CUADERNOS DE LA ALFAL
N° 6 junio 2014: 130-148
ISSN 2218-0761

EL *CORPUS DIACRÓNICO Y DIATÓPICO DEL ESPAÑOL DE AMÉRICA* (*CORDIAM*). PROPUESTA DE TIPOLOGÍA TEXTUAL

CORPUS DIACRÓNICO Y DIATÓPICO DEL ESPAÑOL DE AMÉRICA (*CORDIAM*). A PROPOSAL FOR A CLASSIFICATION OF TEXTUAL TYPOLOGY

VIRGINIA BERTOLOTTI
Universidad de la República, Montevideo
virginia.bertolotti@gmail.com, bertolotti@cordiam.mx

CONCEPCIÓN COMPANY COMPANY
Universidad Nacional Autónoma de México y Academia Mexicana de la Lengua
concepción.company@gmail.com, company@cordiam.mx

Este artículo describe las características y criterios de conformación de un nuevo corpus computarizado, el *Corpus Diacrónico y Diatópico del Español de América*, *CORDIAM* por sus siglas, alojado y sostenido por la Academia Mexicana de la Lengua, que estará próximamente en la red y será de libre acceso. La parte central del trabajo está dedicada a exponer los problemas y decisiones tomadas para elaborar una tipología textual de los más de 3000 documentos que en este momento integran este corpus. Realizamos una propuesta de cuatro grandes tipos textuales, que creemos da cuenta tanto de homogeneidades estructurales internas de esos grupos como de los modos de circulación de los documentos en los virreinatos y/o colonias hispanoamericanas.

Palabras clave: lingüística de corpus, tipología textual, español de América, historia de la lengua española.

This paper describes the characteristics and processes of conformation of a new electronic corpus, the *Corpus Diacrónico y Diatópico del Español de América*, *CORDIAM*, funded by the Academia Mexicana de la Lengua, which will be of open access. The central aim of the paper is to show the criteria by which a textual typology was elaborated for this corpus. The typology is composed of four large discursive types, each of them having structural homogeneity and empirical support of the ways in which Colonial and Viceroyal American documentation spreaded.

Keywords: corpus linguistics, textual typology, American Spanish, history of the Spanish language.

0. INTRODUCCIÓN. LINGÜÍSTICA DE CORPUS E HISTORIA DE LA LENGUA

No es una novedad decir que para hacer investigación en lingüística se requiere, por lo regular, la existencia de corpus o el levantamiento de corpus, entendido este como un conjunto de datos lingüísticos seleccionados con criterios predeterminados y con el fin de

constituir la evidencia empírica para un campo de investigación[1]. Algunas áreas de la investigación lingüística, tales como la historia de la lengua o la variación lingüística, diacrónica y sincrónica, son, además, obligatoriamente disciplinas de corpus. Quizá sea algo más novedoso hacer lingüística con base en corpus digitales, ya que estos le permiten a aquella, al menos desde el último cuarto del siglo pasado, realizar mejores generalizaciones y afinar la evidencia para las explicaciones. Los corpus digitales, como se sabe, están precisamente construidos para usar herramientas informáticas que no sólo permiten manejar grandes cantidades de datos que pueden llevar a generalizaciones fuertes, sino que sofistican las posibilidades de búsqueda y procesamiento de datos y, además, posibilitan el acceso libre y a distancia a través de la web.

En el campo de la historia de la lengua española, en el que se inserta el *Corpus Diacrónico y Diatópico del Español de América*, *CORDIAM*, existen ya, desde hace algunas pocas décadas, corpus computarizados y disponibles en la web, tales como el *Corpus Diacrónico y Diatópico del Español* (*CORDE*) (www.rae.es), el *Corpus del Español* (*CE*) (www.corpusdelespanol.org) o el *Corpus de Documentos y Textos Españoles Anteriores a 1700* (*CODEA*) (http://demos.bitext.com/codea/) y, en desarrollo, el *Corpus Hispánico y Americano en la Red. Textos Antiguos* (*CHARTA*) (http://www.charta.es/), por citar cuatro bien conocidos. Sin embargo, no existía, hasta la creación de *CORDIAM*, un corpus computarizado y disponible en la red diseñado con el objetivo de estudiar la historia del español en América, tanto en sus características lingüísticas, como en su acontecer diacrónico, como en su dialectología histórica, cuanto en su variación textual no literaria. Tiene, por tanto, *CORDIAM*, como enseguida presentaremos, una especificidad dialectal y amplitud cronológica particular de la que carecen los corpus digitales hasta ahora existentes. *CORDIAM* permitirá hacer la historia interna del español de América, con las matizaciones dialectales requeridas y en los varios niveles de la lengua, bien por países, bien por áreas geográficas, bien en sus contrastes internos y respecto del español peninsular; permitirá realizar una dialectología histórica del español de América; contribuirá a hacer una historia general de la lengua española, sin calificativos restrictivos del tipo *americano*, *peninsular*, *boliviano*, etc., si no se desean tales restricciones; permitirá, asimismo, estudiar con mayor precisión los contactos lingüísticos y los movimientos migratorios que han gestado en buena parte la historia externa de la lengua española en este continente; posibilitará acercarse a la oralidad y a variedades no estándares del español –hasta donde un documento escrito permite tal cosa–, y hará posible desde luego aportar nuevas evidencias empíricas para modelar las teorías en Lingüística Histórica.

