las peticiones de ayuda a la beneficencia (Sánchez-Prieto Borja y Vázquez Balonga, 2019), los diálogos (Navarro Gala, 2020), las declaraciones de testigos (Calderón, 2015; Calderón y Vaamonde, 2020), los inventarios de bienes (Morala, 2012; Calderón y Vaamonde, 2020), la correspondencia privada (Fontanella de Weinberg, 1992; Fernández Alcaide, 2009; Arias

 DOI: https://doi.org/10.6035/clr.7200

Álvarez y Hernández Mendoza, 2013; Vaamonde, 2018), los diarios y otros géneros autobiográficos (Rivadeneira y Contreras, 2021; Frühbeck, 2022), etc.

En este marco es, precisamente, en el que se inscribe el proyecto llevado a cabo por el grupo de investigación consolidado *Sociolingüística*, de la Universitat Jaume I, al que pertenecen los autores de este artículo, y que a lo largo de la última década ha reunido un corpus compuesto por catorce millones de registros en el momento de escribir estas líneas.[2] Tanto sus dimensiones como la arquitectura del corpus, integrado en su mayor parte por correspondencia epistolar escrita por miles de españoles de diferente extracción social y dialectal a lo largo de cinco siglos, ha permitido el desarrollo de diversos proyectos de investigación en sociolingüística histórica, más concretamente en la vertiente de esta disciplina enfrentada al análisis minucioso de fenómenos de microvariación y cambio lingüístico, poco explorada hasta la fecha en el estudio del español.[3] Tipológicamente, se configura como un corpus *específico*, por la finalidad de su aprovechamiento, eminentemente sociolingüística; y *secundario*, dada la explotación que en él se hace de textos editados previamente,[4] aunque seleccionados de acuerdo con parámetros rigurosos (para más detalles sobre esta cuestión ver § 5).

La estructura del trabajo queda como sigue. En el apartado 2 repasamos algunas aportaciones relevantes acerca de la presencia de la oralidad en los textos escritos y las decisiones adoptadas en nuestro corpus a partir de ellas (§ 2.1), así como el modo en que tal presencia se visibiliza en otros corpus actualmente disponibles (§ 2.2). A partir de las limitaciones que para nuestro objeto de estudio presentan estas bases de datos, en el apartado 3 se revisan las principales bases científicas que guían su arquitectura. A continuación, en § 4 se justifican las razones de un corpus que, sin renunciar a una posible incorporación futura de anotaciones lingüísticas, se presenta a día de hoy como no anotado. Los criterios que han llevado a la selección de las ediciones son objeto de atención en § 5, un apartado que se completa en § 6 con algunos ejemplos prácticos, derivados de nuestras propias investigaciones, acerca de las posibilidades que ofrecen herramientas externas como *WordSmith Tools* (WS) para la búsqueda de los datos lingüísticos deseados.

2. LA ORALIDAD EN EL ESTUDIO DIACRÓNICO DEL ESPAÑOL

La presencia de la oralidad en la escritura ha adquirido un destacado interés en los estudios de lingüística histórica de las últimas décadas. Sin embargo, la ausencia lógica de fuentes orales hasta tiempos recientes ha supuesto un problema de primer orden para los historiadores de la lengua, quienes se han visto obligados a trabajar con testimonios de naturaleza estrictamente escrita en el intento por aproximarse a esa oralidad. Es la conocida paradoja lavobiana, según la cual: «Historical linguistics can then

[2] Junto a los dos autores que firman el presente trabajo, al grupo de investigación pertenecen en la actualidad los siguientes profesores de la Universitat Jaume I: Isabel Andúgar, Agnese Sampietro, Kim Schulte, Mónica Velando y Javier Vellón, a los que se une el profesor Carles Navarro, de la Universidad de Valladolid. Para las actividades desarrolladas por este grupo en el ámbito de la sociolingüística sincrónica y diacrónica, ver el siguiente enlace:
https://sociolinguisticawe.wixsite.com/sociolinguisticauji

[3] Al mencionado en la nota 1 se añaden los proyectos FFI2010-15280, FFI2013-44614-P y FFI2017-86194-P, financiados igualmente en convocatorias competitivas con cargo a fondos de investigación estatales y europeos.

[4] Los lectores interesados pueden acceder al corpus a través del siguiente enlace:
https://sociolinguisticawe.wixsite.com/sociolinguisticauji/blank-tntpi

be thought of as the art of making the best use of bad data» (Labov, 1994: 11). Sin embargo, estas dificultades no han arredrado a los investigadores y, de hecho, los avances realizados en este ámbito a lo largo de las últimas décadas han sido notables.

El presente apartado se estructura en dos secciones. En el primero (§ 2.1), se reseñan brevemente algunas aportaciones teóricas recientes en las que se ha reflexionado acerca de los rasgos que permiten vislumbrar muestras de la oralidad en la escritura, y que han servido como base para la selección de las tradiciones discursivas incluidas en nuestro corpus. A continuación, en § 2.2 se aborda el modo en que otras bases de datos textuales actualmente en circulación dan cuenta de tales tipologías.

2.1. EL POLO DE LA INMEDIATEZ COMUNICATIVA: LOS TEXTOS DE COMPETENCIA ESCRITA DE IMPRONTA ORAL

Partiendo del modelo del *continuum* concepcional entre la oralidad y la escrituralidad propuesto por Koch y Oesterreicher (2007 [1990]), el polo de la inmediatez comunicativa se define de acuerdo con parámetros diversos, como la privacidad de la comunicación, la libertad temática, la implicación emocional, la cooperación entre los hablantes, la espontaneidad, el dialogismo o el saber compartido entre los interlocutores, entre otros. Algunos años más tarde, Oesterreicher (2004) reflexionaría también acerca de diferentes situaciones comunicativas que favorecen la producción de lo hablado en lo escrito.

Entre las que aquí más nos interesan, Oesterreicher destacaba los textos caracterizados por ser de *competencia escrita de impronta oral*, en los que a menudo es posible advertir una formación incompleta por parte del escritor, un desconocimiento –total o parcial– de la variedad lingüística empleada o un uso inadecuado de las reglas que estructuran el discurso. De ello se derivan ciertos rasgos estructurales, como a) la simplicidad expresiva; b) la existencia de frecuentes descuidos, propios de la espontaneidad y la familiaridad que envuelven a las situaciones comunicativas de este tipo; c) la adaptación de lo escrito a las posibilidades de comprensión del lector; o d) la presencia significativa de formas vernáculas derivadas del origen de los autores, el contacto de variedades y lenguas, etc. Se trata, en definitiva, de testimonios en los que se advierten fenómenos marcados desde el punto de vista dialectal, diastrático o diafásico, y que, en consecuencia, se sitúan en el extremo opuesto a la escrituralidad concepcional y la distancia comunicativa.

Entre las diferentes situaciones comunicativas "ideales" que favorecen la producción de este tipo de textos, en el corpus hemos seleccionado preferentemente la correspondencia privada, escrita por miles de individuos pertenecientes no solo a las élites sociales, sino también a otros grupos sociales menos favorecidos, incluidas gentes escasamente instruidas o semialfabetizadas, y, por tanto, menos condicionadas por las convenciones discursivas y las reglas del estándar.

Investigar la lengua del pasado a partir de epistolarios privados se ha convertido en una tarea cada vez más frecuente, alentada por motivos diversos. Desde un punto de vista estructural, las cartas muestran una serie de rasgos que las sitúan paradigmáticamente en el polo de la inmediatez, como la privacidad, la familiaridad entre los interlocutores, o la implicación emocional (Koch y Oesterreicher, 2007: 29–30). Además, el hecho de que su redacción no se conciba para una publicación posterior garantiza una mayor espontaneidad y, por tanto, una mejor aproximación al habla cotidiana que la reflejada en tradiciones discursivas más formales (van der Wal y Rutten,

 DOI: https://doi.org/10.6035/clr.7200

2013). Todo ello avala la correspondencia como un instrumento apropiado para la investigación diacrónica, en especial para el estudio de fenómenos vernáculos. Ciertamente, no faltan elementos de automatismo expresivo en esas cartas (sobre todo, en secuencias periféricas como encabezamientos y finales), y, en ocasiones, en su interior encontramos variantes sintácticas inicialmente poco previsibles en esta clase de textos (Cano Aguilar, 1996).[5] Pero, aun así:

> […] when persons who have had but limited experience in writing and exposure to the norms of written expression are forced to write nevertheless, their writing reflects many features of their speech fairly accurately: what they do is put their own "imagined" words onto paper, if only with difficulty (Schneider, 2013: 64).

Asimismo, las cartas representan un material excelente para el estudio de la variación y el cambio desde una perspectiva idiolectal, especialmente las de aquellos autores que escribieron su correspondencia durante décadas, lo que resulta muy útil para valorar el papel de los individuos en relación con los cambios lingüísticos coetáneos (Raumolin-Brunberg, 2009; Blas Arroyo, 2022).

