Tres corpus para el español del siglo XVIII *CHARTA*, *CORDIAM* y un corpus jesuítico

Three corpora for 18th century Spanish *CHARTA*, *CORDIAM* and a Jesuit corpus

Marina Albers
Paris-Lodron-Universität Salzburg (Salzburg, Austria), Ludwig-Maximilians-Universität München (München, Alemania)
marina.albers@plus.ac.at

Recibido el 17/10/2023, aceptado el 29/1/2024, publicado el 18/10/2024

Cómo citar este artículo
Albers, Marina 2024. Tres corpus para el español del siglo XVIII. *CHARTA*, *CORDIAM* y un corpus jesuítico. *Studia linguistica romanica* 2024.12, 157-186. https://doi.org/10.25364/19.2024.12.8.

Resumen
Esta contribución se dedica a la comparación de tres corpus electrónicos en vistas a su uso para el estudio del español del siglo XVIII, siglo relativamente desatendido en la diacronía. Por un lado, presentaremos un corpus inédito de documentos jesuíticos de la región rioplatense-paraguaya, que se consulta a través de una base de datos relacional, y, por otro lado, dos subcorpus de comparación extraídos del *CHARTA* y del *CORDIAM*, uno español y otro novohispano. Con el objetivo de deducir de esta comparación las ventajas para la lingüística *con* corpus para usuarios filólogos, realizaremos dos búsquedas ejemplares, así como una amplia contraposición de una serie de parámetros de los tres corpus, incluyendo aspectos informáticos, filológicos y aspectos relacionados con la comodidad y manejabilidad de las superficies de uso desde el punto de vista del usuario.

Abstract
This contribution compares three electronic corpora with regard to their use in studying eighteenth-century Spanish, a period that has been relatively neglected in diachronic research. On the one hand, we present an unpublished corpus of Jesuit documents from the Rioplatense-Paraguayan region, which can be consulted through a relational database, and on the other hand, two comparative sub-corpora extracted from *CHARTA* and *CORDIAM*, one Spanish and the other Novo-Hispanic. To illustrate the advantages of these corpora for corpus linguistics, particularly for philologists, we perform two exemplary searches and conduct a comprehensive comparison of a series of parameters across the three corpora, including computational and philological aspects as well as the comfort and manageability of their user interfaces.

Índice

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

1 Introducción

[1] La lingüística *de* corpus y *con* corpus (Kabatek 2016: 3) ha ganado territorio en la lingüística histórica y en la historia de la lengua en los últimos años. Por un lado, la lingüística *de* corpus pone a disposición un número cada vez mayor de documentos históricos a través de corpus electrónicos disponibles en la web para todos los usuarios. Por otro lado, los investigadores se benefician de estas bases documentales para diversos fines lingüísticos en el campo de la lingüística *con* corpus, dentro de la cual pueden diferenciarse distintas maneras de trabajar con un corpus para ejemplificar y construir hipótesis y teorías, a saber, *corpus-based* o *corpus-driven* (Tognini-Bonelli 2001). La metodología *corpus-based* constituye el proceder clásico de la lingüística, exponiendo, evaluando y ejemplificando una teoría lingüística hecha anteriormente mediante un corpus, mientras que las investigaciones *corpus-driven* se dirigen a la creación de nuevas teorías e hipótesis a partir de la integridad de los datos de un corpus. Ambas direcciones dependen de la existencia de corpus electrónicos con una amplia base documental fielmente elaborada. Los corpus electrónicos históricos son, además de las ediciones de textos, una herramienta imprescindible para el estudio diacrónico de la lengua.

[2] El objetivo de esta contribución consiste en la comparación de tres corpus para el español del siglo XVIII, siglo menos atendido en la historia de la lengua, con el fin de describir, evaluar y contrastar las características y el uso concreto de los tres con fines filológicos. En § 2, esbozaremos brevemente la situación del español dieciochesco en cuanto a su anclaje y representación en los estudios lingüísticos históricos. § 3 estará dedicado a la presentación de un nuevo corpus inédito[1] que consta de documentos jesuíticos redactados durante el siglo XVIII en la Provincia jesuítica del Paraguay, que comprendía en la época colonial tanto Paraguay como parte de los actuales territorios argentinos y uruguayos, además de la parte fronteriza de Brasil. Por lo tanto, usaremos en esta contribución los términos de Río de la Plata y de Paraguay de manera sinónima, refiriéndonos a la antigua Provincia jesuítica del Paraguay. Nos centraremos tanto en la elaboración de una base de datos relacional a partir de los documentos de los jesuitas como en las características de la misma. En § 4, seleccionaremos de dos conocidos corpus electrónicos, el *CHARTA* y el *CORDIAM*, dos subcorpus de comparación que representan el siglo XVIII en dos regiones distintas, esto es, la española y la novohispana, antes de proceder en § 5 a la comparación de dos búsquedas ejemplares en los tres corpus como herramientas para estudiar el español del siglo XVIII. § 6 servirá como balance de las características de elaboración y de uso de los tres corpus estudiados, así como de las ventajas y desventajas desde un punto de vista lingüístico. Cerraremos esta contribución con algunas reflexiones finales en § 7, tanto con

[1] El corpus jesuítico constituye la base para mi tesis de doctorado, cuya publicación contará asimismo con la publicación de los documentos.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

respecto al estudio del español del siglo XVIII en los tres corpus analizados como a una serie de desiderata en el ámbito de la lingüística *de* y *con* corpus.

2 El español dieciochesco

[3] El estudio del español del siglo XVIII ha recibido en el pasado, como bien es sabido, una menor atención en la lingüística histórica, así como en la historia de la lengua, constituyendo, en las palabras de Company Company (2012: 255), «un gran vacío de la investigación diacrónica, posiblemente, […] *el* gran vacío de la diacronía». En oposición con el precedente Siglo de Oro – rico en investigaciones lingüísticas sobre los múltiples cambios e innovaciones de la época áurea –, los trabajos dedicados al siglo XVIII se restringían en el pasado a la fijación y estabilización de la lengua por parte de la Real Academia Española, suponiendo que los cambios lingüísticos se terminaran a mediados del siglo XVII. Sin embargo, los estudios sobre el español dieciochesco de los últimos años han demostrado que sí se produjeron revoluciones lingüísticas aún después del siglo XVII, de ahí que se estableciera el término de *primer español modero*, que abarca aproximadamente el espacio temporal entre 1675 y 1825, para referirse al período de transición entre el español clásico y el moderno (García Godoy 2012b: 9-10; Octavio de Toledo y Huerta 2016: 57-60).

[4] Si bien surgió en los últimos años una serie de investigaciones sobre el español del siglo XVIII[2], en mayor medida sobre las variedades americanas, se sigue manteniendo la deficiencia diacrónica, sobre todo en lo que se refiere a estudios que se enfocan en los cambios morfosintácticos, infrarrepresentados hasta el momento (Guzmán Riverón & Sáez Rivera 2016b: 13). Además, ya que los cambios producidos durante el primer español moderno llevaron a la configuración de determinados rasgos dialectales del español tanto peninsular como americano, así como a la consolidación de la actual división diatópica, hay que hacer hincapié en la relevancia de los estudios dialectológicos (García Godoy 2012b: 11).

[5] No obstante, con el fin de poder estudiar las diferentes variedades del español dieciochesco, es menester basarse en colecciones documentales del siglo XVIII como corpus de investigación. Pese a que el siglo XVIII estaba incluido en una serie de volúmenes que abarcaban, por ejemplo, toda la época colonial – como es el caso de las recopilaciones documentales dirigidas por Fontanella de Weinberg (1993) y Rojas Mayer (2000) sobre las variedades hispanoamericanas de los siglos XVI a XVIII – surgieron también colecciones propias dieciochescas centradas en determinadas regiones y variedades, como lo constituyen las publicaciones de Bertolotti, Coll & Polakof (2010) sobre la Banda Oriental, de Gómez Seibane & Ramírez Luengo (2007) sobre Bilbao o bien de García Aguiar (2014) sobre Málaga.

[2] Los volúmenes de García Godoy (2012a), de Sáez Rivera & Guzmán Riverón (2012) y de Guzmán Riverón & Sáez Rivera (2016a) reúnen un número considerable de trabajos sobre el siglo XVIII.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

[6] En cuanto al español dieciochesco de la región rioplatense, en su amplia definición – incluyendo el actual Uruguay, Paraguay, así como parte de los vastos territorios argentinos, que están en el centro de nuestra contribución – contamos hasta el momento con cuatro estudios exhaustivos, a saber, acerca de Buenos Aires (Fontanella de Weinberg 1984) y la Banda Oriental (Elizaincín, Malcuori & Bertolotti 1997), que se dedican exclusivamente al siglo XVIII, además de los trabajos sobre Tucumán (Rojas Mayer 1985) y Santa Fe (Donni de Mirande 2004), que incluyen el lapso temporal del siglo XVI al XIX. Cabe destacar que carecemos de un estudio acerca del actual territorio paraguayo en el siglo XVIII, de ahí que nuestro corpus jesuítico, que presentaremos en § 3, pueda llenar este vacío.

[7] Si bien consultar las ediciones que contienen las compilaciones documentales resulta, quizás, el acceso más directo al texto y, por consiguiente, al estudio diacrónico de la lengua, la manera más aprovechada en el siglo XIX es recurrir a los corpus históricos disponibles en la web para todos los usuarios. La ventaja de la mayoría de estos corpus electrónicos, además de disponer de una base documental de mayor amplitud y de una elaboración informática compleja, consiste en la posibilidad de periodización y de selección de las áreas que sean del interés del investigador, es decir que los corpus electrónicos ofrecen asimismo la oportunidad de estudiar el español del siglo XVIII con una base documental más amplia y diversa. Por lo tanto, procederemos en los próximos apartados a la comparación de un nuevo corpus jesuítico inédito, que presentaremos a continuación, con dos corpus electrónicos, el *CORDIAM* y el *CHARTA*, en lo que se refiere a los documentos dieciochescos incluidos en ellos.

3 El corpus rioplatense de los jesuitas

[8] La lingüística *de* corpus, que tiene como objetivo la creación de corpus electrónicos, se divide en dos fases, esto es, una primera fase meramente filológica y una segunda, de índole informática y computacional (Kabatek 2016: 3; Calderón Campos 2019: 42).

[9] Antes de adentrarnos en la presentación del nuevo corpus jesuítico, cabe aclarar que no pretendemos ni la generación de un corpus electrónico de acceso abierto ni una propuesta de cómo construir un corpus histórico, como lo plantea por ejemplo Torruella Casañas (2017), sino que nuestro objetivo consiste únicamente en describir el modo de proceder que hemos elegido en el marco de la tesis doctoral en lingüística para tratar los datos documentales[3]. No obstante, ya que la base de datos relacional elaborada, que contiene los documentos jesuíticos, puede considerarse, a pesar de su tamaño, acceso y uso de herramientas informáticas re-

[3] Agradezco al revisor anónimo sus comentarios acerca de otros mecanismos computacionales de mayor eficacia, así como sobre las herramientas informáticas que pueden beneficiar la creación de un corpus electrónico, que tendré en cuenta a la hora de dedicarme a la elaboración de un corpus de acceso abierto. En esta contribución, sin embargo, en la que nos proponemos exclusivamente describir lo que hasta el momento se hizo, los objetivos son de naturaleza meramente filológica.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

ducidos, un corpus electrónico, presentaremos a continuación ambas fases de elaboración del corpus dieciochesco de los jesuitas.