Este trabajo tiene dos objetivos, uno general y uno específico. El primero es presentar este nuevo corpus digital, la motivación para constituirlo y sus características generales. El segundo, y central, es mostrar la tipología textual que hemos realizado para agrupar los documentos que hasta ahora integran *CORDIAM* y los ejes que subyacen a esta propuesta.

El trabajo, además de esta breve introducción, contiene cuatro secciones. En las dos primeras, §2 y §3, exponemos las características de *CORDIAM* y los criterios y decisiones para su constitución y estructura. La sección 4, la más extensa, aborda la propuesta de tipología textual antes mencionada y los criterios para llevarla a cabo. Cierran unas conclusiones en §5.

[1] Una definición más amplia y precisa fue propuesta por Tognini-Bonelli: *"A corpus can be defined as a collection of texts assumed to be representative of a given language put together so that it can be used for linguistics analysis. Usually the assumption is that the language stored in a corpus is naturally-occurring, that is gathered according to explicit design criteria, with specific purpose in mind, and with a claim to represent larger chunks of language selected according to a specific typology"* (Tognini-Bonelli 2001: 2)

1. LAS RAZONES PARA LA CREACIÓN DE *CORDIAM*

Es un hecho bien sabido que el español en América suele estar infrarrepresentado o, incluso, en lo que respecta a algunas zonas, no representado en las bases de fuentes disponibles electrónicamente para el estudio del desarrollo histórico de la lengua española. Esto lleva a que la descripción de la variación diatópica y/o diacrónica de la lengua materna de aproximadamente el 90% de la población hispanohablante actual sea un capítulo aún pendiente en la lingüística hispánica en general y, particularmente, en la diacrónica. Las carencias sobre información diacrónica americana se explican, a nuestro modo de ver, al menos por tres motivos, los dos primeros complementarios. En primer lugar, la ausencia de información sobre la historia de la lengua española se debe en buena parte a que no existen descripciones lingüísticas de la gran mayoría de los países hispanohablantes americanos, y, en consecuencia, mucho menos hay, con algunas excepciones, descripciones históricas de esas variedades. En segundo lugar, y este es el problema mayor, las carencias de información diacrónica americana se deben en gran medida a la falta de documentación lingüístico-histórica, transcrita y editada con rigor filológico, fácilmente accesible para la comunidad científica interesada. Un tercer motivo, menor, de tal desatención es, sin duda, que hasta hace relativamente poco la investigación en historia del español y en gramática histórica interna de nuestra lengua había estado caracterizada por un notable castellanocentrismo, explicable, en parte, por las dos razones anteriores. Es un hecho también sabido que la literatura americana virreinal y/o colonial, aunque constituye una fuente de información lingüística muy importante, debe ser tomada con suma precaución para hacer lingüística histórica, ya que las obras literarias americanas, hasta casi el siglo XX, seguían en gran parte modelos lingüísticos peninsulares que no dejan aflorar la identidad lingüística de las diversas variantes dialectales americanas, o eran, por el contrario, obras costumbristas que exacerban los estereotipos lingüísticos más populares, tal es el caso de muchas novelas americanas de la segunda mitad del siglo XIX.