Desde otro punto de vista, las cartas proporcionan también detalles contextuales sumamente relevantes para el análisis sociolingüístico, como las relaciones de poder y solidaridad entre los participantes, su condición social y dialectal, etc. (Nevalainen y Raumolin-Brunberg, 2017 [2003]). En nuestro caso, a todo ello hay que sumar la excepcional circunstancia histórica que supusieron las migraciones de españoles a América, lo que generó un caudal inmenso de cartas a uno y otro lado del Atlántico, en las que se trasladaban al papel historias personales colmadas de afectividad.

Aunque en proporciones mucho menores (para más detalles, ver § 3), el corpus se completa con un segundo grupo de textos, caracterizados también por una concepción oralizante de la escritura. Se trata de un conjunto de obras de carácter autobiográfico, correspondientes a diversas tradiciones discursivas. Entre ellas, sobresalen los diarios, a los que se suma una pequeña representación de otros géneros, como memorias de servicios, apuntes y notas personales, libros de familia, libros de cuentas, relaciones y crónicas de soldados, etc. Sin embargo, en el corpus no se han incluido otros textos autobiográficos, como, por ejemplo, aquellos de carácter espiritual que tanto éxito tuvieron en la España del Siglo de Oro, ya que, en el mejor de los casos, en ellos podemos encontrar una escritura en "estilo llano" que, sin embargo, responde a un interés fundamentalmente estilístico a cargo de escritores "profesionales" (Oesterreicher, 2004: 754). Del mismo modo, tampoco forman parte del corpus otros documentos en los que, también por razones interesadas, se traslada al papel una especie de oralidad simulada, como sucede con el género picaresco en el periodo áureo, las comedias y sainetes populares entre los siglos XVII y XIX, o textos similares, en los que, como destacaba Oesterreicher (2004: 756), «es el autor del texto, o sea, la conciencia lingüística del autor, la que selecciona ciertos rasgos lingüísticos considerados característicos de la lengua hablada». Pese a que autores como Culpeper y Kytö (2010: 16) avalan las ventajas de esta clase de textos en el estudio de la oralidad por su carácter interaccional, el hecho de cubrir un espectro social amplio o disponer de ellos en cantidades razonables –variables, en todo caso, según los géneros y periodos históricos–, a nuestro juicio, estas se ven superadas

[5] Culpeper y Kytö (2010: 16) han destacado otros inconvenientes, como el hecho de que las cartas no respondan propiamente a una verdadera interacción verbal cara a cara.

por las limitaciones reseñadas, lo que justifica que no se hayan considerado en nuestra compilación.

En el siguiente apartado veremos cuál es la presencia de estas tradiciones discursivas –*speech-related texts* en la terminología de Culpeper y Kytö (2010)– en algunos corpus diacrónicos actualmente disponibles.

2.2. REPRESENTACIÓN DE LA ORALIDAD EN OTROS CORPUS

Comenzamos esta revisión por los dos principales corpus diacrónicos de referencia del español, el conocido *Corpus Diacrónico del Español* (CORDE), desarrollado por la Real Academia Española, y el *Corpus del Español* (CdE), compilado bajo la dirección del profesor Mark Davies. Como se ha mencionado ya, ambos se han convertido en herramientas imprescindibles para la investigación histórica del español, si bien muestran importantes desequilibrios tanto en relación con los periodos representados, como en los géneros y tradiciones discursivas que los integran.

Los textos que conforman el CORDE son fundamentalmente de carácter formal, con una presencia masiva de géneros ubicados en la distancia comunicativa. De estos, los documentos literarios (novelas, cuentos, obras dramáticas, poesía, etc.) constituyen un 44 % del total, mientras que los no literarios –científicos, tecnológicos, jurídico–administrativos, didácticos, entre otros– representan el 56 % restante. No obstante, en este último grupo, el CORDE integra tres ámbitos temáticos que podrían vincularse con las tradiciones discursivas que aquí nos interesan. Por un lado, recopila una cifra no despreciable de cartas (en torno a cuatro mil), con casi seis millones de palabras en su interior, lo que representa un 2,4 % del total de registros del corpus. Lamentablemente, sin embargo, la utilidad de estos materiales para el estudio sociolingüístico es limitada, pues la mayor parte de esa correspondencia disponible fue redactada por individuos pertenecientes a las élites sociales y culturales de cada época, mientras que la escrita por otros sectores sociales es mucho menor. La misma limitación presenta un segundo bloque de *memorias y diarios*, que supera los tres millones y medio de palabras (1,5 % del total), y en el que, además, junto a la presencia de obras prototípicas de estos géneros autobiográficos, aparecen otras que difícilmente encajan en tales categorías, como las cartas cruzadas entre individuos nuevamente adscritos a las élites (Rodríguez Puente, 2018: 98–101). Por último, otros ámbitos, como los de *prosa dramática breve y extensa* o *prosa narrativa: diálogos y misceláneas*, con cerca de dos millones y medio de palabras (1 %), no dejan de ser textos literarios en los que prima el interés estilístico de los autores por "imitar" la oralidad, antes que un verdadero reflejo de esta.

Al igual que el CORDE, el *Corpus del Español* se presenta como una fuente muy útil para el estudio diacrónico del español, por los más de 20.000 textos recopilados y los cien millones de palabras que albergan en su interior, así como algunos recursos computacionales de los que carece el corpus académico,[6] como la lematización y el etiquetado morfosintáctico de las formas lingüísticas, lo que facilita las búsquedas. Ahora bien, además de no ofrecer una relación completa de esos textos –lo que, afortunadamente, sí hace el CORDE–, estos proceden de nuevo mayoritariamente de universos discursivos propios de la distancia comunicativa, como los géneros jurídicos y

[6] No obstante, el más reciente corpus para el *Diccionario histórico de la lengua española* (CDH) incorpora en su versión nuclear la lematización de una parte de los materiales del CORDE, así como un sistema de preanotación morfosintáctica llevada a cabo con herramientas de software libre (*Freeling*).

 DOI: https://doi.org/10.6035/clr.7200

administrativos, los textos científico-técnicos o los documentos literarios, entre otros. Únicamente en el siglo XX es posible encontrar testimonios de la oralidad, aunque no prototípicamente cercanos al polo de la inmediatez, como sucede con entrevistas, transcripciones de congresos, discursos, etc.

Los problemas reseñados en estos grandes corpus de referencia –junto a otros en los que no entraremos por cuestiones de espacio y oportunidad– han impulsado en los últimos años la creación de nuevas bases de datos textuales, cuyas dimensiones, más reducidas, se compensan con avances nítidos en otras esferas, entre las que se halla el interés por incluir géneros discursivos más cercanos a la oralidad. Diversos grupos de investigación han venido trabajando en los últimos años en esta empresa, como los proyectos CHARTA (Isasi, Pierazzo y Spence, 2020), CODEA +2022 (Sánchez-Prieto, 2012), *CorLexIn* (Morala, 2012), CORDEREGRA (Calderón, 2015), *Oralia diacrónica del español* (Calderón y Vaamonde, 2020), CORDIAM (Bertolotti y Company, 2014) o *Post Scriptum* (Vaamonde, 2018), por mencionar solo algunos de los más conocidos. Algunos de ellos presentan tipologías textuales de todo tipo, desde textos jurídico-administrativos a literarios y cronísticos, pasando por documentos privados (CHARTA, CODEA, CORDIAM), al tiempo que otros se concentran justamente en los de este último tipo. Así, el corpus *CorLexIn* se especializa en inventarios de bienes del siglo XVII repartidos por toda la geografía española. Textos similares, redactados entre 1492 y 1833, se incluyen también en CORDEREGRA y *Oralia diacrónica*, aunque con límites geográficos diferentes en estos dos proyectos sucesivos impulsados desde la Universidad de Granada (provincias del antiguo Reino de Granada en el primer caso, que en el segundo se extienden también al centro-norte peninsular). *Oralia diacrónica*, por su parte, amplía la nómina de tipologías discursivas de impronta oral para incluir también certificaciones periciales a cargo de cirujanos. Finalmente, un proyecto con especiales conexiones con el presentado en estas páginas es *Post Scriptum*, *Archivo digital de escritura cotidiana en Portugal y España en la Edad Moderna*, un corpus integrado completamente por correspondencia privada hallada en expedientes judiciales entre 1517 y 1833, y con una representación cercana al millón de palabras para cada lengua.

Pese a las ventajas indudables que presentan estos corpus, no solo por las tipologías consideradas en ellos, sino también por los avances indiscutibles en materia de edición, presentación y etiquetado de los materiales, algunas de sus características los convierten en herramientas incompletas para nuestros intereses. De ahí la necesidad de compilar un nuevo corpus que dé satisfacción a estos y cuyas bases científicas exponemos en el siguiente apartado.