3.1 Fase filológica

[10] El primer paso de la lingüística *de* corpus consiste en la selección documental, fase fundamental para la futura elaboración del corpus electrónico. Los documentos de nuestro corpus jesuítico proceden del Archivo General de la Nación de Buenos Aires (AGN), donde, en un proyecto de colaboración entre la Pontificia Universidad Católica Argentina y el Centro Universitario de Digitalización de Documentos e Investigación, fueron escaneados más de 5600 documentos redactados en el ámbito de las actividades de la Compañía de Jesús en la Provincia jesuítica del Paraguay.

[11] De la totalidad de los documentos escaneados, hemos seleccionado aquellos redactados por los miembros de la orden que hayan nacido en los territorios de la Provincia del Paraguay (cf. Storni 1980), de manera que los 29 escribientes de nuestro corpus puedan considerarse criollos rioplatenses. El resultante corpus consta de un total de cien documentos, que contienen 87 cartas de índole oficial, destinadas en su gran mayoría a otros jesuitas, y 13 recibos de pago, actas, contratos y otros textos que se caracterizan por su impronta jurídico-administrativa. Los documentos de naturaleza mayoritariamente epistolar fueron redactados entre 1728 y 1765 en parte de los actuales territorios de Paraguay, Argentina, Uruguay y Brasil.

[12] Después de concluir la selección de los documentos escaneados, el siguiente paso consiste en la transcripción. En cuanto al tipo de edición, hemos optado por la edición paleográfica de los textos del corpus, respetando fielmente los usos gráficos originales, y no por una presentación crítica ni normalizada, con el fin de poder estudiar tanto las particularidades gráficas, como reflejo de la evolución fónica de la época, como las variantes ortográficas (cf. Calderón Campos 2019: 43-44; Torruella Casañas 2017: 171; Sánchez Lancis 2022: 35). Además, se tuvieron en cuenta los parámetros de transcripción elaborados para el *CORDIAM*, de ahí que hayamos desatado las abreviaturas mediante cursivas y los superíndices mediante redondas. Al final de la fase filológica, los cien documentos jesuíticos del siglo XVIII conforman un corpus paleográfico con un total de 31074 palabras.

3.2 Fase informática

[13] Los siguientes pasos computacionales, que iremos describiendo para el corpus jesuítico sin pretensión de ejemplaridad, versan alrededor del tratamiento de los datos y del etiquetado del corpus (Kabatek 2016: 3). De este modo, las transcripciones convertidas en formato TXT fueron subidas a una base de datos relacional del tipo SQL, que constituirá la plataforma de acceso y de consulta[4], si bien sabemos que el uso de otros mecanismos hubiera llevado a un resultado más

[4] Agradezco a Christian Riepl del equipo informático de la Ludwig-Maximilians-Universität München por su generosa ayuda en la elaboración de la base de datos.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

provechoso. Desde esta base en la web, hemos realizado los procesos de separación y numeración de las formas ortográficas en tokens (tokeneización), que llegan en nuestro corpus a un número de 34159, y de lematización, esto es, la asignación de un lema a cada uno de los tokens, en nuestro caso de forma automática. Sin embargo, ya que la lematización no fue realizada sobre una edición normalizada, sino sobre la transcripción paleográfica y, por lo tanto, no estándar, fue imprescindible controlar la clasificación de los lemas manualmente, también con respecto a la desambiguación, es decir, la asignación de tokens como *fue* al lema *ir* o *ser*, respectivamente. Este paso nos permitirá luego encontrar todas las variantes gráficas y formas flexivas vinculadas con un lema buscado, por ejemplo, *embiado*, *envie*, *embio*, *enviare*, etc. para el lema *enviar* (cf. Calderón Campos 2019: 45-46). Adicionalmente, pueden utilizarse comodines, en este caso el símbolo *%*, para sustituir uno o bien varios caracteres, tanto a nivel de token como a nivel de lema. Por lo tanto, la búsqueda por el token *fue%*, por ejemplo, arroja diversos resultados, tales como *fueron*, *fuese*, *fuera*, pero también formas pertenecientes a otros lemas, como *fuerte*, *fuerza*, *fuego*, etc. Además, hemos elaborado en la base de datos una tabla propia con trigramas, esto es, la secuencia de tres tokens en una cadena, basada en el principio de n-gramas (cf. Davies & Parodi 2022: 19).

[14] La siguiente tarea trata del etiquetado morfosintáctico, que consiste en la asignación de una categoría gramatical, en nuestro caso una clase de palabra, a cada uno de los tokens. En el leguaje computacional suele emplearse el término de *part-of-speech tagging* (*POS tagging*) para referirse a esta forma de etiquetado, que, al final posibilita búsquedas de mayor grado de abstracción, tales como la búsqueda por combinaciones de determinadas clases de palabras, por ejemplo, *determinante* + *posesivo*, facilitada por los trigramas (cf. Calderón Campos 2019: 45-46). Pese a que el *POS tagging* automático, en nuestro caso el programa *TreeTagger* (cf. Moreno-Sandoval 2022: 405), trabaja sobre los lemas ya controlados, cabe desambiguar los casos dudosos y revisar la asignación de las clases de palabras de manera manual.

[15] Concluida la fase informática, la superficie de uso de la base de datos elaborada cuenta, en formato de tabla, con los textos transcritos en su edición paleográfica separados en tokens, el número del token, los lemas y las clases de palabras correspondientes, así como de información adicional sobre la fuente archivística, el folio, el párrafo y la línea del manuscrito de cada uno de los tokens, ordenados en columnas, como puede observarse a continuación:

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

id_token	token	quelle Quelle	folio Folio	absatz Absatznummer	zeile Zeilennummer	lemma	wortart
3477	Resebi	AG442	f.751r	2	2	recibir	V
3478	la	AG442	f.751r	2	2	el	ART
3479	de	AG442	f.751r	2	2	de	PREP
3480	Vuestra	AG442	f.751r	2	2	vuestro	POS
3481	Reverencia	AG442	f.751r	2	2	Reverencia	NP
3482	y	AG442	f.751r	2	2	y	CONJ
3483	me	AG442	f.751r	2	2	me	P PERS
3484	alegro	AG442	f.751r	2	2	alegrar	V
3485	se	AG442	f.751r	2	2	se	P PERS
3486	halle	AG442	f.751r	2	2	hallar	V
3487	libre	AG442	f.751r	2	2	libre	ADJ
3488	de	AG442	f.751r	2	2	de	PREP
3489	su	AG442	f.751r	2	2	su	POS
3490	catarro	AG442	f.751r	2	2	catarro	N

Figura 1: La superficie de uso de la base de datos

3.3 Características de la base de datos

[16] En la tabla 1 a continuación pueden apreciarse las características propias de la base de datos jesuítica, de acuerdo a los parámetros establecidos por Calderón Campos (2019):

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

Parámetro	Corpus jesuítico
Cronología	- 1728-1765
Extensión geográfica	- Provincia jesuítica del Paraguay (Paraguay, Argentina, Uruguay, Brasil)
Tipología textual	- cartas - documentos jurídico-administrativos (listas, contratos, etc.)
Tamaño del corpus	- 100 documentos - 34159 tokens - promedio de palabras por texto: 310,74
Ediciones	- edición paleográfica - permite expresiones regulares (comodín %)
Anotación lingüística	- lematización - etiquetado: *POS tagging*
Características particulares	- base de datos estructurada - separación de los datos metatextuales de los textuales

Tabla 1: Características del corpus jesuítico rioplatense

Como puede deducirse de la tabla, nuestra base de datos dieciochesca constituye un corpus tanto reducido en tiempo, espacio y tamaño como homogéneo con respecto a la tipología textual y al origen social y étnico de los escribientes. Sin embargo, la edición paleográfica, así como la fiel anotación lingüística posibilitan el estudio del español rioplatense en el siglo XVIII en diferentes niveles lingüísticos, a saber, morfosintáctico, léxico y gráfico-fónico. Por lo tanto, puede considerarse una herramienta de utilidad para el fin de la investigación, que consiste en ampliar la base documental rioplatense de la época y en estudiar el español dieciochesco.

[17] Cabe señalar, no obstante, que carecemos en la base de datos del contexto sintáctico de las ocurrencias, visto que los resultados contienen exclusivamente el token buscado, aislado de su contexto sintáctico, lo cual constituye la gran deficiencia de la base de datos. Para poder analizar el token en su contexto mayor, es imprescindible acceder a la transcripción del respectivo documento en otro formato, lo que, si bien resulta poco práctico, es una tarea realizable con un corpus de tamaño tan limitado.

4 Los subcorpus de comparación: *CHARTA* y *CORDIAM*

[18] Disponemos de una serie de corpus históricos de la lengua española disponibles en la web, tanto en lo que concierne al español peninsular como al español americano y sus respectivas variedades[5]. Para nuestro objetivo de investigación, hemos optado, por una parte, por el *CHARTA*, y por otra parte, por el *COR-*

[5] Las contribuciones de Bertolotti & Company Company (2022), Calderón Campos (2019) y Sánchez Lancis (2022), entre otras, comparan gran parte de los corpus existentes desde diferentes perspectivas.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

DIAM, donde seleccionaremos el período de 1700 a 1800 con tal de poder contrastar las particularidades lingüísticas de la época de manera suprarregional[6].

[19] No obstante, antes de adentrarnos en la descripción detallada de los subcorpus de comparación, presentaremos brevemente las características[7] de los corpus de acuerdo a los parámetros aplicados en la tabla 1 (cf. Sánchez Lancis 2022: 38; Calderón Campos 2019: 46-49, 56; Bertolotti & Company Company 2022: 50-56; Company Company & Bertolotti 2021: 582-585):

Parámetro	***CHARTA***	***CORDIAM***
Cronología	- 1200-1800 - siglo más representado: XIII	- 1494-1905 - siglo más representado: XIX
Extensión geográfica	- América, Asia y España - sobre todo España	- América (24 países)
Tipología textual	- documentos de diversa tipología - sobre todo formales (legislativa, notarial)	- *CORDIAM* documentos - *CORDIAM* prensa - *CORDIAM* literatura
Tamaño del corpus	- 2076 textos - siglo XVIII: 94 textos	- 20444 textos - 15391056 palabras - siglo XVIII: 4487 textos, 3135063 palabras
Ediciones	- facsímil - transcripción paleográfica - edición crítica - permite expresiones regulares (comodines *, ?)	- facsímil parcialmente disponible - transcripción paleográfica - permite expresiones regulares (comodines *, ?)
Anotación lingüística	- no lematizado - no etiquetado	- lematización parcial/en proceso - no etiquetado
Características particulares	- visualización en triple formato - análisis estadístico	- metadatos cronológicos, geográficos, textuales y sociales - datos cuantitativos - PDF de cada documento

Tabla 2: Características de los corpus *CHARTA* y *CORDIAM*

La tabla 2 visualiza las características de ambos corpus según los parámetros de Calderón Campos (2019) y pone en evidencia el mayor tamaño – tanto en lo que

[6] Véase con fines de comparación el trabajo de Octavio de Toledo y Huerta (2016), quien estudia el *CORDE* con respecto al análisis sintáctico en el primer español moderno.