Por lo que respecta a la documentación lingüística filológica americana, la situación, afortunadamente, está cambiando de forma acelerada. En efecto, existen ya numerosos esfuerzos filológicos y ecdóticos, institucionales e individuales, que han dado lugar a diversas colecciones de documentos. Algunas de ellas han sido impulsadas institucionalmente, como es el caso de los surgidos en el marco de la entonces *Comisión para el Estudio del Español de América* de la Asociación de Lingüística y Filología de la América Latina (ALFAL), que dio lugar a las compilaciones de documentación americana de Fontanella de Weinberg (1993) y Rojas (2000, 2001 y 2008), el caso del *Proyecto Documentos Lingüísticos de la Nueva España*, radicado en la Universidad Nacional Autónoma de México, que ha producido ya bastantes frutos (Company 1994), Rivero Franyutti (2000), Reyna (2006), Melis y Rivero (2008) y Ramírez Quintana (2013), o el caso del proyecto *Cibola* de documentos virreinales e independientes del suroeste de los Estados Unidos y de algunas zonas del norte de México que contribuyeron a la colonización y conformación lingüística hispánica de esa zona, proyecto coordinado por Jerry Craddock de la Universidad de California Berkeley, que contiene un número importante de documentos transcritos con criterios ecdóticos, subidos en la red y de libre acceso (http://scholarship.org/uc/rcrs_ias_ucb_cibola). Otras colectas documentales, la mayoría, son producto del trabajo de investigación individual o de grupos más pequeños de investigadores. Tal es el caso de Baranowski (s/d)[2], Bertolotti, Coll y Polakof (2010, 2012),

[2] Baranowski, como también Sanz-Sánchez integran *Cibola Project* y colaboran con CORDIAM de manera individual, aportando acorde con las normas de *CORDIAM* documentos de *Cibola* y otros adicionales.

Carrera de la Red (s/d), Díaz Collazos y Ortiz Vanegas (s/d), Egido (s/d), Elizaincín, Malcuori y Bertolotti (1997), Enguita (s/d), Fernández Alcaide (2009), Fernández Lávaque (s/d), Guzmán (s/d), Huamanchumo (2011), Martínez Martínez (2007), Masih (2009), Mendoza (2000), Parodi (coord. s/d), Postigo de de Bedia y Díaz de Martínez (2009), Ramírez Luengo (2011, s/d), Rivarola (2006), Sanz-Sánchez (s/d), Stéfano y Tejera (2006), Zabalegui (s/d).

Como surge de estas referencias, se trata, en muchas ocasiones, de materiales aún no publicados o publicados en editoriales universitarias, que carecen, como es sabido, de circulación amplia y adecuada. Estas condiciones han impedido que el enorme esfuerzo realizado para crear la infraestructura filológica necesaria para el estudio del español en América haya podido ser aprovechado por toda la comunidad académica.

Para dar solución a este impedimento, merced al trabajo de un equipo de especialistas y con el generosísimo aporte de los investigadores citados arriba provenientes de diversas universidades americanas y europeas, se está construyendo el repositorio documental electrónico llamado *CORDIAM*. Aportará *CORDIAM* a la comunidad científica internacional la infraestructura necesaria para poder avanzar sustancialmente en la historia del español de América, tanto en la historia interna como en su historia externa. Permitirá, como ya señalamos, hacer la dialectología histórica del español en América, posibilitará hacer una gramática histórica del español abarcadora, habilitará la construcción de una historia lingüística que no sea la historia del español estandarizado, y, sin duda, proveerá evidencia empírica nueva y firme para la Lingüística Histórica.

2. LAS CARACTERÍSTICAS Y EL PROCESO DE CONSTRUCCIÓN DE CORDIAM

CORDIAM es un corpus computarizado de acceso libre y abierto y es un corpus de corpus. Como todo corpus lingüístico, es una herramienta de infraestructura para la investigación; contiene un sistema de búsqueda y procesamiento diseñado para el análisis lingüístico; contiene, a diferencia de otros corpus, una plantilla de metadatos de relevancia lingüístico-histórica. Está constituido por documentos americanos exclusivamente; todos ellos se caracterizan por ser no literarios y no periodísticos; todos están recabados directamente de archivo; el corpus tiene una profundidad histórica de 400 años: 1493-1904; abarca los 19 países hispanohablantes de América más el suroeste de Estados Unidos (los territorios pertenecientes antiguamente a la Nueva España), Jamaica, Haití y Guyana; por lo tanto, contiene documentación procedente de 23 actuales países americanos. Veamos con detalle cada una de estas características.

Es un corpus de acceso libre porque cualquier persona podrá consultar sus contenidos sin necesidad de registro o de pago de tasa alguna. Es abierto porque se prevé que siga creciendo con nuevos materiales de los investigadores-colaboradores integrantes de *CORDIAM* o mediante la incorporación de nuevos investigadores-colaboradores con sus textos. En la actualidad, y por estar en etapa de construcción tanto en cuanto al procesamiento de los documentos como en cuanto a la programación, es de acceso limitado. Su gestión económica e institucional está en manos de la Academia Mexicana de la Lengua; las directoras de *CORDIAM* son las autoras de este artículo. Su apertura oficial está prevista para noviembre de 2014 con motivo del Congreso General de las Academias de la Lengua, que tendrá lugar en México.