3. BASES PARA LA COMPILACIÓN DEL CORPUS

La delimitación clara de los objetivos representa una clara ventaja metodológica a la hora de seleccionar los textos de un corpus específico (Enrique Arias, 2009; Sánchez-Prieto, 2012). Y así, de la misma forma que algunos se conciben como base para la elaboración de diccionarios o gramáticas históricas de la lengua –el caso, entre nosotros, del *Corpus del nuevo diccionario histórico del español* (CDH)– el que presentamos en estas páginas está diseñado para proporcionar materiales suficientemente amplios y representativos para el análisis sociolingüístico de variables –morfosintácticas, léxicas, unidades fraseológicas, marcadores discursivos, etc.– que, por su propia naturaleza, ofrecen a menudo una baja recurrencia en el discurso. Además de acudir a textos cercanos al polo de la inmediatez como forma más adecuada de aproximarnos al habla cotidiana

del pasado, el corpus precisa de tres requisitos indispensables: a) una profundidad diacrónica suficiente, que haga posible la comparación entre cortes históricos diferentes en la evolución de una determinada variable lingüística; b) un plus de exhaustividad; y c) el acceso a la mayor información contextual posible.

Por lo que al primer aspecto se refiere, los textos compilados en el corpus abarcan casi cinco siglos, en concreto desde 1492 a 1960. En estos límites no se contempla, pues, el periodo medieval, sin duda, esencial para entender la evolución del español, pero cuyo análisis presenta dificultades añadidas –fiabilidad de las dataciones, contextualización fragmentaria o incompleta, documentos conservados a cargo casi exclusivamente de las élites, etc.– que sin duda se agravan en las tradiciones discursivas aquí consideradas. Aun así, el tramo temporal contemplado en el corpus incluye períodos decisivos en la periodización del español, que van desde el primer español clásico, a la lengua de mitad del siglo XX. De hecho, a diferencia de otros corpus, que se detienen en diversos momentos del siglo XIX, en nuestro caso hemos decidido prolongar este último límite tras comprobar que algunos patrones de variación y cambio lingüístico experimentan modificaciones significativas en las primeras décadas del XX respecto al español decimonónico, con independencia de que la norma del idioma se hubiera consolidado ya en buena medida una centuria antes.

Ahora bien, por amplio que sea el lapso temporal estudiado, un análisis efectivo de la variación tan solo es posible si contamos con muestras suficientemente extensas en cada periodo, que permitan examinar el condicionamiento lingüístico y extralingüístico con ciertas garantías. Más arriba hemos visto ya algunas de las dificultades que plantean los corpus más extensos para el estudio de este condicionamiento, especialmente el de carácter extralingüístico. Así, pese a presentar un número nada desdeñable de cartas, estas se hallan irregularmente representadas en el CORDE, por no hablar del hecho de que, en su inmensa mayoría, han sido escritas por las élites, además de no ofrecer información suficiente sobre los diferentes tipos de misivas, un dato fundamental para analizar la variación estilística. Claro que las dificultades para el análisis diastrático y diafásico todavía son mayores en el CdE, pues ni siquiera se ofrece una relación completa de los textos incluidos. Por su parte, otros corpus más recientes, mucho más cuidadosos a la hora de proporcionar esta clase de informaciones, además de modélicos en su rigor filológico y metodológico, presentan el problema de una representación insuficiente de ciertas variables caracterizadas por una escasa presencia en el discurso. Este es el caso, por ejemplo, de *Post Scriptum*, cuyos novecientos mil registros para el español, repartidos en más de dos mil trescientas cartas, se quedan cortos para analizar la evolución de algunas variables morfosintácticas. Así, una búsqueda en esos materiales de la variación entre las perífrasis *deber* y *deber de + infinitivo* en el español clásico ofrece cifras insuficientes para el estudio sociolingüístico. Por ejemplo, para el siglo XVI, el corpus arroja tan solo 24 muestras de la variable, lo que impide realizar un análisis de regresión que permita evaluar la potencial relevancia de ciertos factores lingüísticos, estilísticos y sociales que hemos estudiado en otro lugar (Blas Arroyo, 2016). Además, una simple comparación de las frecuencias de cada una de las variantes (*deber de*= 14; *deber*= 10) ofrece una visión distorsionada de la relevancia que cada una de estas formas ha tenido a lo largo de la historia.

Para asegurar una representación, si no óptima, al menos apropiada para la realización de estudios sociolingüísticos sobre variables morfosintácticas, léxicas o discursivas, así como un acercamiento al cambio lingüístico con suficiente profundidad histórica, nuestro objetivo inicial fue la compilación de al menos dos millones de registros

 DOI: https://doi.org/10.6035/clr.7200

por centuria. En el momento presente, este objetivo se ha cumplido ya con creces, como se puede comprobar en la Tabla 1, en la que se detalla, además, el número de archivos y escritores recogidos en cada caso.

Tabla 1. Algunas magnitudes del corpus

	N.º archivos	N.º registros	N.º escritores
Siglo XVI	2.195	2.407.913	1.395
Siglo XVII	2.314	2.611.363	1.075
Siglo XVII	4.156	3.284.210	2105
Siglo XIX	2.436	2.472.211	1.223
Siglo XX (h. 1960)	3.909	3.504.274	1.015
Total	15.010	14.279.971	6.813

El corpus muestra algunos desequilibrios muestrales, entre los que sobresale una cierta sobrerrepresentación de los siglos XVIII y XX (no solo en la cantidad de palabras disponibles, sino también –al menos en el XVIII– en el número de escritores), y una menor representación del resto (en todo caso, siempre por encima del límite proyectado). El peso de las tradiciones discursivas compiladas en el corpus es, sin embargo, muy diferente. Como se puede ver en la Tabla 2, la correspondencia epistolar representa un 86 % del total, frente a un 14 % para los géneros autobiográficos. Además, esa desproporción se agudiza en los primeros siglos.

Tabla 2. Distribución de los materiales del corpus por tradiciones discursivas

	Correspondencia		Resto	
	N.º registros	%	N.º registros	%
Siglo XVI	2.108.046	95	96.550	5
Siglo XVII	2.313.517	89	297.846	11
Siglo XVIII	2.879.504	88	404.706	12
Siglo XIX	1.888.899	77	583.312	23
Siglo XX (h. 1960)	2.938.819	84	565.455	16
Total	12.128.785	86	1.947.869	14

Como es lógico, hay también mucha más diversidad autoral en la correspondencia epistolar que en los géneros autobiográficos, en los que la nómina de escritores es reducida. Ahora bien, dentro de la correspondencia encontramos también importantes diferencias en el número de misivas a cargo de cada autor, con cifras que van desde la carta aislada en un extremo, a decenas de ellas en el opuesto. Pese a ello, por lo general hemos optado por no limitar la representación de las cartas en los epistolarios más nutridos. Además de sacar provecho de todos los recursos disponibles, algunas herramientas estadísticas habituales en la práctica variacionista, como el análisis de regresión de efectos mixtos, permiten aquilatar más adelante la relevancia de los factores considerados en el análisis, filtrando los resultados a través del tamiz de la variación individual. Por lo demás, esta variación idiolectal resulta sumamente interesante, no solo para valorar la (in)congruencia de las distribuciones individuales y colectivas, o la existencia de potenciales cambios a lo largo de la vida de los individuos, sino también para comparar tales distribuciones con las que estos mismos individuos presentan en otras

tradiciones discursivas más formales, ámbitos de estudio todos ellos que cuentan con escasos precedentes en el análisis del español.