[7] Los datos de ambos corpus corresponden a la fecha de consulta, el 28 de septiembre de 2023. Todos los posibles cambios posteriores a esta fecha no podrán tenerse en cuenta.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

se refiere al número de textos, palabras y tipos textuales como a la extensión geográfica – del *CORDIAM*. Con respecto a las ediciones de los textos, el *CHARTA* se destaca por su triple formato, esto es, la posibilidad de acceder al facsímil, a la transcripción paleográfica, así como a la edición crítica, formato que falta en el *CORDIAM*, pese a que un gran número de textos dispone ahí de la versión facsimilar. Los dos corpus electrónicos carecen de una anotación lingüística general, a saber, no son etiquetados morfosintácticamente y la lematización, que falta en el *CHARTA*, está en proceso en el *CORDIAM*. El *CHARTA* dispone de un análisis estadístico que incluye las frecuencias absolutas y relativas de cada token del respectivo documento, mientras que el *CORDIAM* ofrece datos cuantitativos del total de palabras sobre el que se lleva a cabo la búsqueda, lo cual permite calcular la frecuencia relativa. Además, el *CORDIAM* se destaca por poseer para cada texto 14 tipos de metadatos[8] de índole geográfica, social, histórica, textual etc., de los cuales siete pueden ser considerados en la búsqueda. Gracias a estos metadatos puede caracterizarse con mayor exactitud una variedad histórica concreta también desde el punto de vista sociolingüístico, ya que permiten conocer las características del escribiente del texto, así como el escenario comunicativo (Arias Álvarez & Hernández Mendoza 2016: 387).

[20] En cuanto al siglo XVIII, que se encuentra en el enfoque de nuestro estudio, el corpus *CHARTA* contiene un total de 94 documentos, de los cuales 78 son de origen español, diez de Venezuela, dos de Cuba y uno de Ecuador, Colombia, Puerto Rico y El Salvador, respectivamente, de ahí que España constituya con un 83 % de los textos el país más representado. Las tipologías textuales versan alrededor de documentos jurídico-administrativos tales como actas y declaraciones, cartas de compraventa, textos legislativos, etc. El corpus americano *CORDIAM*, por el contrario, dispone de 4487 textos dieciochescos, por lo que resulta imprescindible una restricción del total de documentos con el fin de poder realizar búsquedas puntuales. Si contrastamos los textos del *CORDIAM* con el corpus jesuítico, conviene buscar la mayor comparabilidad posible con respecto a los escribientes, que son criollos en nuestro corpus rioplatense. Por consiguiente, limitaremos el subcorpus del *CORDIAM* a documentos del siglo XVIII redactados por criollos, lo cual proporciona la siguiente distribución por países:

Argentina: 10	El Salvador: 21	Paraguay: 11
Bolivia: 14	Guatemala: 42	Perú: 10
Chile: 17	Honduras: 15	Uruguay: 12
Colombia: 1	México: 24	Venezuela: 3
Costa Rica: 1	Nicaragua: 74	Total: 255

Tabla 3: Número de documentos escritos por criollos, siglo XVIII, según países (*CORDIAM*)

[8] Estos son, para cada documento consultado: país, toponimo actual, topónimo histórico, adscripción histórico-administrativa, autoría, sexo, pertenencia étnica, número de palabras, síntesis, etc. (Company Company & Bertolotti 2021: 582).

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

En la tabla 3 queda manifiesto que, al tratar de extraer de los documentos dieciochescos escritos por criollos un subcorpus que sea de la mayor homogeneidad posible en cuanto a su extensión geográfica, destacan con un porcentaje de un 69 % los textos de origen novohispano, es decir, los que proceden de las regiones pertenecientes en la época colonial al virreinato de la Nueva España. Esta restricción nos permite, en consecuencia, crear un propio subcorpus de comparación que contiene un total de 177 documentos con 92577 palabras, distribuidos entre 74 de Nicaragua, 42 de Guatemala, 24 de México, 21 de El Salvador, 15 de Honduras y uno de Costa Rica.

[21] En resumen, hemos seleccionado de los dos corpus elegidos para la comparación suprarregional uno de origen español del *CHARTA*, con un total de 74 documentos, y uno novohispano del *CORDIAM*, que contiene 177 documentos redactados por criollos, ambos exclusivamente del siglo XVIII. Procederemos a continuación a la búsqueda de dos fenómenos lingüísticos que resultan de interés en el primer español moderno, uno gráfico y otro de índole morfosintáctica, en los tres corpus.

5 Comparación de búsquedas en los tres corpus

[22] En la base de datos relacional que contiene nuestros documentos jesuíticos, disponemos de dos opciones de búsqueda. Por un lado, existe la posibilidad de recurrir al motor de búsqueda propio de la base de datos, que permite introducir uno o más datos requeridos: lema, token, clase de palabra, fuente, etc., aparte de la posibilidad de usar el comodín *%*. Por otro lado, dado que se trata de una base de datos de tipo SQL, pueden realizarse búsquedas en el lenguaje SQL. En cambio, los subcorpus ofrecen motores de búsqueda propios del *CHARTA* y del *CORDIAM*, que permiten únicamente realizar búsquedas por formas ortográficas, dado que (aún) no son ni lematizados ni etiquetados.

[23] Las particularidades lingüísticas que hemos escogido como meros ejemplos para el análisis comparativo en los tres corpus son, por una parte, la vacilación entre <b> y <v> (y <u> con valor consonántico), que pese a la regulación por parte de la Real Academia Española en el proemio del *Diccionario de autoridades* en 1726, donde fueron prescritos los criterios etimológicos, pervivió en la grafía del siglo XVIII. Por otra parte, hemos optado por la combinación de diferentes determinantes en el margen izquierdo del sintagma nominal, fenómeno cuya existencia en el siglo XVIII se asocia con la tradición medieval, así como con la influencia de la tradición discursiva jurídico-administrativa. De este modo, trataremos de contrastar los tres corpus históricos en lo que respecta a su función como herramienta para el estudio de las características del español dieciochesco.

5.1 Vacilación entre <b> y <v>

[24] En primer lugar, nos centraremos en la búsqueda por la vacilación entre <b> y <v> en unos lemas modelo, elegidos arbitrariamente por la gran cantidad de palabras afectadas, para ilustrar las diferencias en los tres corpus de comparación.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

Hemos seleccionado *gobernar*, *gobernador*, *gobierno*, etc., que de acuerdo a la raíz etimológica latina *gubern-* suelen escribirse con <b> desde que su ortografía fue normalizada por la Real Academia Española. Sin embargo, pueden suponerse otras posibles grafías tales como *govierno*, *gouierno*, etc., que cabe tener en cuenta a la hora de la búsqueda.

[25] La base de datos que contiene los cien textos de los jesuitas criollos de la Provincia del Paraguay ofrece, como hemos expuesto arriba, dos maneras de búsqueda. Por una parte, se puede introducir en el motor, que trabaja sobre la edición paleográfica, pero lematizada, en la categoría *lema* la forma *gob%*, permitiendo de esta manera, con el comodín *%*, que aparezcan todas las formas ortográficas (tokens) que empiecen con *gob-*, *gov-*, *gou-*, etc. Por otra parte, puede aplicarse la siguiente búsqueda SQL:

(1)
```
SELECT *
FROM tokens
WHERE lemma LIKE 'gob%'
```

Ambas versiones de búsqueda llevan al mismo resultado que arroja un total de 18 ocurrencias correspondientes a los criterios de búsqueda, como puede observarse en la figura a continuación que ilustra la superficie con los resultados:

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

id_token	token	quelle Quelle	folio Folio	absatz Absatznummer	zeile Zeilennummer	lemma	wortart
381	Governador	AG104	f.167	2	3	Gobernador	NP
2458	governaba	AG311	f.569v	3	43	gobernar	V
3056	Governador	AG311	f.570v	3	78	Gobernador	NP
11736	govierna	IdL430	f.633r	2	8	gobernar	V
13360	gobierno	IR847	f.1182r	2	13	gobierno	N
16010	Governador	JIU102	f.165r	6	12	Gobernador	NP
16817	governarla	JIU99	f.161r	2	3	gobernar	V
17359	Governador	JJdP299	f.453v	4	18	Gobernador	NP
17797	Governador	JJdP417	f.630r	4	35	Gobernador	NP
19486	Governador	JNA282	f.430r	2	5	Gobernador	NP
20677	Governador	LdT285	f.433b.r	6	41	Gobernador	NP
20698	Governador	LdT285	f.433b.r	6	44	Gobernador	NP
23789	Governador	LdT778	f.1097r	8	11	Gobernador	NP
23900	Governador	LdT778	f.1097r	15	21	Gobernador	NP
23970	Governador	LdT778	f.1097r	18	28	Gobernador	NP
25021	govierna	MB374	f.734r	2	8	gobernar	V
25074	govierne	MB374	f.734r	2	12	gobernar	V
34107	Govierno	TdT341	f.653v	4	24	gobierno	N

Figura 2: Los resultados de búsqueda en la base de datos

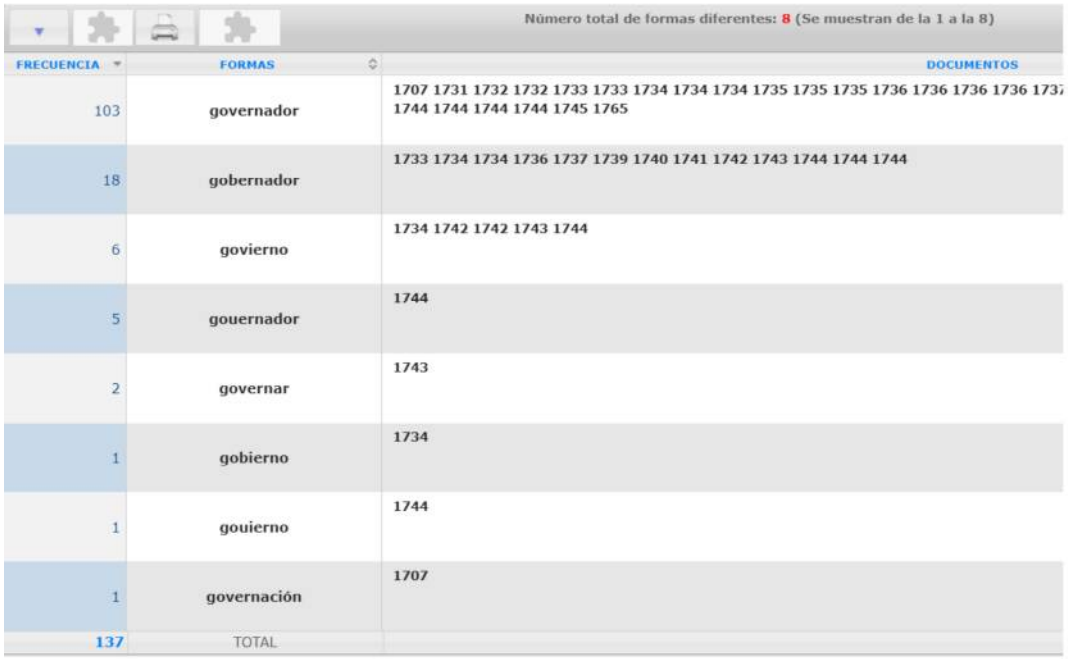
De la segunda columna, que muestra los tokens, puede deducirse que 17 de las 18 ocurrencias muestran una grafía con <v>, lo cual corresponde a un 94 %. Además, gracias a la lematización y el *POS tagging*, contamos con informaciones adicionales acerca de cada uno de los tokens en las columnas *lemma* ('lema') y *wortart* ('clase de palabra').