En la sección anterior hacíamos referencia a un conjunto de corpus construidos con el objetivo de estudiar la diacronía de la lengua española en América y que ahora integran

CORDIAM. Por esta razón, es posible decir que *CORDIAM* es un corpus de corpus, esto es, compila conjuntos documentales ya existentes y hace disponibles, mediante un tratamiento y sistematización informática adecuados, tales materiales para la comunidad académica. Sin embargo, *CORDIAM* no es la mera suma de todos ellos sino bastante más: es un corpus codificado y estandarizado de forma tal que ofrece acceso masivo a datos lingüísticos históricos americanos, así como también a datos geográficos e históricos del documento mismo y de quién lo elaboró, así como también datos diversos sobre los escribientes de tales documentos, como veremos más adelante.

Todos los documentos integrados en *CORDIAM* provienen de corpus recogidos y editados críticamente con el objeto de estudiar alguna zona hispanohablante de América. Todos los documentos fueron escritos en América. Todos los materiales, además, están recabados directamente de archivos americanistas de diversa índole (históricos, religiosos, administrativos, entre otros), existentes en América o en Europa; es decir, *CORDIAM* no integra colecciones documentales preexistentes recabadas por historiadores, sociólogos o antropólogos cuya finalidad no era estudiar la lengua en este continente. En la actualidad los documentos ya cedidos para su uso en *CORDIAM* provienen de 58 archivos o repositorios documentales diferentes.

CORDIAM contiene únicamente textos no literarios y no periodísticos. Como es bien sabido, algunos de los documentos que se conservan en los archivos, a diferencia de los textos periodísticos, los literarios o de los recogidos por historiadores, permiten a los investigadores acercarse mejor a la «inmediatez comunicativa», según el término acuñado por Koch y Oesterreicher (1990; *cf.* también Oesterreicher 1996), y son, por ello, una mejor fuente donde encontrar textos positivamente ponderados entre los muy diversos tipos que permiten mejor hacer historia de la lengua (Oesterreicher, Stoll y Wesch 1998).

Los escritores son en general nacidos en América, sin embargo, como es esperable, se incluyen documentos de españoles en los momentos más cercanos a la llegada europea, en la que todavía no había población nacida en América, y para el caso de Argentina, Chile y Uruguay, se incorporan también algunos textos en español de hispanohablantes no nacidos en América, porque no se entendería la constitución e identidad lingüística de estos tres países sin los fuertes movimientos migratorios de europeos a esas zonas en el siglo XIX.

El corpus abarca una profundidad histórica de 400 años (1493 a 1904), esto es, la época de la conquista, la época colonial / virreinal y el siglo de las independencias. Desde el punto de vista de la geografía histórica, contiene documentos de todos los actuales países americanos de habla española, así como también de territorios que fueron colonizados por el imperio español aunque no formen parte actualmente de lo que suele reconocerse como Hispanoamérica, como es el caso de Jamaica, Haití y Guayana y de territorios pertenecientes al sur de los Estados Unidos de América. En cuanto a información administrativa histórica americana, todos los documentos de *CORDIAM* contienen la adscripción virreinal –abarca los cuatro virreinatos: Nueva España, Perú, Nueva Granada y Río de La Plata–, colonial –por ejemplo, Capitanía General de La Habana, que, como se sabe, nunca estuvo integrada en un virreinato–, y muchos documentos incorporan información de la audiencia o de la gobernación donde fueron gestionados. En resumen, *CORDIAM* contiene información de cuatro ámbitos administrativos y jurídicos históricos americanos: virreinatos, audiencias, capitanías y gobernaciones.

De acuerdo con los datos de los investigadores que han cedido sus corpus a *CORDIAM*, este cuenta en la actualidad con más de 3000 documentos de extensión muy variada. Incluye desde breves billetes hasta textos cronísticos y juicios transcriptos en su totalidad, por lo cual, la cantidad de palabras es un mejor indicador de la extensión del corpus. El acervo documental

CORDIAM superará los cuatro millones y medio de palabras en la primera etapa, en su apertura oficial.