Finalmente, junto a la profundidad histórica y el tamaño, un requisito fundamental para la compilación de un corpus que sirva como base para la investigación en sociolingüística histórica estriba en la obtención de la mayor cantidad de información extratextual posible. Frente a las limitaciones que en este sentido ofrecen los grandes corpus de referencia, y que impiden controlar una serie de variables extralingüísticas esenciales para el análisis, en nuestro caso se contemplan los siguientes factores en la codificación de los textos:

-*Extensión geográfica*. Dadas las dificultades que entrañaría la compilación de un corpus panhispánico mínimamente representativo de las principales variedades dialectales del español con los recursos disponibles, el presente se limita al español europeo. Aun así, en la codificación se toman en consideración algunas divisiones dialectales que han hecho fortuna en los últimos tiempos, y que parten de una hipótesis según la cual las innovaciones lingüísticas se transmiten a través del espacio geográfico (Rodríguez Molina, 2010; Fernández Ordóñez, 2011). Así, en un reciente estudio acerca de la evolución del queísmo en los últimos cinco siglos (Blas Arroyo y Velando, 2022), hemos atendido a dos ejes geográficos cada vez más frecuentes en los trabajos de dialectología histórica, y que distinguen entre variedades norteñas y centro-meridionales, por un lado, y variedades orientales y occidentales, por otro. Lógicamente, para ello es necesario conocer el lugar donde se redactaron los documentos, ya sea de primera mano, a través de textos autógrafos, ya por medio de mediadores diversos, como escribanos, pendolistas, etc. que trasladaban al papel el dictado de los verdaderos autores. Dados los altos niveles de analfabetismo durante los primeros siglos, esta práctica fue habitual, sobre todo en la correspondencia epistolar mantenida en los estratos sociales más humildes. Así sucedió, por ejemplo, con tantos emigrantes que desde América intentaban comunicarse con sus allegados a este lado del océano (Fernández Alcaide, 2009; Stangl, 2013). Pese a ello, algunos estudios han comprobado que la influencia que pudieron ejercer tales intermediarios fue menor, y que no perturba la representatividad de los datos, al menos en los niveles gramatical y discursivo. Así, se ha visto, por ejemplo, en relación con el inglés antiguo (Bergs, 2005; Pahta y Jucker, 2011), pero también en algunos trabajos sobre el español (Blas Arroyo, 2016; Calderón, 2019).

Por comunidades históricas, algunas regiones resultan más favorecidas que otras, aunque estas representaciones pueden cambiar con el devenir de la historia. Así, en los primeros siglos encontramos un número elevado de extremeños, lo que no debe extrañar si consideramos la relevancia de estos en la colonización de América. Sin embargo, sus cifras a partir del siglo XVIII disminuyen considerablemente. De hecho, hasta esa centuria, la representación más nutrida corresponde, con diferencia, a territorios del antiguo reino de Castilla, como Castilla la Vieja, Castilla-La Mancha y Andalucía, a los que se suman numerosos testimonios de otros territorios norteños, como el País Vasco, Navarra o Asturias. Por el contrario, los documentos procedentes de los territorios de la antigua Corona de Aragón son menos cuantiosos en los Siglos de Oro, como significativamente menor fue el papel que tuvieron estos en las primeras etapas de la empresa colonizadora. Sin embargo, a partir del periodo dieciochesco, la procedencia dialectal de los textos tiende a equilibrarse.

Lamentablemente, no siempre ha sido posible conocer el origen del escritor o del escriba que actúa como intermediario. Del total de textos disponibles, hemos podido localizar la procedencia geográfica de un 70 % aproximadamente de los casos, si bien la

 DOI: https://doi.org/10.6035/clr.7200

distribución de estos ha sido irregular, con cifras más elevadas conforme se avanza en el tiempo. Por ejemplo, en los documentos del siglo XX, esas proporciones se acercan al 100 %, y no lejos de estos números se sitúan también los textos decimonónicos. Sin embargo, las lagunas en este sentido son mucho mayores en los documentos del siglo XVI.

-*Cronología*. Junto a la distribución por siglos a la que hacíamos referencia más arriba, cada uno de los textos se ha codificado de acuerdo con el año de su redacción, y en los casos en que esta datación no era precisa, en la década en que tuvo lugar. Por el contrario, se han descartado aquellos documentos en los que tal información se desconocía. Afortunadamente, las dificultades que la datación experimenta en el periodo medieval en un corpus como el CORDE (Rodríguez Molina y Octavio de Toledo y Huerta, 2017), se han podido salvar en la mayoría de las ocasiones. En el análisis cuantitativo, la consideración de este factor como una variable continua representa un instrumento muy útil para evaluar las tendencias evolutivas en los periodos tomados como referencia en cada caso.

-*Estatus social*. Frente al carácter unidimensional de los textos a los que se ha enfrentado tradicionalmente la lingüística histórica, salidos casi siempre de la pluma de los estratos más elevados de la sociedad, las tradiciones discursivas cercanas al polo de la inmediatez comunicativa permiten el acceso a diferentes sociolectos, un hecho especialmente relevante cuando se comprueba que muchos fenómenos de variación y cambio lingüístico tuvieron en su origen a grupos sociales diversos. Lógicamente, la sociolingüística histórica no puede contar con el nivel de precisión con que trabaja la sociolingüística sincrónica contemporánea, entre otras razones porque el conocimiento de la historia es a menudo fragmentario. Aun así, los avances en historiografía social de las últimas décadas avalan algunas divisiones sociales que, adaptadas a las particularidades de cada etapa histórica, y sin incurrir, por tanto, en anacronismos contraproducentes (Bergs, 2012), permitirían distinguir al menos varios grupos sociolectales. Inicialmente, para cada uno de los periodos estudiados se propone una división tripartita, en la que se distinguen tres estratos: superior, intermedio y bajo. Con todo, en los periodos en los que la estratificación social resulta más abrupta, como sucede en los siglos XVI y XVII, esa distribución puede simplificarse en función de las tendencias de variación observadas, distinguiendo a este respecto únicamente entre las élites sociales, por un lado, y el resto de la sociedad, por otro.

-*Sexo*. Pese al interés que esta variable despierta siempre, dada su comprobada responsabilidad en numerosos desenlaces de variación y cambio lingüístico, la escasez de muestras escritas a cargo de mujeres, especialmente en periodos remotos, supone un considerable problema de representatividad en las investigaciones de sociolingüística histórica. Lamentablemente, nuestro estudio no es una excepción, dado que los textos escritos por mujeres con que contamos apenas superan el 13 % sobre el total. Además, presentan algunos desequilibrios muestrales, con una mayor representación en las etapas extremas –español clásico y contemporáneo– que en el periodo dieciochesco. Se trata, en definitiva, de un problema importante, que a día de hoy parece de difícil solución –en buena medida porque la incorporación de la mujer a la escritura fue más tardía que la de los hombres–, aunque cabe confiar en que pueda paliarse en un plazo no demasiado lejano con el hallazgo y la edición de nuevos materiales. Con los mimbres actuales, sin embargo, la posibilidad de encontrar resultados estadísticamente significativos a partir de la diferenciación generolectal resulta problemática. Aun así, puede ser interesante evaluar el sentido de las potenciales diferencias frecuenciales, especialmente si estas se repiten de manera recurrente en variables y periodos distintos.

-*Tenor*. Aunque la concepción general de las tradiciones discursivas incluidas en el corpus las acerca al polo de la inmediatez, una revisión más detenida de los textos muestra diferentes grados de formalidad en su redacción. Para la configuración de este predictor estilístico, todos los textos se codifican de acuerdo con dos parámetros diferentes. En primer lugar, y con carácter general, la temática básica abordada en ellos. En segundo término, aunque en este caso reservado a la correspondencia epistolar, el tipo de relación que se establece entre el emisor del texto y sus destinatarios. La combinación entre ambos parámetros ofrece tres tipos de documentos, situados en otras tantas posiciones en un eje imaginario de formalidad. En un extremo, se sitúan las cartas entre familiares o miembros de unidades similares –esposos, padres e hijos, amantes, amigos íntimos, etc.–, en las que se abordan temas íntimos y solidarios, y en las que, por tanto, resulta previsible una mayor presencia de variantes vernáculas o cambios desde abajo. En el extremo opuesto, se ubica, por el contrario, la correspondencia entre individuos situados en puntos diferentes del eje del poder (cartas dirigidas de un superior a un inferior, o viceversa), y en la que se ventilan asuntos instrumentales, lo que hipotéticamente dará lugar a una mayor profusión de variantes estándares y cambios desde arriba. Por último, a caballo entre ambos extremos ideamos un punto intermedio en el que caben cartas de naturaleza diferente a las anteriores,[7] así como los textos autobiográficos, en los que es frecuente la aparición del plano más personal de la comunicación, pero en los que, a diferencia de la correspondencia epistolar, no hay un destinatario definido y, por tanto, falta el componente esencial de la dialogicidad.

-*Contexto migratorio*. Dada la relevancia que para la compilación del corpus posee la extensísima correspondencia mantenida entre individuos situados a uno y otro lado del océano, los textos se codifican también de acuerdo con el lugar de su redacción. Especialmente interesante como marco para la discusión acerca de los fenómenos de reestructuración, simplificación o koineización que se han aventurado para las situaciones de contacto intenso entre variedades dialectales diferentes (Penny, 2000; Tuten, 2003; Blas Arroyo, 2021 *a*), los textos del corpus se clasifican en dos grupos. Por un lado, se encuentran aquellos que se escribieron en América, previsiblemente en situaciones de contacto cotidiano con individuos de otras procedencias. Por otro, los documentos escritos desde España, en condiciones dialectales más convencionales, y expuestas previsiblemente a una menor influencia interdialectal.