[26] En los dos subcorpus de comparación del *CHARTA* y del *CORDIAM*, que no son ni lematizados ni etiquetados morfosintácticamente, la tarea de identificar las formas ortográficas correspondientes pareciera ser de mayor dificultad, dado que cabe saber de antemano las posibles grafías que teóricamente puede llegar a haber. No obstante, los comodines permiten búsquedas bastante generales, de ahí que la forma *go*ern** resulte adecuada para incluir las potenciales formas ortográficas *govierno*, *gouernador*, *gobernaron*, etc[9].

[9] La búsqueda por lemas en el *CORDIAM*, que está en proceso, no toma todavía en consideración las variantes gráficas, por lo que la búsqueda por un lema como *gobern** proporciona exclusivamente las ocurrencias escritas con <b>, de acuerdo al lema.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

[27] La búsqueda por *go*ern** en los documentos de origen español en el *CHARTA*, así como en los textos novohispanos escritos por criollos, ambos del siglo XVIII, nos lleva a los resultados que veremos en los extractos de las superficies de uso de los corpus:



Número total de formas diferentes: 8 (Se muestran de la 1 a la 8)

FRECUENCIA	FORMAS	DOCUMENTOS
103	governador	1707 1731 1732 1732 1733 1733 1734 1734 1734 1735 1735 1735 1736 1736 1736 1736 1737 1744 1744 1744 1744 1745 1765
18	gobernador	1733 1734 1734 1736 1737 1739 1740 1741 1742 1743 1744 1744 1744
6	govierno	1734 1742 1742 1743 1744
5	gouernador	1744
2	governar	1743
1	gobierno	1734
1	gouierno	1744
1	governación	1707
137	TOTAL	

Figura 3: Los resultados de búsqueda en el *CHARTA*

go*ern* Buscar Búsqueda simple

Buscando: *go*ern** donde Siglo es *18* y País actual es *CR* o *GUA* o *HON* o *MEX* o *NIC* o *PAN* o *SAL* y Tipo textual es cualquiera de Documentos y Autor (datos étnicos) es *criollo*

Mostrados solo 20 casos. Encontrados 50 casos en 30 (de 177) textos que contienen 20 295 (de 92 577) palabras.

Resultado generado el sábado 30 de septiembre de 2023, 15:44:57 CEST

1 Primero Anterior 1 Siguiente Ultimo 3

1	18 NIC ADM	…a dado el Maestro de Campo Don, Miguel de Camargo	Gouernador	de esta prouinçia assistiendo perssonalmentte co…
2	18 NIC ADM	…ni peligro, a dichos padres, en los pueblos deste	gouierno	, procurando en ttodo Y por ttodo, la mayor onrra …
3	18 NIC JUR	…s avctos paran en poder de su merced el señor /	governador	de la provincia, quien tiene encargado (…) cru…
4	18 NIC JUR	…rced el ma(…) / de campo don Miguel de Camargo,	gouernador	de lo politico y militar / desta probincia por s…
5	18 NIC CAR	…General de la Caualleria Don Joseph Calbo de Lara	Gobernador	en ynterin de la Prouincia de Nicaragua; ymforma …
6	18 NIC CAR	…a ocasión a Vuestra Magestad de como estoy en el	gobierno	de esta Probincia, desde onze de Junio de el año…
7	18 MEX JUR	…reales, donde / le mando el senor ttheniente al	governador	y alcaldes que lo / llevasen a la cárzel, y el s…
8	18 MEX JUR	…/ me sean testtigos cómo, aviéndole mandado al	govierno	/ que llevase preso a dicho Cárdenas, no quiso o…
9	18 NIC ADM	…Y a uista del Comun asierto, y aPlausso Con que a	Gouernado	Tres Veses Su rrelixion de Predicadores En las Pr…
10	18 NIC CRO	…, lo á manifestado desde que tomó posession del	Govierno	de su Cathedral, que átropellando, El quebrantto…
11	18 NIC ADM	…ro cauildo diferentes / consultas de matherias de	gouierno	, assi / en lo espiritual como en lo temporal, d…
12	18 NIC ADM	…fue nombrado por maestro / de seremonias, y en el	gouierno	de este vuestro / obispo hizo el supra dicho dife…
13	18 NIC ADM	…nta a Vuestra magestad de auer quedado solo en el	Gouierno	d[e] ella por auer fallesido. El Dean, Arcediano …
14	18 NIC ADM	… Yglesia Cuia difucion Causo el quedar solo en el	gouierno	d[e] ella, y Obispado, por auer acontesido la mis…
15	18 MFX ADM	o de México, del Consejo de su magestad, virrey,	governador	/ y capitán general de esta Nueva España y pres…
16	18 MEX ADM	…blo de /20 Santa Maria, ocurrió a este superior	gobierno	expresando si / nuestramente (sic) el que mis par…
17	18 MEX ADM	…n posesión por despacho (…) / de este superior	gobierno	, notifique a dicha doña Juana exhiba / el despac…
18	18 MEX ADM	…nobar, remita ambas diligencias a este superior /	gobierno	citadas las partes para que vien de su derecho. Y…
19	18 MEX JUR	…otibo a que por omision aya havido quexa a ningun	governador	/ ni alcalde mayor. Y que tampoco a dado motibo a…
20	18 MEX JUR	…gavelas (sic) que los / an querido apensionar los	governadores,	alcaldes y mandones, / como en algunos y los mà…

Documento DLNECM116:

Ejemplo | Metadatos | PDF | Facsimil

1. **Nombre:** 116
2. **Siglo:** 18
3. **Año:** 1716
4. **Autor (datos étnicos):** criollo
5. **Autor (hombre o mujer):** hombre
6. **Autógrafo:** no
7. **País actual:** MEX
8. **Topónimo actual:** Papantla
9. **Topónimo histórico:** Papantla
10. **Adscripción histórica:** Audiencia de México, Virreinato de la Nueva España
11. **Tipo textual:** Documentos jurídicos
12. **Archivo:** Archivo General de la Nación, México, Criminal 77, ff. 32r-33v. Comienza en la línea 4.
13. **Número de folios:** 2
14. **Número de palabras aproximado:** 870
15. **Créditos:** Chantal Melis, Agustín Rivero Franyutti, con la colaboración de Beatriz Arias Álvarez, Documentos lingüísticos de la Nueva España. Golfo de México, México: Universidad Nacional Autónoma de México, 2008.
16. **Facsimilar disponible:** si
17. **Síntesis:** Denuncia de un hombre criollo, Juan José Pérez, contra Alonso de Cárdenas, por desacato a un teniente.

Figura 4: Los resultados de búsqueda en el *CORDIAM*

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

Mientras que los resultados del *CHARTA* aparecen *ad hoc* agrupados según formas ortográficas idénticas, lo cual permite identificar rápidamente las que correspondan al criterio de búsqueda, como puede verse en la figura 3, el *CORDIAM* carece de este orden, por lo que cabe filtrar manualmente las formas ortográficas que son de interés (figura 4). Ambos corpus ofrecen información cuantitativa respecto al número total de ocurrencias, 50 en el caso del *CORDIAM* y en el *CHARTA* 137, si bien se trata solamente de 8 formas distintas. La ventaja del *CORDIAM* consiste en el hecho de que el primer paso de la búsqueda arroja instantáneamente tanto el contexto sintáctico de cada una de las ocurrencias como, al marcar una de las ocurrencias, los metadatos existentes en la ventana a la derecha. En cambio, para obtener estas mismas informaciones en el *CHARTA*, que ilustraremos en la figura 5, cabe elegir una forma y acceder a los resultados:

DOCUMENTO	FECHA	PROVINCIA	CONTEXTO PRECEDENTE	FORMA	CONTEXTO SIGUIENTE
CODEMA-0310 {h1va:21}	1734	Málaga	a <...> ciudad el excelentisimo señor don Alexandro de la Mota, teniente general de los exérsitos de su magestad, al	**govierno**	politico y militar de esta plaza. Y porque su excelencia quiere tomar la posesión este dia para lo que han
CODEMA-0310 {h2ra:12}	1734	Málaga	la carta orden del tenor siguiente: Siendo el ánimo del rey que vuestra excelencia pase a servir por aora el	**govierno**	militar y politico de Málaga y a encargarse también del mando de aquella costa, lo partisipó a vuestra excelencia de
CODEMA-0334 {h2ra:14}	1742	Málaga	de este año, en la villa de Madrid ante los señores del consejo de su magestad estando haciendo sala de	**govierno**	en la posada del eminentisimo señor cardenal de Molina, obispo de esta ciudad, y governador del consexo, el expresado señor
CODEMA-0335 {h2ra:14}	1742	Málaga	de este año, en la villa de Madrid ante los señores del consejo de su majestad, estando haciendo sala de	**govierno**	en la posada del excelentisimo señor cardenal de Molina, obispo de esta ciudad y gobernador del consexo, el expresado señor
CODEMA-0337 {h2ra:6}	1743	Málaga	sido sitado a todos los cavalleros rexidores que se hallan en esta ciudad para, en él, dar la posesión del	**govierno**	militar y politico interino al señor don Antonio Manzo Maldonado. Dieron fee los porteros haber hecho dicha zitasión. La ziudad,

Figura 5: Los contextos sintácticos en el *CHARTA*

En la figura 5 aparecen los contextos precedentes y siguientes, así como los metadatos respectivos a la ocurrencia escogida, que en comparación con el *CORDIAM* contienen poca información detallada, a saber, ni la tipología textual ni informaciones sobre el escribiente.