La localización de estos metadatos, indispensables para la investigación sociolingüística, se obtiene a través de diferentes vías. En la mayoría de los casos, son los propios editores quienes proporcionan tal información. Con todo, cuando ello no es así, la lectura atenta de los textos ofrece detalles biográficos y contextuales de gran valor para desentrañar este tipo de datos.

4. RAZONES PARA UN CORPUS NO ANOTADO

Un debate creciente en la lingüística es el valor que aportan los corpus anotados, en los que se incluyen informaciones relevantes, tanto de carácter contextual, como, sobre

[7] Así ocurre, por ejemplo, con las cartas entre iguales en las que se tratan asuntos no íntimos ni personales –como las cartas de negocios entre socios y clientes, etc.–, pero también con la correspondencia entre algunos miembros de la familia extendida (parientes lejanos, etc.), en la que el tono y la temática abordados son de naturaleza mucho menos solidaria que en el primer grupo de misivas, ya que en ella priman objetivos básicamente instrumentales (pedir ayuda económica, solicitar un favor, etc.).

 DOI: https://doi.org/10.6035/clr.7200

todo, estructural (lematizaciones de palabras, etiquetados morfológicos y sintácticos, etc.). Inicialmente, no todos los usuarios de la lingüística de corpus son partidarios de incorporar tales informaciones, y de hecho los hay que recelan de esta clase de bases de datos, que, en su opinión, no solo contaminan los textos, sino que, al tiempo, dificultan innecesariamente el análisis (Sinclair, 2004). Hoy son, sin embargo, cada vez más los partidarios de que los corpus –sincrónicos y diacrónicos– cuenten con diferentes niveles de anotación, no solo porque la incorporación de estos datos facilita la búsqueda de los fenómenos que interesan al analista, sino también porque los avances informáticos permiten manejar los distintos niveles sin perturbar en exceso tales búsquedas (Vaamonde, 2018; Calderón, 2019). Claro que esto último requiere de un nivel de formación tecnológica que no está al alcance de cualquiera, lo que explica que, por lo general, los escasos corpus diacrónicos que a día de hoy presentan algún grado de anotación reserven esta parte a empresas externas, sin duda mucho más duchas en la resolución de aspectos técnicos, pero, como contrapartida, con una formación lingüística limitada, lo que pueden entorpecer la tarea asignada.[8] Si a ello unimos las limitaciones económicas a las que se enfrentan tantos equipos de investigación (como el nuestro), la situación aboca –por el momento– a la compilación de un corpus no anotado.

Ahora bien, como recuerda oportunamente Schulte (2009), existen razones que justifican seguir utilizando este tipo de corpus. Sin duda, una de ellas es la mencionada economía de recursos, pero hay también otros motivos de peso. El principal es, a su juicio: «the lack of availability of sufficiently large annotated corpora» (p. 167). Y ello es lo que ocurre, precisamente, en nuestro caso. Como hemos visto más arriba, existen ya corpus que recogen tradiciones discursivas cercanas al polo de la inmediatez comunicativa, pero estos no satisfacen las necesidades de la sociolingüística histórica, ya sea por la insuficiente información contextual que brindan –además de otros problemas en la interfaz de búsqueda y el acceso a los textos (de Benito, 2019)–, ya sea por la relativa escasez de materiales lingüísticos que proporcionan, como ocurre con corpus técnicamente superiores, pero cuyas dimensiones resultan todavía incompletas, al menos para nuestros intereses. En tales casos, advierte Schulte (2009: 170): «the only viable alternative may be to use a non-annotated corpus». El modo de hacerlo posible en nuestro caso se explica en los siguientes apartados.

5. CRITERIOS PARA LA SELECCIÓN DE LAS EDICIONES

Por las razones esgrimadas, y sin renunciar en el futuro a ciertos niveles de anotación de los materiales, optamos en este proyecto por la compilación de un corpus "crudo", del que obtendremos la información requerida mediante el empleo de las herramientas informáticas necesarias (sobre estas, ver § 6, a continuación). Además, se trata de un corpus *secundario*, en el sentido de que, en su confección, se utilizan materiales textuales ajenos, de manera que los compiladores somos tan solo responsables del proceso de selección y preparación electrónica (Sánchez-Prieto, 2012: 13).

Este carácter no es en absoluto novedoso, y de hecho preside la arquitectura de algunos de los corpus más difundidos, como sucede con CORDE, CdE y CORDIAM para el español, o el corpus CICA para la historia del catalán, entre otros. En todo caso, lo que

[8] Quizá ello explica que las tareas de etiquetado sintáctico se hallen todavía en un estado embrionario en la mayoría de los casos, a diferencia de otras informaciones de carácter morfológico, inicialmente más fáciles de resolver.

puede diferenciar unos de otros son los criterios elegidos para la selección de las obras. Así, CORDIAM emplea únicamente ediciones realizadas por historiadores de la lengua, mientras que los criterios del CdE son bastante más laxos (de Benito, 2019).

Como norma general, en nuestro caso tan solo hemos seleccionado aquellos documentos en los que existe una declaración explícita por parte de los editores acerca de los criterios que se han empleado para la transcripción, un aspecto fundamental para la confección del corpus (Taavitsainen y Fitzmaurice, 2007: 21–22). A partir de aquí, el rango de posibilidades varía en función del grado de detalle que interese al investigador. En los manuscritos, especialmente de los primeros siglos, abunda una notable variedad ortográfica, con un interés filológico indiscutible. Por poner un ejemplo: en las búsquedas de la variación entre las formas adverbiales *ansí* y *así* en el español clásico, en los textos encontramos las siguientes variantes gráficas: *asi, assi, así, assj, asy, ansi, ansí, ansj, ansy* (Blas Arroyo, 2021 *b*). Sin duda, esta variación es especialmente relevante en los estudios de carácter fonético o gráfico, pero no tanto en otros niveles del análisis, por no hablar de que constituye uno de los problemas principales para la anotación automática de textos históricos. Así las cosas, considerando que nuestro objetivo principal se centra en el estudio de variables que van más allá del nivel fónico, en la selección de los textos hemos primado las ediciones críticas frente a las paleográficas en aquellos casos en que disponíamos de ambas. En las primeras, los editores adaptan la acentuación y la puntuación a las normas contemporáneas, con el objeto de facilitar la lectura, al tiempo que simplifican diversas convenciones del texto original que pueden dificultar la búsqueda posterior de los datos lingüísticos, como la segmentación de palabras, el desarrollo de las abreviaturas, etc. Asimismo, en estas ediciones se reduce la variabilidad gráfica, siempre que esta no tenga relevancia fonética. Así, grafías como 'i', 'j', 'u' y 'v' se simplifican a menudo, aunque, lógicamente, no otras que pueden encerrar fenómenos específicos, como el seseo, el yeísmo, etc. Aun así, algunos trabajos presentan transcripciones semidiplomáticas, en las que la edición crítica aparece complementada por símbolos paratextuales, como inicios y finales de línea, de página, etc. A pesar del interés que este tipo de datos posee desde un punto de vista filológico, considerando nuestros objetivos, y con el fin de no entorpecer la localización del mayor número de formas posibles de las variables lingüísticas, hemos procedido a eliminar manualmente esos símbolos.

Como es lógico, en la selección de los materiales han tenido preferencia las ediciones a cargo de lingüistas, historiadores de la lengua y filólogos. Ahora bien, no todas las ediciones que salen del ámbito de la filología son inservibles a los efectos del análisis lingüístico. De hecho, no son pocos los historiadores sociales que en los últimos tiempos han sacado a la luz trabajos realmente valiosos, en los que se llevan a cabo transcripciones que, aun con las salvedades reseñadas más arriba, permiten la investigación en disciplinas como la dialectología o la sociolingüística históricas. La colaboración entre filólogos e historiadores interesa a ambas ramas del saber (Stangl, 2013), y de ella se pueden obtener beneficios mutuos.

6. HERRAMIENTAS PARA EL ACCESO A LOS DATOS

Una vez seleccionados los textos, se procede a su digitalización, y mediante el reconocimiento de caracteres –extraordinariamente mejorado en los últimos tiempos– se transforman en archivos de texto, los más adecuados a día de hoy para su manipulación posterior a través de programas de búsqueda automática. Con todo, antes de llegar al

 DOI: https://doi.org/10.6035/clr.7200

formato .txt, en algunos casos se realiza una escala previa en archivos .rtf, de los que, mediante las herramientas del procesador de textos, es posible eliminar toda la información paratextual prescindible, como las marcas ya reseñadas (inicios y finales de línea, de página, etc.), así como las notas al pie de página de las ediciones críticas, los metadatos al comienzo de las cartas, etc.