[28] Una vez identificadas las formas que corresponden a los criterios de búsqueda, tanto el *CHARTA* como el *CORDIAM* permiten acceder a las transcripciones y eventualmente a los facsímiles de los textos – en el *CHARTA*, además a la edición crítica –, que pueden ser de relevancia en el marco de un estudio de la grafía o braquigrafía. En el *CHARTA*, el práctico triple formato es reproducido en una misma pantalla, como ilustra la figura 6:

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

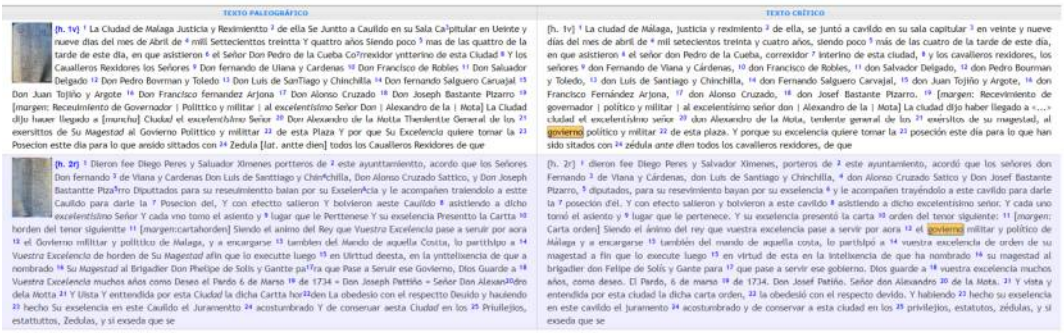


Figura 6: Representación del triple formato en el *CHARTA*

5.2 Determinantes en el sintagma nominal

[29] Después de haber contrastado la búsqueda por un fenómeno gráfico en los tres corpus de comparación, nos adentraremos en segundo lugar en la combinación de diferentes determinantes dentro del sintagma nominal a nivel morfosintáctico. Como modelo, buscaremos por la combinación de *demostrativo + posesivo* (+ *sustantivo*) con el fin de abarcar casos como *este mi*, *este tu*, *este su*, *esos mis*, etc., búsqueda que se complica por la cantidad de diferentes demostrativos y personas gramaticales.

[30] La base de datos de los documentos jesuíticos dispone, como ya hemos señalado arriba, de dos opciones de búsqueda, que conviene realizar esta vez sobre los trigramas, que se encuentran en otra tabla. De este modo, podemos introducir en el motor de búsqueda la combinación de las tres clases de palabras que nos interesa, esto es, *demostrativo (DEM) + posesivo (POS) + sustantivo (N)*. El lenguaje SQL permite realizar la misma búsqueda:

(2)
```
SELECT *
FROM triple_tokens_wortart
WHERE awortart LIKE 'DEM' AND bwortart LIKE 'POS' AND cwortart
LIKE 'N'
```

El resultado de ambas formas de búsqueda es el de ocho ocurrencias, de las cuales siete corresponden a los criterios, como puede apreciarse en el fragmento de la tabla a base de los trigramas, que cuenta con las columnas *token* y *wortart* ('clase de palabra') para cada una de las posiciones a, b y c buscadas:

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

aID_token	bID_token	cID_token	atoken	btoken	ctoken	awortart	bwortart	cwortart
3434	3435	3436	esso	nuestro	credito	DEM	POS	N
7381	7382	7383	esta	mi	carta	DEM	POS	N
13754	13755	13756	esta	su	gente	DEM	POS	N
14069	14070	14071	esta	su	gente	DEM	POS	N
17995	17996	17997	esa	su	redusion	DEM	POS	N
23725	23726	23727	Este	mi	ofizio	DEM	POS	N
24024	24025	24026	este	mi	ofizio	DEM	POS	N
24942	24943	24944	esta	nuestra	Provincia	DEM	POS	N

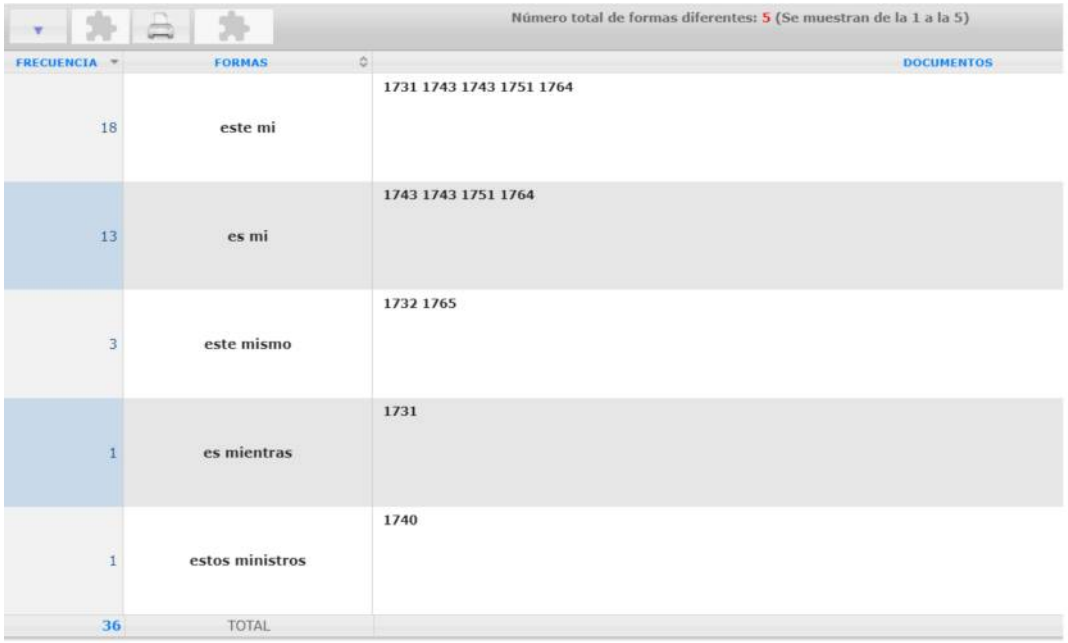
Figura 7: Los resultados de búsqueda en la base de datos

En la figura 7 pueden observarse las diferentes combinaciones que aparecen en el corpus de los documentos jesuíticos. Cabe volver a mencionar que, para ver el contexto sintáctico entero, hay que recurrir a la transcripción paleográfica en un documento aparte. Sin embargo, destacan las ventajas de la base de datos lematizada y etiquetada a través del *POS tagging*, que permiten realizar tales búsquedas abstractas por clases de palabras.

[31] En los dos subcorpus de comparación del *CHARTA* y del *CORDIAM*, que carecen de este etiquetado morfosintáctico, resulta necesario recurrir a los comodines para realizar búsquedas de mayor abstracción. Por lo tanto, es imprescindible llevar a cabo un gran número de búsquedas con el fin de obtener resultados para cada una de las combinaciones posibles[10], que hay que conocer de antemano: *es* mi**, *es* tu**, *es* su**, *es* nuestr**, etc. Las dos figuras a continuación ilustran los resultados que se obtienen de la búsqueda por *es* mi** en el *CHARTA* (figura 8) y el *CORDIAM* (figura 9):

[10] Además, cabe tener en cuenta las eventuales variaciones gráficas.

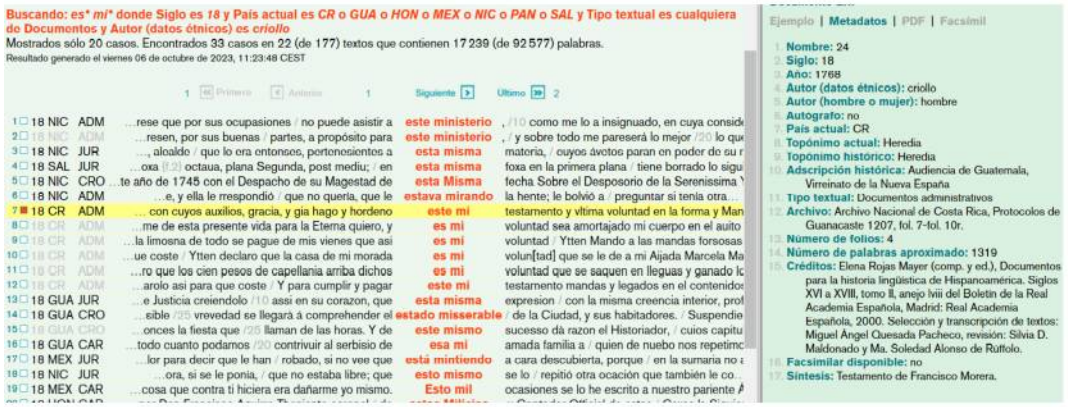
DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815



Número total de formas diferentes: 5 (Se muestran de la 1 a la 5)

FRECUENCIA	FORMAS	DOCUMENTOS
18	este mi	1731 1743 1743 1751 1764
13	es mi	1743 1743 1751 1764
3	este mismo	1732 1765
1	es mientras	1731
1	estos ministros	1740
36	TOTAL	

Figura 8: Los resultados de búsqueda en el *CHARTA*



Buscando: es* *mi** donde Siglo es *18* y País actual es *CR* o *GUA* o *HON* o *MEX* o *NIC* o *PAN* o *SAL* y Tipo textual es cualquiera de Documentos y Autor (datos étnicos) es *criollo*

Mostrados sólo 20 casos. Encontrados 33 casos en 22 (de 177) textos que contienen 17 239 (de 92 577) palabras.

Resultado generado el viernes 06 de octubre de 2023, 11:23:48 CEST

Ejemplo | Metadatos | PDF | Facsímil

1. Nombre: 24
2. Siglo: 18
3. Año: 1768
4. Autor (datos étnicos): criollo
5. Autor (hombre o mujer): hombre
6. Autógrafo: no
7. País actual: CR
8. Topónimo actual: Heredia
9. Topónimo histórico: Heredia
10. Adscripción histórica: Audiencia de Guatemala, Virreinato de la Nueva España
11. Tipo textual: Documentos administrativos
12. Archivo: Archivo Nacional de Costa Rica, Protocolos de Guanacaste 1207, fol. 7-fol. 10r.
13. Número de folios: 4
14. Número de palabras aproximado: 1319
15. Créditos: Elena Rojas Mayer (comp. y ed.), Documentos para la historia lingüística de Hispanoamérica. Siglos XVI a XVIII, tomo II, anejo lviii del Boletín de la Real Academia Española, Madrid: Real Academia Española, 2000. Selección y transcripción de textos: Miguel Ángel Quesada Pacheco, revisión: Silvia D. Maldonado y Ma. Soledad Alonso de Ruffolo.
16. Facsimilar disponible: no
17. Síntesis: Testamento de Francisco Morera.