Para la localización de las variables lingüísticas hemos empleado hasta la fecha *WordSmith Tools* (Scott, 1996-2012), un programa cuya interfaz y manejo intuitivo facilita la búsqueda de información.[9] Además de proporcionar datos frecuenciales relevantes para la investigación lingüística, como la frecuencia absoluta y normalizada de las palabras en el corpus, o su tokenización –imprescindible para la correcta identificación de las unidades léxicas–, el principal interés de la aplicación estriba en su versatilidad para localizar todas las formas posibles de una misma variable lingüística. El procedimiento para hacerlo es diverso. El más sencillo corresponde a aquellos fenómenos en los que la variación tiene un importante componente léxico, como sucede con las ocurrencias que nos han servido como base para el estudio de algunos dobletes adverbiales en diversos periodos de la historia del español, como *ansí/así*; *agora/ahora*; *allí/allá*. En tales casos, las búsquedas tienen un éxito muy elevado desde el primer ensayo. Esta es, por ejemplo, la secuencia de caracteres que, introducida en la casilla "Search Word" de WS, permite localizar en menos de un minuto todas las ocurrencias de la variación *ansí/así* en los más de dos mil archivos que componen el corpus del siglo XVI:

ans^/asi/así/asj/assi/assí/assj/asy/assy

Tras la ordenación de los resultados obtenidos (N= 5.210), la tabla de concordancias muestra que apenas 22 no corresponden a la variable objeto de análisis.

Las cosas son algo más complicadas para algunas variables sintácticas como la mencionada variación entre las perífrasis *deber* y *deber de + infinitivo*, en las que interviene la flexión verbal, por un lado, y la posibilidad de que las cadenas de caracteres coincidan con otras palabras. Con todo, tras el oportuno entrenamiento, y la ayuda adicional de un recurso avanzado de búsqueda ("Exclude if search is or context contain"), el programa permite localizar sin excesivos problemas todas las ocurrencias de ambas expresiones. Estas son las cadenas de búsqueda para todas las combinaciones posibles en las que intervienen *deber* y *deber de + infinitivo*, respectivamente:[10]

dev* *r/deb* *r/dev* *ll*/deb* *ll*/dev* *rl*/deb* *rl*/dev* *r^e/deb* *r^e

dev* de *r/deb* de *r/dev* de *ll*/deb* de *ll*/dev* de *rl*/deb* de *rl*/dev* de *r^e/deb* de *r^e

Y esta, la lista de excepciones que permite evitar el cruce con otras palabras en las búsquedas:

[9] En un futuro próximo, deseamos explorar las posibilidades de otras herramientas, como *Sketch Engine*, diseñada para trabajar con grandes corpus y que, al mismo tiempo, permite etiquetar automáticamente los materiales escritos mediante CQL, lo que ayudaría a realizar búsquedas cada vez más sofisticadas.

[10] Por cierto, los resultados de la búsqueda ascienden a 754 y 164 ocurrencias, respectivamente, muy lejos, pues, de las proporciones que veíamos anteriormente en un corpus como *Post Scriptum* (N= 14/10).

deba^o/debat*/de^oçi*/de^oci*/devo*/debo*/debu*

Otras variables sintácticas requieren de búsquedas más complejas, en las que corresponde decidir qué elementos se exploran y cómo se realizan tales exploraciones. Por ejemplo, ante la ausencia de anotación sintáctica, para nuestro estudio sobre la evolución del queísmo en la historia (Blas Arroyo y Velando, 2022) tuvimos que adoptar algunas decisiones en este sentido. Dado que la cabeza de las construcciones queístas (y de sus contrapartidas preposicionales) está en la lengua en un número muy amplio e indeterminado, optamos por limitar el análisis a las estructuras más recurrentes en los diferentes periodos en que dividimos nuestra investigación (para ello la función "Wordlist" fue determinante). Con todo, el problema no terminaba ahí, pues el enlace (*que*) entre el núcleo de la construcción y la subordinada podía aparecer en posiciones diversas: más frecuentemente unido al núcleo, pero en no pocas ocasiones también a distancia, un hecho que podría tener relevancia para explicar la variación y que, por tanto, interesaba investigar. En estas circunstancias, por ejemplo, la localización en el corpus de la alternancia *seguro que/seguro de que* obligó a combinar la cadena "s^gur*" en la casilla de búsqueda básica, con dos recursos avanzados: a) el de exclusión, al que nos referíamos más arriba (para evitar, por ejemplo, la localización de palabras como "seguridad" o "seguramente"), y b) una nueva herramienta contextual "Context word(s) & context search horizons". En la casilla correspondiente a esta última, se introducía la combinación *"q*"*[11] y se solicitaba su ubicación en un número máximo de palabras a la derecha del núcleo, una cifra que, tras las comprobaciones oportunas, advertimos que nunca iba más allá de la posición R6. Ello nos permitió obtener ejemplos tanto de la variante preposicional (1), como de las alternativas queístas (2) y (3), y, además, con variaciones en la posición del enlace *que* con respecto al adjetivo, como puede advertirse al comparar los dos últimos fragmentos:

(1) ansí no se perderá real y está todo ello en la bolsa ***seguro de que*** oy en un año no deverá casi nada (*Die Korrespondenz spanischer Emigranten aus Amerika*, 1574)
(2) y estad ***seguro q*** os tengo gran voluntad… (*Carta de Luisa de Cárdenas para Pompeo Amoroso*, 1588)
(3) Y sed ***seguro***, e yo os lo prometo, ***que*** mientra que en esa tierra estuvierdes… (*Carta de Diego de Ordás a Francisco Verdugo*, 1530)

Otra de las funcionalidades que presenta WS es el acceso a un contexto amplio de las búsquedas, que además se puede incrementar a demanda, permitiendo llegar incluso al texto completo. Esto último facilita la codificación lingüística de las variables, especialmente cuando intervienen factores discursivos cuya consideración podría verse en peligro de otro modo. Uno de esos factores es, por ejemplo, el *priming*, entendido como la tendencia a reutilizar material lingüístico que el hablante acaba de emplear en el discurso. En el estudio sobre el queísmo, puede interesar, por ejemplo, investigar las tendencias asimilatorias (o disimilatorias) que el empleo de una forma tan frecuente como *que* puede ejercer cuando aparece en el contexto previo a la variable. En el fragmento (4), a continuación, se puede ver cómo la presencia del *que* anterior es más cercana que en (5). Pues bien, interpretada como una variable continua en el análisis de regresión, la

[11] Al objeto de recoger todas las variantes gráficas del enlace presentes en el corpus, especialmente en los siglos XVI y XVII: *que, qve, qe, q…*

 DOI: https://doi.org/10.6035/clr.7200

distancia en palabras de ese *que* podría darnos una respuesta acerca de ese potencial condicionante:

> (4) Y que él y sus servidores y aficionados no desean otra cosa ***que*** quedar ***seguros que*** vuestra santidad no haya de inquietar ni molestar a su majestad en sus estados y reinos (*Carta del gran duque de Alba*, 1556)
> (5) […] sea servido presentar a las personas ***que*** le nonbraremos porque sin duda podra estar su magestad ***seguro que*** procuraremos descargar su real conçiençia… (*Carta de Gaspar de Ávalos al licenciado Fuentes*, 1533)

No se puede ocultar que el trabajo con este tipo de herramientas requiere de una labor de filtrado manual inevitable, que en el caso de una variable como la alternancia entre variantes queístas y preposicionales se ve agravada por la selección de expresiones sintácticas que, aunque formalmente idénticas, son estructuralmente distintas. Por continuar con la variable ejemplificada en los últimos párrafos, la tabla de concordancias de *seguro (de) que* localiza ocurrencias de otras estructuras sintácticas que nada tienen que ver con el queísmo, como construcciones comparativas (*más segura que*), consecutivas (*tan seguro que*), adjetivas (*al puerto seguro que*), colocaciones (*ten por seguro que*), etc. Ahora bien, incluso en estos casos, la posibilidad de ordenar las búsquedas de acuerdo con criterios contextuales diversos, y el hecho de que WS resalte en color azul las unidades buscadas facilita considerablemente la tarea del filtrado. Así, la simple ordenación de las concordancias a partir de la palabra que precede al adjetivo y su resalte en rojo permite valorar con facilidad si los ejemplos precedidos de adverbios como "más" o "tan" no son en realidad sino manifestaciones de otro tipo de oraciones que nada tienen que ver con el queísmo.[12]

Sea como sea, este examen exhaustivo de las concordancias, por enojoso que sea, no solo está a años luz del lento proceder con que se desenvolvía el investigador no hace tanto tiempo, sino que además resulta imprescindible para aquilatar la fiabilidad de los datos. En definitiva, ante la duda, siempre será preferible revisar y volver a revisar los resultados, que dar estos por válidos sin excesivas comprobaciones. Por lo demás, y como sentencia adecuadamente Schulte (2009: 180) en relación con un método similar, aplicado a otro corpus no anotado: «The amount of inaccuracy generated by the method of analysis proposed in this paper is, in fact, in most cases insignificant in comparison with the degree of natural, content-dependent frequency variation present in any corpus».