Figura 9: Los resultados de búsqueda en el *CORDIAM*

[32] La figura 8 contiene los resultados de búsqueda del *CHARTA* y muestra las diferentes secuencias de formas que corresponden a los criterios de búsqueda, además del número de ocurrencias, dentro de las cuales resulta fácil filtrar aquellos que son relevantes. En este caso, se trata de 18 ocurrencias de *este mi* en cinco documentos distintos, como puede deducirse de la indicación de los años a la derecha. En el *CORDIAM* (figura 9), en cambio, cabe seleccionar dentro de las 33 ocurrencias que son proporcionadas, las combinaciones que correspondan a los criterios, ya que los resultados no están ordenados según formas, sino cronológi-

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

camente[11]. Si bien este primer paso de filtración requiere un esfuerzo – que dependiendo de la cantidad de ocurrencias puede llegar a ser grande –, el *CORDIAM* ofrece la posibilidad de acceder automáticamente a los metadatos a la derecha, así como al contexto sintáctico en el que está incrustada la ocurrencia, que incluso puede expandirse al seleccionar la opción *ejemplo* en la ventana lateral:

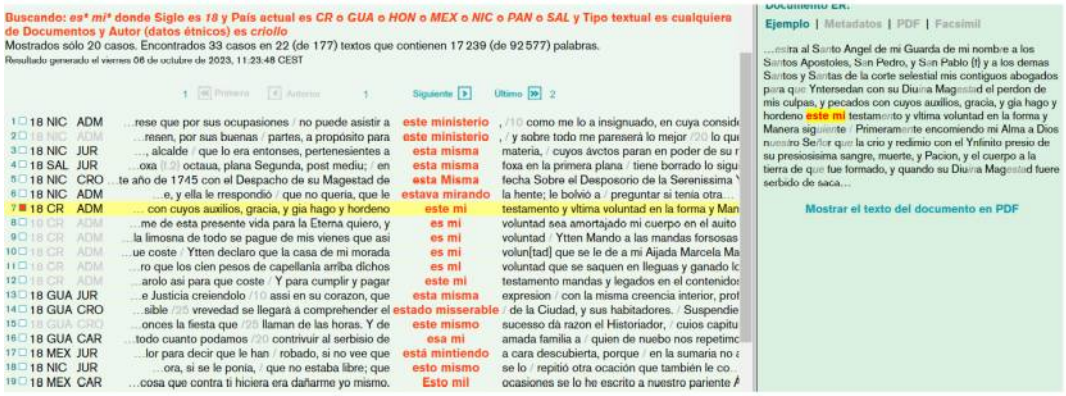


Figura 10: El contexto ampliado en el *CORDIAM*

De este modo, el formato del *CORDIAM* resulta ser de mayor comodidad en cuanto a búsquedas morfosintácticas. Como ya hemos señalado arriba, en el *CHARTA* es necesario elegir la forma buscada para poder ver el contexto sintáctico en un segundo paso.

6 Balance

[33] Después de haber contrastado la base de datos de textos rioplatenses de los jesuitas con los subcorpus dieciochescos del *CORDIAM* y del *CHARTA* a base de dos ejemplos de búsqueda, trataremos de resumir las características en cuanto al análisis gráfico, por un lado, y respecto a la morfosintaxis, por otro, con el objetivo de deducir las ventajas y desventajas de cada uno de los tres corpus para el investigador que persigue un fin filológico a través de la lingüística *con* corpus.

[34] La tabla 4 a continuación reúne una serie de parámetros que, de acuerdo a la comparación que acabamos de hacer, se consideran relevantes en el ámbito de la lingüística *con* corpus del español del siglo XVIII:

[11] Existe también la posibilidad de ordenar los resultados por año, país, tipología textual o alfabéticamente.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

Parámetro	Corpus jesuítico	*CHARTA*	*CORDIAM*
Agrupación por formas ortográficas	no	sí	no
Visualización del contexto sintáctico	no	sí (en un segundo paso)	sí (instantáneamente)
Información cuantitativa a) número de resultados b) frecuencia de las formas c) número de formas diferentes d) número de textos y palabras	 sí no no no	 sí sí sí no	 sí no no sí
Metadatos sociolingüísticos	6-7	2-3	5-6
Ediciones	doble formato	triple formato	doble formato
Procesamiento de las ocurrencias	descarga de las formas	no	descarga de las ocurrencias en su contexto sintáctico
Lematización	sí	no	en progreso
Etiquetado	sí	no	no

Tabla 4: Comparación de las características de los tres corpus

El primer parámetro que hemos escogido versa alrededor de la agrupación de los resultados por formas ortográficas idénticas a la hora de la búsqueda por una forma, formato del que dispone exclusivamente el *CHARTA*. Ante el objetivo de estudiar la ortografía, las variaciones gráficas, las secuencias de formas, etc. del español dieciochesco, este formato ofrece la ventaja de poder distinguir instantáneamente las formas ortográficas iguales y de obtener información acerca de la frecuencia absoluta de cada una de ellas. En consecuencia, resulta fácil proceder a los análisis cuantitativos de las diferentes variantes gráficas, de combinaciones de formas ortográficas, etc., mientras que tanto el *CORDIAM* como la base de datos de los textos jesuitas exigen un paso preliminar, a saber, la selección y agrupación de las formas ortográficas que son de interés.

[35] En segundo lugar, tanto el *CHARTA* como el *CORDIAM* ofrecen el contexto sintáctico precedente y siguiente a la forma buscada. El *CORDIAM* es el corpus que se destaca por la práctica visualización del contexto en el primer momento de la búsqueda, además de la posibilidad de acceder al contexto extendido a través de la opción *ejemplo* en la ventana lateral, como hemos visto en la figura 10. Por lo tanto, este formato resulta beneficioso sobre todo en el estudio morfosintáctico, ya que permite ver y analizar la forma buscada no de manera aislada, sino en su contexto sintáctico mayor desde el principio, sin tener que realizar otros pasos, como en el caso del *CHARTA*. En este, dado que muestra primero las diferentes formas ortográficas agrupadas, cabe seleccionar en un segundo paso o bien la palabra, para acceder a los contextos sintácticos de la totalidad de las ocurrencias de la palabra escogida, o bien el año a la derecha para llegar a las ediciones del texto completo, de ahí que requiera más pasos antes de poder observar el con-

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

texto sintáctico. En comparación con los dos corpus electrónicos, que facilitan el contexto mayor, la base de datos de los jesuitas carece de esta posibilidad, por lo que puede considerarse el formato menos apto en este sentido. Visto que la base de datos no ilustra sino los tokens que corresponden al criterio de búsqueda, aparte de una serie de informaciones adicionales, el lema y la clase de palabra, cabe abrir la transcripción entera en un documento propio, lo cual constituye la gran deficiencia de la base de datos.

[36] En cuanto a las informaciones cuantitativas disponibles en la superficie de uso, los tres corpus difieren. La única información que es facilitada por todos los corpus es el número total de ocurrencias correspondientes a los criterios que se han elegido para la búsqueda. El *CHARTA* resulta ser el único corpus que indica tanto la frecuencia de cada una de las formas ortográficas diferenciadas como el número de formas diferentes encontradas en la búsqueda, lo cual se debe a la agrupación por formas idénticas (cf. figura 3). No obstante, pueden obtenerse estas mismas informaciones en el *CORDIAM* y la base de datos de los textos rioplatense tras una selección manual de las formas que resultan de interés. La ventaja del *CORDIAM* respecto a la información cuantitativa consiste en el hecho de que informa al usuario acerca del número de documentos y de palabras sobre los que fue realizada la búsqueda y, adicionalmente, acerca del número de documentos que contienen la forma buscada (Bertolotti & Company Company 2022: 52)[12]. Estos números, que faltan en los otros corpus, pueden revelar informaciones relevantes en lo que respecta a frecuencias relativas, la vitalidad, etc. del fenómeno gráfico o morfosintáctico buscado.

[37] Los tres corpus se diferencian además en cuanto a los metadatos disponibles, de ahí que quepa centrarse en una serie de metadatos que puedan ser de utilidad para caracterizar la historia de la lengua, en nuestro caso el siglo XVIII, así como el contexto sociolingüístico que tanto influye en la representación de una variedad histórica a través de textos escritos. De acuerdo a Arias Álvarez & Hernández Mendoza (2016: 387-391), tanto las características del escribiente (origen dialectal y étnico, sexo, etc.) como del escenario comunicativo (tipología textual, inmediatez o distancia comunicativa, tema del texto, receptor, etc.) entran en juego, por lo que los metadatos disponibles pueden en menor o mayor grado contribuir a la caracterización del español del siglo XVIII en las tres regiones estudiadas. La siguiente tabla ilustra los metadatos sobre los escribientes y los escenarios comunicativos de los que disponen los tres corpus:

[12] Esta información se presenta en la línea arriba de los resultados de búsqueda, en el formato «encontrados x casos en x (de $x_{\text{total de búsqueda}}$) documentos que contienen x (de $x_{\text{total de búsqueda}}$) palabras» (*CORDIAM*, cf. figuras 4, 9, 10).

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

Metadatos	Corpus jesuítico	*CHARTA*	*CORDIAM*
Lugar de redacción del texto	sí	sí	sí
Año del texto	sí	sí	sí
Tipología textual	no/sí	no	sí
Región de origen del escribiente	sí	no	no
Sexo del escribiente	sí	no	sí
Origen étnico del escribiente	sí	no	sí
Destinatario (en el caso de cartas)	sí	a veces	a veces

Tabla 5: Comparación de los metadatos sociolingüísticos de los tres corpus

La totalidad de los tres corpus contrastados proporciona información acerca del lugar y el año en el que fue redactado el texto, constituyendo estos metadatos cronológicos y geográficos la base imprescindible para la selección de los subcorpus. Es archiconocida la relevancia que se deriva de la tipología textual para la lingüística histórica, por lo que los tres corpus poseen información al respecto. No obstante, únicamente el *CORDIAM* ilustra la tipología textual en el momento en el que se muestran los resultados de búsqueda, mientras que la naturaleza de la base de datos jesuítica es conocida solamente de antemano. Si bien el *CHARTA* ofrece la opción de escoger una cierta tipología textual al introducir la búsqueda, no la proporciona en los resultados.

[38] En cuanto a las informaciones sobre el escribiente, resulta difícil corroborar datos, ya que se trata en la mayoría de los documentos de autores sin mayor relevancia histórica. Por lo tanto, en el caso del *CORDIAM* aparecen en la ventana lateral el sexo y la pertenencia étnica del autor, así como, en *síntesis*, el nombre del mismo en algunos documentos, mientras que el *CHARTA* carece de estas informaciones. El nombre del escribiente y del destinatario pueden leerse, sin embargo, en los casos de cartas, cuando se accede a la transcripción entera. La base de datos de los jesuitas es el corpus que posee el mayor número de datos acerca del escribiente (y destinatario de las cartas), gracias al catálogo de Storni (1980), que reúne un sinfín de informaciones acerca de los jesuitas activos en la Provincia jesuítica del Paraguay. En consecuencia, hemos obtenido datos sobre la región y el país de origen de los escribientes y destinatarios, su origen étnico, etc., además de informaciones relacionadas con los cargos dentro de la Compañía de Jesús, las cuales pueden ser de relevancia a la hora de reconstruir el contexto comunicativo concreto, por ejemplo en vistas a la relación entre autor y destinatario.

[39] Otro parámetro relevante trata de las ediciones de las que disponen los corpus estudiados, ya que un corpus ideal tendría que ofrecer una edición múltiple, es decir, tanto la versión paleográfica y crítica como el acceso al facsímil (Kabatek 2016: 7). Este criterio es cumplido únicamente por el *CHARTA*, que cuenta con un triple formato en una misma pantalla, como ilustra la figura 6. Mientras que la edición crítica, que facilita la lectura, se considera útil para estudios morfo-

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

sintácticos y sintácticos, la transcripción paleográfica y la versión facsimilar ofrecen grandes ventajas para el estudio gráfico y braquigráfico (Calderón Campos 2019: 43-45). El corpus *CORDIAM*, en cambio, carece de la edición crítica. Sin embargo, la transcripción paleográfica, que indica además las abreviaturas desatadas, etc., fue realizada para todos los documentos a base de los mismos criterios y el facsímil se encuentra disponible para la mayoría de los documentos. Este doble formato es el que contiene también la base de datos de los textos jesuíticos.