7. CONCLUSIONES

En las páginas anteriores hemos presentado los fundamentos de un corpus diacrónico compilado por el grupo de investigación *Sociolingüística*, de la Universitat Jaume I, que a lo largo de la última década ha servido como base para la realización de numerosos estudios de sociolingüística histórica del español. Compuesto íntegramente

[12] Somos conscientes de que unas búsquedas basadas exclusivamente en criterios formales presentan limitaciones para la localización de otras variables sintácticas, y que estas podrían mejorar mediante recursos como la lematización o la anotación morfológica y sintáctica. Así se ha destacado, por ejemplo, en relación con el estudio de las formas auxiliares irregulares en las perífrasis verbales (Garachana y Artigas, 2012), el voseo, tantas veces camuflado en las desinencias verbales (Díaz Bravo, 2018) o la concordancia de objeto (García Salido y Vázquez Rozas, 2012). Aunque no imposible, en estos casos la multiplicación de las búsquedas necesarias dificulta sobremanera la localización de los datos lingüísticos que interesan.

por tradiciones discursivas cercanas al polo de la inmediatez comunicativa, como mejor forma de acercarnos al habla cotidiana de tiempos pretéritos, el proyecto se configura como un corpus específico –esto es, destinado a satisfacer unos objetivos concretos– y secundario, ya que se nutre de ediciones preparadas por otros autores, aunque cuidadosamente seleccionadas.

Tras analizar algunas lagunas que a día de hoy presentan todavía los corpus disponibles para el estudio sociolingüístico del pasado –insuficiente contextualización extralingüística en unos casos o escasez de los materiales disponibles en otros–, en el artículo se repasan las bases que han servido para la compilación de un corpus que ha seguido creciendo con el paso del tiempo mediante la incorporación progresiva de nuevos textos. Estas bases se fundamentan en tres principios básicos: una profundidad histórica suficiente, que permita seguir la evolución de los procesos de variación y cambio lingüístico en diferentes momentos de la historia del español; una representación lo más exhaustiva posible, que devuelva un número suficientemente amplio de las variables lingüísticas estudiadas en cada periodo; y una contextualización extralingüística adecuada, acorde también con los objetivos de la sociolingüística histórica. A partir de estos mimbres, el corpus supera hoy los catorce millones de registros, a razón de más de dos millones por centuria, escritos entre finales del siglo XV y la primera mitad del siglo XX. De estos, una gran mayoría corresponde a documentos extraídos de la correspondencia epistolar, y el resto, aunque en proporciones inferiores, a textos autobiográficos, como diarios personales, memorias de servicio, libros de cuentas, crónicas de soldados, etc.

Limitado por razones metodológicas al español europeo, el corpus da voz a cerca de siete mil individuos de diversa extracción geográfica y social. Con representación de todas las regiones españolas, aunque con distribuciones diferentes en distintos momentos de la historia, y fuertemente condicionado por el fenómeno de la emigración a América, en el corpus encontramos textos escritos o dictados por individuos de estratos sociales diferentes, desde las élites sociales y culturales en un extremo, a representantes de profesiones manuales y gentes del común, pasando por diversos grados intermedios. Al mismo tiempo, las temáticas y relaciones diversas entre los participantes en el proceso comunicativo permiten vislumbrar diferencias en el orden diafásico, que habilitan este factor para el análisis sociolingüístico.

En su estado actual, se presenta como un corpus no anotado, dadas las dificultades para obrar de otro modo con ediciones realizadas por autores ajenos al grupo de investigación y que, en prácticamente todos los casos, están protegidas por derechos de autor. Para la incorporación de estas al corpus se han preferido las ediciones críticas, en las que el editor interviene con algunas adaptaciones y simplificaciones (acentuación, puntuación y, en algunos casos, ciertas grafías sin valor fonético). Aun así, no se han descartado ediciones más complejas, en las que se emplean diversos símbolos paratextuales, aunque todos estos se han eliminado en el proceso de digitalización de los documentos previo a su conversión en archivos de texto. Son, finalmente, estos últimos los que sirven como base para la realización de las búsquedas, que llevamos a cabo mediante el auxilio de uno de los programas de concordancias más versátiles y utilizados en la actualidad, *WordSmith Tools*. En el artículo, se ejemplifican las principales herramientas para tales búsquedas, que, aun necesitadas de un filtrado manual por parte del investigador, arrojan resultados en los que la pérdida de información es mínima.

Como ha advertido Kabatek (2016: 10), la teoría del cambio lingüístico ha ido identificando en los últimos tiempos un número creciente de factores lingüísticos

 DOI: https://doi.org/10.6035/clr.7200

(sintácticos, semánticos, fónicos y pragmáticos) y extralingüísticos (grupos sociales, individuos, tradiciones discursivas), lo que ha supuesto un importante cambio de paradigma en el quehacer de la lingüística histórica. En sus palabras, que reproducimos a continuación:

> Una lingüística histórica con una base de datos fiables más amplia es precisamente la que produce los análisis más complejos y completos de las evoluciones y permite que nos acerquemos más a la reconstrucción adecuada del cambio. Por otro lado, resulta evidente que no todos los factores tienen el mismo peso en cada cuestión empírica concreta y que la tarea del lingüista no consiste únicamente en la recolección de datos y la enumeración de factores, sino en su ponderación e interpretación. Nos hallamos, pues, en una fase de la lingüística histórica en la que hay más complejidad, más datos y más factores de lo que solía haber, pero también nuevas posibilidades de ordenar los datos y de presentarlos.

La sociolingüística histórica persigue, justamente, tales objetivos, y de ahí la necesidad de contar con corpus que, aun con las limitaciones señaladas, ofrezcan como contrapartida datos suficientemente amplios y representativos para desentrañar la jerarquía de esos factores.

REFERENCIAS BIBLIOGRÁFICAS

Arias Álvarez, Beatriz y Hernández Mendoza, Juan Antonio (2013). Importancia de la incorporación de los parámetros diastráticos y diafásicos en la elaboración del corpus electrónico del español colonial mexicano. *Scriptum digital*, 2, 5–20.

Bergs, Alexander (2005). *Social Networks and Historical Sociolinguistics. Studies in Morphosyntactic Variation in the Paston Letters.* De Gruyter.

Bergs, Alexander (2012). The Uniformitarian Principle and the Risk of Anachronisms in Language and Social History. En M. Hernández-Campoy y J. Camilo Conde-Silvestre (Eds.), *The Handbook of Historical Sociolinguistics* (pp. 80–98). Blackwell.

Bertolotti, Virginia y Company Company, Concepción (2014). El corpus diacrónico y diatópico del español de América (CORDIAM). Propuesta de tipología textual. *Cuadernos del ALFAL*, 6, 130–148.

Blas Arroyo, José Luis (2016). The rise and fall of a change from bellow in Early Modern Spanish: The periphrasis «deber de + infinitive» in texts of linguistic immediacy. *Journal of Historical Linguistics*, 6(1), 1–31. DOI: 10.1075/jhl.6.1.01bla.

Blas Arroyo, José Luis (2021a). El contacto interdialectal a debate: análisis comparativo de seis fenómenos de variación en textos de inmediatez comunicativa escritos en España y América entre los siglos XVI y XVIII. *Revista Internacional de Lingüística Iberoamericana*, 38(2), 199–236.

Blas Arroyo, José Luis (2021b). Apogeo y declive de ansí en los Siglos de Oro: nuevos datos desde la sociolingüística histórica. *Boletín de Filología*, 56(1), 263–299.

Blas Arroyo, José Luis (2022). Patterns of individual variation and change in Golden Age Spanish. Analysis of three linguistic variables in a private correspondence corpus. *Folia Linguistica Historica*, 43(4), 1–40. https://doi.org/10.1515/folia-2022-2024.

Blas Arroyo, José Luis y Velando, Mónica (2022). *El queísmo en la historia: variación y cambio lingüístico en el régimen preposicional del español (siglos XVI-XXI).* Walter De Gruyter. https://doi.org/10.1515/9783110766851-00.

Calderón Campos, Miguel (2015). *El español del reino de Granada en sus documentos (1492-1833). Oralidad y escritura*. Peter Lang.

Calderón Campos, Miguel (2019). Los corpus del español clásico y moderno: Entre la filología y la lingüística computacional. *Revista de lingüística teórica y aplicada*, 57(2), 41–64.

Calderón Campos, Miguel y Vaamonde Dos Santos, Gael (2020). Oralia diacrónica del español: un nuevo corpus de la edad moderna. *Scriptum digital*, 9, 167–189.

Cano Aguilar, Rafael (1996). Lenguaje 'espontáneo' y retórica epistolar en cartas de emigrantes españoles a Indias. En T. Kotschi, W. Oesterreicher y K. Zimmermann (Eds.), *El español hablado y la cultura oral en España e Hispanoamérica* (pp. 375–404). Iberoamericana.

Caravedo, Rocío (1999). *Lingüística del corpus. Cuestiones teórico-metodológicas aplicadas al español*. Universidad de Salamanca.

Clavería Nadal, Gloria (2012). Corpus diacrónicos: nuevas perspectivas para el estudio de la historia de la lengua. En E. Montero Cartelle y C. Manzano Rovira (Coords.), *Actas del VIII Congreso Internacional de Historia de la Lengua Española*, 1, 405–420.

Culpeper, Jonathan y Kytö, Merja (2010). *Early Modern English dialogues. Spoken interaction as writing (Studies in English Language).* Cambridge University Press.

Davies, Mark. (CdE). *Corpus del español* [en línea]. Disponible en: https://www.corpusdelespanol.org/

De Benito Moreno, Carlota (2019). Los corpus del español desde la perspectiva del usuario lingüista. *Scriptum digital*, 8, 1–21.

Di Tullio, Ángela y Resnik, Gabriela (2019). Diario de un soldado: una fuente para la reconstrucción de la oralidad rioplatense del siglo XIX. *VI Congreso de la Red Internacional CHARTA*.

Díaz Bravo, Rocío (2018). Las Humanidades Digitales y los corpus diacrónicos en línea del español: problemas y sugerencias. En E. Romero Frías y L. Bocanegra Barbecho (Eds.), *Ciencias Sociales y Humanidades Digitales Aplicadas* (pp. 562–686). Universidad de Granada.

Enrique Arias, Andrés (2009). Lingüística de corpus y diacronía de las lenguas iberorromances. En A. Enrique Arias (Ed.), *Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus* (pp. 11–21). Iberoamericana/Vervuert.

Enrique Arias, Andrés (2012). Dos problemas en el uso de corpus diacrónicos del español: perspectiva y comparabilidad. *Scriptum digital: revista de corpus diacrònics i edició digital en llengües iberoromàniques*, 1, 85–106.

Fernández Alcaide, Marta (2009). *Cartas de particulares en Indias del siglo XVI*. Iberoamericana.

Fernández Ordóñez, Inés (2011). *La lengua de Castilla y la formación del español*. Real Academia Española.

Fontanella de Weinberg, María Beatriz (1992). *Documentos para la historia lingüística de Hispanoamérica*. Boletín de la Real Academia Española.

 DOI: https://doi.org/10.6035/clr.7200

Frühbeck, Nicolás M. (2022). La autobiografía confesional de los siglos XVI y XVII: una propuesta. *Janus: estudios sobre el Siglo de Oro*, 11, 449–473.

García Salido, José María y Vázquez Rozas, Victoria (2012). Los corpus diacrónicos como instrumento para el estudio del origen y distribución de la concordancia de objeto en español. *Scriptum Digital,* 1, 67–84.

Garachana Camarero, Mar y Artigas, Esther (2012). Corpus digitales y palabras gramaticales. *Scriptum digital*, 1, 37–65.

Isasi, Santiago; Pierazzo, Elena y Spence, Paul (2020). *Edición digital de documentos antiguos: marcación XML-TEI basada en los criterios CHARTA*. Universidad de Sevilla.

Kabatek, Johannes (2013). ¿Es posible una lingüística histórica basada en un corpus representativo? *Iberoromania: Revista dedicada a las lenguas y literaturas iberorrománicas de Europa y América,* 77, 8–28.

Kabatek, Johannes (2016). Un nuevo capítulo en la lingüística histórica iberorrománica: el trabajo crítico con los corpus. En J. Kabatek y C. de Benito Moreno (Eds.), *Lingüística de Corpus y Lingüística Histórica Iberorrománica* (pp. 1–18). De Gruyter.

Koch, Peter y Oesterreicher, Wulf (2007 [1990]). *Lengua hablada en la Romania: español, francés, italiano*. Gredos.

Labov, William. 1994. *Principles of Linguistic Change. Internal Factors*. Blackwell.

Morala, José Ramón (2012). Léxico e inventarios de bienes en los Siglos de Oro. En G. Clavería Nadal, M. Freixas, M. Prat Sabater y J. Torruella (Eds.), *Historia del léxico: perspectivas de investigación* (pp. 199-218). Iberoamericana/Vervuert.

Navarro Gala, Rosario (2020). *La voz armada del soldado español Alonso de Medina (1549)*. Vervuert.

Nevalainen, Terttu y Raumolin-Brunberg, Helen. 2017 [2003]. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. Pearson Education.

Octavio de Toledo y Huerta, Álvaro S. y Pons Rodríguez, Lola (2017). *Textos para la historia del español. Tomo X: Queja política y escritura epistolar durante la Guerra de la Independencia: documentación de la Junta Suprema Central en el AHN. Selección, edición y estudio lingüístico.* Universidad Alcalá de Henares.

Oesterreicher, Wulf (2004). Textos entre inmediatez y distancia comunicativas. El problema de lo hablado en lo escrito en el Siglo de Oro. En R. Cano (Ed.), *Historia de la lengua española* (pp. 729–769). Ariel.

Otte, Enrique (1988). *Cartas privadas de emigrantes a Indias, 1510-1616*. Escuela de Estudios Hispano-Americanos.

Pahta, Päivi y Jucker, Andreas (2011). *Communicating Early English Manuscripts. (Studies in English Language).* Cambridge University Press.

Penny, Ralph (2000). *Gramática histórica del español*. Ariel.

Raumolin-Brunberg, Helena (2009). Lifespan changes in the language of three early modern gentlemen. *Pragmatics and beyond. New series*, 183, 165–196.

Real Academia Española: Banco de datos (CORDE) [en línea]. *Corpus diacrónico del español.* Disponible en: https://corpus.rae.es/cordenet.html

Rivadeneira-Valenzuela, Marcela y Contreras-Gutiérrez, Alejandra (2021). En el nombre de Dios Todopoderoso»: los tratamientos nominales en la «Relación autobiográfica de Úrsula Suárez (1666-1749). *Rilce. Revista de Filología Hispánica*, 37(1), 162–88.

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

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

Rodríguez Puente, Paula (2018). En busca de lo hablado en lo escrito en los corpus diacrónicos del español: una comparativa con los corpus anglosajones. *E-Scripta Romanica*, 5, 89–127.

Sánchez-Prieto Borja, Pedro (2012). Desarrollo y explotación del "Corpus de Documentos Españoles Anteriores a 1700" (CODEA). *Scriptum digital*, 1, 5–35.

Sánchez-Prieto Borja, Pedro y Vázquez Balonga, Delfina (2019). *La beneficencia madrileña. Lengua y discurso en los documentos de los siglos XVI al XIX*. Ediciones Complutense.

Schneider, Edgar W. (2013). Investigating Historical Variation and Change in Written Documents: New Perspectives. En J. K. Chambers y N. Schilling (Eds.), *The Handbook of Language Variation and Change* (pp. 57–81). Wiley-Blackwell.

Schulte, Kim (2009). Using non-annotated diachronic corpora: benefits, methods and limitations. En A. Enrique Arias (Ed.), *Diacronía de las lenguas iberorromances: nuevas aportaciones desde la lingüística de corpus* (pp. 169–182). Iberoamericana/Vervuert.

Scott, Mike (1996-2012). *WordSmith Tools (Version 5.0)* [Software].

Sinclair, John (2004). *Trust the text: Language, corpus and discourse*. Routledge.

Stangl, Werner (2013). Un cuarto de siglo con Cartas privadas de emigrantes a Indias. Prácticas y perspectivas de ediciones de cartas transatlánticas en el Imperio español. *Anuario de estudios americanos*, 70(2), 703–736.

Taavitsainen, Irma y Fitzmaurice, Susan (2007). Historical pragmatics: What it is and how to do it. En S. Fitzmaurice e I. Taavitsainen (Eds.), *Methods in Historical Pragmatics* (pp. 11–36) De Gruyter.

Tuten, Donald N. (2003). *Koineization in Medieval Spanish*. De Gruyter.

Vaamonde, Gael (2015). Limitaciones en el uso de corpus diacrónicos del español. Nuevas aportaciones desde el proyecto de investigación. *Post Scriptum*. *E-Aesla,* 1, 1–10.

Vaamonde, Gael (2018). Escritura epistolar, edición digital y anotación de corpus. *Cuadernos del Instituto Historia de la Lengua*, 11, 139–164.

van der Wal, Marijke y Rutten, Gijsbert (2013). Change, contact and conventions in the history of Dutch. *Taal en Tongval*, 65(1), 97–123.