[40] Antes de llegar a los parámetros de naturaleza informática, nos detendremos brevemente en las opciones de procesar las ocurrencias que son de interés del usuario. Mientras que el *CHARTA* no permite descargar las ocurrencias escogidas[13] – solo se pueden copiar y pegar en un documento aparte – para luego procesar y guardarlas, la base de datos y el *CORDIAM* ofrecen diferentes posibilidades. La base de datos jesuítica, por un lado, permite seleccionar dentro del total de ocurrencias una cantidad individual y luego exportar y guardar estos resultados en diversos formatos, tales como PDF, XML o DOCX. La figura 11 ilustra la descarga en PDF de una serie de ocurrencias de la búsqueda realizada en § 5.1, que se reproduce en el mismo formato de la base de datos, a saber, en una tabla con las mismas columnas:

ID_token	token	quelle	folio	absatz	zeile	lemma	wortart
2458	governaba	AG311	f.569v	3	43	gobernar	V
11736	govierna	IdL430	f.633r	2	8	gobernar	V
16817	governarla	JIU99	f.161r	2	3	gobernar	V
25021	govierna	MB374	f.734r	2	8	gobernar	V
25074	govierne	MB374	f.734r	2	12	gobernar	V

Figura 11: Las ocurrencias descargadas de la base de datos

Por otro lado, el *CORDIAM* constituye el corpus de más elaboración en este sentido, ya que permite descargar el documento paleográfico entero en formato PDF, así como el facsímil como imagen, a través de las opciones *PDF* y *facsímil* en la ventana a la derecha. En lo que se refiere a los resultados de búsqueda, el corpus dispone de varias opciones para guardar las ocurrencias, por ejemplo, todas las ocurrencias visibles o las marcadas en un archivo XLSX. A continuación, veremos el extracto del archivo XLSX descargado con una serie de ocurrencias seleccionadas de la búsqueda por la forma *go*ern**:

[13] El facsímil, en cambio, puede guardarse como imagen.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

	Concordancia	Siglo	Año	País acti	Autor (h	Autor i	Tipo textual	Cómo citar
2	...a dado el Maestro de Campo Don, Miguel de Camargo **Gouernador** de esta prouinçia assistiendo perssonalmentte co...	18	1704	NIC	hombre	criollo	Documentos admini	[Año 1704, Nicaragua, Docu
3	...ni peligro, a dichos padres, en los pueblos deste **gouierno**, procurando en ttodo Y por ttodo, lo mayor onra ...	18	1704	NIC	hombre	criollo	Documentos admini	[Año 1704, Nicaragua, Docu
4	...rced el ma(...) / de campo don Miguel de Camargo, **gouernador** de lo politico y militar / desta probincia por s...	18	1705	NIC	hombre	criollo	Documentos jurídic	[Año 1705, Nicaragua, Docu
5	...ro cauildo diferentes / consultas de matherias de **gouierno**, assi / en lo espiritual como en lo temporal, d...	18	1726	NIC	hombre	criollo	Documentos admini	[Año 1726, Nicaragua, Docu
6	...fue nombrado por maestro / de seremonias, y en el **gouierno** de este vuestro / obispo hizo el supra dicho dife...	18	1726	NIC	hombre	criollo	Documentos admini	[Año 1726, Nicaragua, Docu
7	...nta a Vuestra magestad de auer quedado solo en el **Gouierno** d[e] ella por auer fallesido. El Dean, Arcediano ...	18	1726	NIC	hombre	criollo	Documentos admini	[Año 1726, Nicaragua, Docu
8	...Yglesia Cula difucion Causo el quedar solo en el **gouierno** d[e] ella, y Obispado, por auer acontesido la mis...	18	1726	NIC	hombre	criollo	Documentos admini	[Año 1726, Nicaragua, Docu
9	...s sesenta y ocho años. / E yo dicho theniente de **Gouernador** certifico conosco al otorgante y de que asi lo di...	18	1768	CR	hombre	criollo	Documentos admini	[Año 1768, Costa Rica, Docu
10	...y vno de los / que a estas juntas concurren es el **gouernador** dc / esta prouincia, don Domingo Cauello, que diz ...	18	1771	NIC	hombre	criollo	Documentos admini	[Año 1771, Nicaragua, Docu
11	...ado sus poderes a sugeto de sagacidad que huuiera **gouernado** sus derechos con juicio, mas sin embargo siempre ...	18	1788	MEX	hombre	criollo	Documentos entre g	[Año 1788, México, Docume
12	...lmirante / don Thomás Marcos Duque de / Estrada, **gouernador** de la prouincia / de Nicaragua, parte de Nueva / ...	18	s/d	NIC	hombre	criollo	Documentos admini	[Año s/d, Nicaragua, Docum

Figura 12: Las ocurrencias descargadas del *CORDIAM*

Este archivo, que permite seguir procesando los datos del *CORDIAM*, contiene no solamente la ocurrencia con el contexto sintáctico precedente y siguiente, sino también determinados metadatos (siglo, año, país actual, sexo y pertenencia étnica del escribiente, tipología textual), así como indicaciones para citar los ejemplos del corpus.

[41] Con respecto a la lematización, ya hemos señalado que el único corpus completamente lematizado lo constituye la base de datos que contiene los documentos rioplatenses. El proceso informático fue llevado a cabo de manera automática con un estricto control manual, dado que el automatismo no trabajó sobre una edición normalizada, la cual facilitaría la lematización automática, sino sobre la transcripción paleográfica. De este modo, hemos podido garantizar la asignación correcta a los lemas, incluso de formas ortográficas divergentes de la norma actual. En el *CORDIAM*, la lematización automática está en progreso actualmente, es decir que, mientras que ya son lematizadas correctamente las formas ortográficas 'correctas', el programa está siendo mejorado en cuanto al reconocimiento de variaciones gráficas (Bertolotti & Company Company 2022: 56). Una fiel lematización tiene la ventaja de facilitar enormemente las búsquedas, tanto desde el punto de vista ortográfico – la búsqueda por un lema arrojaría todas las variantes gráficas existentes – como desde la morfosintaxis, visto que no haría falta tener en cuenta las posibles grafías. El único corpus que no dispone de una lematización es el *CHARTA*, por lo que constituye en este aspecto el corpus menos beneficioso.

[42] El otro procesamiento informático trata del etiquetado morfosintáctico, esto es, la asignación de las formas ortográficas a clases de palabras, herramienta de la que carecen tanto el *CHARTA* como el *CORDIAM*. La base de datos jesuítica, sin embargo, cuenta con un *POS tagging* automático, que fue igualmente controlado de manera manual a causa de las divergencias de la norma moderna que presenta la edición paleográfica. Gracias al etiquetado, pueden realizarse búsquedas abstractas por clases de palabras o combinaciones de las mismas sin tener que recurrir a los comodines o tomar en consideración la totalidad de representantes de una clase de palabra a la hora de la búsqueda. Por lo tanto, cabe admitir que un etiquetado morfosintáctico, que en el mejor de los casos se basa en una edición normalizada, tiene el enorme potencial de facilitar las búsquedas morfosintácticas y de mejorar, de esta manera, las búsquedas que se llevan a cabo en el ámbito de

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

la lingüística *con* corpus, de ahí que un etiquetado morfosintáctico pudiera ser provechoso y ventajoso para los usuarios tanto del *CHARTA* como del *CORDIAM*.

[43] En resumen, hemos observado que mientras más ediciones, más metadatos concretos y más elaboración informática, etc. tenga un corpus electrónico, mayores serán los beneficios para los usuarios. No obstante, la calidad de un corpus no depende exclusivamente de la cantidad, sino asimismo de la calidad de las herramientas, por ejemplo respecto a la superficie de uso, al procesamiento de los documentos y resultados, a la visualización de las ocurrencias y los metadatos, así como de las opciones de las que dispone. Si bien el *CHARTA* ofrece por ejemplo una triple edición, datos cuantitativos, metadatos, etc., resulta a veces poco práctico en su uso, ya que no permite al usuario ver la tipología textual o descargar los resultados, por ejemplo. El *CORDIAM*, en cambio, se destaca por ser más fácil de usar y por contar con un mayor número de metadatos y de opciones para procesar y guardas los resultados, las transcripciones, los facsímiles, etc. No obstante, ambos corpus carecen de una elaboración informática completa con respecto a la lematización, que está en progreso en el *CORDIAM*, y el etiquetado morfosintáctico. Bien es verdad que la base de datos de los jesuitas cuenta con estas herramientas informáticas, pero está limitada por la falta de visualización del contexto sintáctico. Cabe admitir que la lematización y el etiquetado facilitarían las búsquedas en el *CHARTA* y el *CORDIAM* y mejorarían aún más ambos corpus electrónicos.

7 Reflexiones finales

[44] Tras de la contraposición de los tres corpus para el español del siglo XVIII, recapitularemos brevemente los resultados de esta contribución en calidad de cierre. El siglo XVIII constituye, como hemos expuesto en § 2, un vacío en la investigación diacrónica, que se está llenando paulatinamente a través de estudios lingüísticos dieciochescos que se basan, por una parte, en compilaciones documentales, que muchas veces versan alrededor de una determinada variedad dialectal, y por otra, en corpus electrónicos con una base textual de mayor amplitud, que pueden ser aprovechados de la misma manera.

[45] Ante este objetivo, hemos expuesto en § 3 un corpus dieciochesco inédito que contiene cien documentos, mayoritariamente epistolares, de la región rioplatense, redactados por jesuitas criollos. La base de datos relacional fue elaborada para uso propio tras los dos pasos obligatorios de la lingüística *de* corpus: la fase filológica, que abarca la selección y transcripción de los documentos, así como la fase informática, que incluye la tokeneización, lematización y el etiquetado morfosintáctico (*POS tagging*) de la base de datos SQL. El resultante corpus dieciochesco de la Provincia jesuítica del Paraguay, reducido en tamaño y extensión y homogéneo en cuanto a los escribientes y la tipología textual, constituye una herramienta bien elaborada y útil ante el fin de ampliar la base documental para el siglo XVIII rioplatense-paraguayo, sin que pretenda constituir un corpus histórico como tal.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

[46] En § 4 hemos contrastado, en primer lugar, una serie de parámetros en los dos corpus de comparación, el *CHARTA* y el *CORDIAM*, que nos reveló la mayor extensión del *CORDIAM*. Hemos seleccionado en ambos un subcorpus que contiene documentos del siglo XVIII de España (*CHARTA*) y de la Nueva España (*CORDIAM*), redactados por criollos en el último caso, que servirán para un análisis suprarregional.

[47] Nos hemos centrado en § 5 en dos búsquedas ejemplares, una de naturaleza gráfica y otra morfosintáctica, en los tres corpus con el objetivo de comparar las posibilidades de búsqueda y las superficies de uso desde la perspectiva del usuario. La ventaja de la base de datos, en contraste con el *CHARTA* y el *CORDIAM*, consiste en el hecho de que la lematización, así como el etiquetado morfosintáctico, facilitan las búsquedas porque no es necesario tener en cuenta las variantes ortográficas. En cambio, los dos corpus electrónicos requieren el uso de comodines para lograr buscar de manera más abstracta y, dependiendo de los fenómenos buscados, un mayor número de búsquedas. Mientras que el contexto sintáctico de las ocurrencias proporcionadas puede ilustrarse tanto en el *CORDIAM* (*ad hoc*) como en el *CHARTA* (tras marcar el resultado), la base de datos carece de esta posibilidad, lo cual resulta ser quizá la mayor deficiencia. El corpus *CHARTA* se destaca, además, por la agrupación de los resultados según formas ortográficas y el triple formato (facsímiles, ediciones paleográficas y críticas) y el *CORDIAM* por la cantidad de metadatos y las numerosas opciones útiles en la ventana lateral.

[48] § 6 sirvió de balance para contrastar las ventajas y desventajas de cada uno de los corpus en cuanto a determinados parámetros de uso. De esta comparación hemos deducido que mientras mayor sea el número de metadatos, ediciones, variables cuantitativas, opciones de procesamiento, y sobre todo de procesamiento informático, mejor puede ser aprovechado el corpus por el investigador. Sin embargo, es imprescindible que el manejo de las superficies de uso sea, además, cómodo y fácil, de ahí que el *CORDIAM* resultara el corpus de mayor comodidad y disposición clara para fines lingüísticos. El aspecto que pudiera mejorar aún tanto el *CHARTA* como el *CORDIAM* lo constituyen la lematización y el etiquetado morfosintáctico, que se consideran de gran utilidad en la base de datos de los textos jesuíticos.

[49] A través de los parámetros que hemos contrastado en esta contribución podemos llegar a la conclusión de que una inversión en el procesamiento informático en los corpus electrónicos disponibles podría significar un avance de enorme relevancia para los usuarios filólogos, también en vistas a la combinación de diferentes corpus para un estudio más completo, lo que solo puede lograrse si disponen de una base filológica e informática común. En lo que respecta a las direcciones futuras de los corpus electrónicos, cabe seguir y aprovechar el constante avance en la lingüística computacional, como lo constituyen la optimización de la anotación automática, de algoritmos de aprendizaje de los programas, el desarrollo de herramientas de procesamiento del lenguaje natural, de procesamiento a través de

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

la inteligencia artificial, etc., dado que los progresos en estos campos pueden llegar a enriquecer también la elaboración informática de corpus históricos para la lingüística[14].

[50] En lo que se refiere a los corpus electrónicos para el estudio del español del siglo XVIII, hay que tener en cuenta la representación desequilibrada de las diferentes regiones hispanas, de ahí que quepa seguir contribuyendo con más documentación dieciochesca, sobre todo proveniente de regiones hasta el momento infrarrepresentadas, para igualar la representación geográfica en los corpus. Además, sería ventajoso si dispusiéramos de un mayor número de metadatos sociolingüísticos para poder reconstruir y, de esta manera, entender el contexto comunicativo de cada uno de los documentos para luego deducir aspectos relevantes para las variedades del español del siglo XVIII.

[14] La parte III del volumen publicado por Parodi, Cantos-Gómez & Howe (2022) reúne una serie de contribuciones innovadoras acerca de metodologías y herramientas informáticas novedosas para los corpus electrónicos.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

Abreviaturas y referencias bibliográficas

AGN = Archivo General de la Nación. Fondos AR-AGN.DE/CJ, Sala 9.

Arias Álvarez & Hernández Mendoza 2016 = Beatriz Arias Álvarez, Juan A. Hernández Mendoza 2016. Argumentos dialectológicos y sociolingüísticos que ayudan a la caracterización del español en la nueva España en el siglo XVI. Johannes Kabatek (ed.). *Lingüística de corpus y lingüística histórica iberorrománica*. De Gruyter, 385-400.

Bertolotti, Coll & Polakof 2010 = Virginia Bertolotti, Magdalena Coll, Ana C. Polakof 2010. *Documentos para la historia del español en el Uruguay. Vol. 1. Cartas personales y documentos oficiales y privados del siglo XVIII*. Universidad de la República.

Bertolotti & Company Company 2022 = Virginia Bertolotti, Concepción Company Company 2022. Corpus diacrónicos del español de las Américas. Giovanni Parodi, Pascual Cantos-Gómez, Chad Howe (eds.). *Lingüística de corpus en español. The Routledge Handbook of Spanish corpus linguistics*. Routledge, 45-58.

Calderón Campos 2019 = Miguel Calderón Campos 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. https://revistas.udec.cl/index.php/rla/article/view/1573.

CHARTA = Pedro Sánchez-Prieto Borja (ed.) 2011-. *Corpus hispánico y americano en la red*. https://www.redcharta.es.

Company Company 2012 = Concepción Company Company 2012. El español del siglo XVIII. Un parteaguas lingüístico entre México y España. María T. García Godoy (ed.). *El español del siglo XVIII. Cambios diacrónicos en el primer español moderno*. Peter Lang, 255-291.

Company Company & Bertolotti 2021 = Concepción Company Company, Virginia Bertolotti 2021. Para una historia del español de América. El Corpus diacrónico y diatópico del español de América (CORDIAM). Santiago Muñoz Machado (ed.). *Crónica de la lengua española 2021*. Editorial Planeta, 580-592.

CORDE = Real Academia Española (ed.) 2008. *Corpus diacrónico del español*. http://corpus.rae.es/cordenet.html.

CORDIAM = Concepción Company Company, Virginia Bertolotti (eds.) 2016-. *Corpus diacrónico y diatópico del español de América*. https://www.cordiam.org/.

Davies & Parodi 2022 = Marc Davies, Giovanni Parodi 2022. Constitución de corpus crecientes del español. Giovanni Parodi, Pascual Cantos-Gómez, Chad Howe (eds.). *Lingüística de corpus en español. The Routledge Handbook of Spanish corpus linguistics*. Routledge, 13-32.

Donni de Mirande 2004 = Nélida E. Donni de Mirande 2004. *Historia del español en Santa Fe del siglo XVI al siglo XIX*. Academia Argentina de Letras.

Elizaincín, Malcuori & Bertolotti 1997 = Adolfo Elizaincín, Marisa Malcuori, Virginia Bertolotti 1997. *El español de la Banda Oriental en el siglo XVIII*. Universidad de la República.

Fontanella de Weinberg 1984 = María B. Fontanella de Weinberg 1984. *El español bonaerense en el siglo XVIII*. Universidad Nacional del Sur.

Fontanella de Weinberg 1993 = María B. Fontanella de Weinberg (ed.) 1993. *Documentos para la historia lingüística de Hispanoamérica. Siglos XVI a XVIII. Vol. 1*. Real Academia Española.

García Aguiar 2014 = Livia C. García Aguiar 2014. *El español del siglo XVIII. Edición y estudio de un corpus de documentación municipal malagueña*. Tesis doctoral, Universidad de Málaga. http://hdl.handle.net/10630/8313.

García Godoy 2012a = María T. García Godoy (ed.) 2012. *El español del siglo XVIII. Cambios diacrónicos en el primer español moderno*. Peter Lang.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815

García Godoy 2012b = María T. García Godoy 2012. Introducción. María T. García Godoy (ed.). *El español del siglo XVIII. Cambios diacrónicos en el primer español moderno*. Peter Lang, 9-18.

Gómez Seibane & Ramírez Luengo 2007 = Sara Gómez Seibane, José L. Ramírez Luengo (eds.) 2007. *El castellano de Bilbao en el siglo XVIII. Materiales para su estudio. Documentos lingüísticos del País Vasco*. Universidad de Deusto.

Guzmán Riverón & Sáez Rivera 2016a = Martha Guzmán Riverón, Daniel M. Sáez Rivera (eds.) 2016. *Márgenes y centros en el español del siglo XVIII*. Tirant Humanidades.

Guzmán Riverón & Sáez Rivera 2016b = Martha Guzmán Riverón, Daniel M. Sáez Rivera 2016. Introducción. Martha Guzmán Riverón, Daniel M. Sáez Rivera (eds.). *Márgenes y centros en el español del siglo XVIII*. Tirant Humanidades, 11-20.

Kabatek 2016 = Johannes Kabatek 2016. Un nuevo capítulo en la lingüística histórica iberorrománica: el trabajo crítico con los corpus. Introducción a este volumen. Johannes Kabatek (ed.). *Lingüística de corpus y lingüística histórica iberorrománica*. De Gruyter, 1-17.

Moreno-Sandoval 2022 = Antonio Moreno-Sandoval 2022. Etiquetadores morfosintácticos para corpus en español. Giovanni Parodi, Pascual Cantos-Gómez, Chad Howe (eds.). *Lingüística de corpus en español. The Routledge Handbook of Spanish corpus linguistics*. Routledge, 404-418.

Octavio de Toledo y Huerta 2016 = Álvaro S. Octavio de Toledo y Huerta 2016. Aprovechamiento del CORDE para el estudio sintáctico del primer español moderno (ca. 1675-1825). Johannes Kabatek (ed.). *Lingüística de corpus y lingüística histórica iberorrománica*. De Gruyter, 57-89.

Parodi, Cantos-Gómez & Howe 2022 = Giovanni Parodi, Pascual Cantos-Gómez, Chad Howe (eds.). *Lingüística de corpus en español. The Routledge Handbook of Spanish corpus linguistics*. Routledge.

Rojas Mayer 1985 = Elena M. Rojas Mayer 1985. *Evolución histórica del español en Tucumán entre los Siglos XVI y XIX*. Universidad Nacional de Tucumán.

Rojas Mayer 2000 = Elena Rojas Mayer (ed.) 2000. *Documentos para la historia lingüística de Hispanoamérica. Siglos XVI a XVIII. Vol. 2*. Real Academia Española.

Sáez Rivera & Guzmán Riverón 2012 = Daniel M. Sáez Rivera, Martha Guzmán Riverón (eds.) 2012. *El español del siglo XVIII*. *Cuadernos dieciochistas* 13. https://revistas.usal.es/dos/index.php/1576-7914/issue/view/647.

Sánchez Lancis 2022 = Carlos Sánchez Lancis 2022. Corpus diacrónicos del español de España. Giovanni Parodi, Pascual Cantos-Gómez, Chad Howe (eds.). *Lingüística de corpus en español. The Routledge Handbook of Spanish corpus linguistics*. Routledge, 33-44.

Storni 1980 = Hugo Storni 1980. *Catálogo de los jesuitas de la provincia del Paraguay (Cuenca del Plata) 1585-1768*. Institutum Historicum.

Tognini-Bonelli 2001 = Elena Tognini-Bonelli 2001. *Corpus linguistics at work*. Benjamins.

Torruella Casañas 2017 = Joan Torruella Casañas 2017. *Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación en lingüística*. Peter Lang.

TreeTagger = Helmut Schmid 1994-. *TreeTagger – a part-of-speech tagger for many languages*. https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

DOI: https://doi.org/10.25364/19.2024.12.8 ISSN: 2663-9815