CORDIAM se está construyendo sobre tres pilares: los investigadores-colaboradores de *CORDIAM*, el equipo de informáticos, y el equipo de filólogos. El pilar fundamental es, sin lugar a dudas, el de los investigadores americanos, europeos y norteamericanos que han autorizado el uso informático de sus materiales y, en la mayoría de los casos, han elaborado los metadatos[3].

Cada uno de los documentos está acompañado por una plantilla de metadatos asociados que tiene los siguientes campos: nombre del documento, siglo, año, autor (datos étnicos), autor (datos sexuales), autógrafo, país actual, topónimo actual, topónimo histórico, adscripción histórica, tipo textual, archivo, número de folios, número de palabras aproximado, créditos, facsimilar disponible, síntesis. En (1) presentamos un ejemplo de la plantilla de metadatos con los campos completos.

1. Nombre del documento: 1. Denuncia de un bachiller sobre una bolsa con hierbas que se le halló a una mulata

   Siglo: 18

   Año: 1704

   Autor (datos étnicos): mestizo

   Autor (datos sexuales): hombre

   Autógrafo: sí

   País actual: MEX

   Topónimo actual: Texcoco, Estado de México

   Topónimo histórico: Tezcuco

   Adscripción histórica: Virreinato de la Nueva España

   Tipo textual: Documentos jurídicos

   Archivo: Archivo General de la Nación (México), Inquisición, vol. 727, exp. 24, foja 552r.

   Número de folios: 1

   Número de palabras aproximado: 383

   Créditos: Paloma Paula Reyna Vázquez, *El siglo XVIII en el Altiplano Central de México. Materiales para su estudio. Edición crítica, estudio filológico, introducción y notas*, tesis de licenciatura inédita, México: Universidad Nacional Autónoma de México, 2006.

   Facsimilar disponible: sí

   Síntesis: Denuncia de un bachiller sobre una bolsa con hierbas que se le halló a una mulata.

Los datos de la plantilla son incluidos con objetivos diversos: el registro y la reproducción del proceso de investigación (nombre del documento, el archivo y los créditos), la contextualización diacrónica, diatópica y diastrática del documento (siglo, año, datos étnicos, país actual, topónimo actual, topónimo histórico, adscripción histórica) y las características del documento (autógrafo, tipo textual, número de folios, número de palabras aproximado). Los metadatos permitirán al investigador hacer búsquedas desde los distintos ángulos temáticos

[3] Para la elaboración de los metadatos no provistos por los investigadores-colaboradores y para la homogeneización se contará con el asesoramiento de historiadores especializados en la geografía histórica y la historia político-administrativa de América.

contenidos en la plantilla y permitirán, asimismo, hacer búsquedas cruzadas geográficas, cronológicas, textuales y/o étnicas de distintos tipos.

Los documentos con sus metadatos son incluidos en los programas computarizados de almacenamiento, búsqueda y procesamiento diseñados para el análisis lingüístico por los doctores Alexander Gelbukh y Grigori Sidorov del Instituto Politécnico Nacional (México). Por cuestiones de comodidad en la redacción, nos referimos a este conjunto de programas informáticos y a su interfaz gráfica como CORDIAM-WEB. Al momento de la redacción de este artículo ya han sido editados e incluidos en CORDIAM-WEB 1713 documentos que suman más de un millón de palabras. El promedio de palabras por documento es 710 palabras: el más corto contiene 58 y el más largo contiene 29,086.

Previamente a su inclusión en CORDIAM-WEB los documentos son tratados filológicamente y son revisados en cuanto a algunos aspectos filológicos, ecdóticos y diplomáticos. El objetivo de este proceso de revisión es optimizar el aprovechamiento informático por parte de los usuarios. Por un lado, hay datos que se perderían por la existencia de abreviaturas, tachados o sobreescritos o que se perderían por mantener la unión de palabras del documento original. En el caso de las abreviaturas, hemos desatado en cursivas todas las abreviaturas de aquellos corpus que carecen de desatado. En caso de no haberlo hecho, la solución para encontrarlas es realizar tantas búsquedas como posibles abreviaturas, generando siempre la sensación de que quizá se esté perdiendo información por no haber previsto una abreviatura posible, además de que hay abreviaturas que son más logográficas que alfabéticas. Por otro lado, en el caso de la continuidad entre palabras, hemos separado según los usos gráficos del español actual, ya que un corpus sin separación de palabras según los usos ortográficos actuales, dificulta la realización de las búsquedas. Los *sandhis* comunes en el español antiguo, tales como *desto*, *daqueste*, etc. han sido respetados, porque cualquier usuario especializado podrá fácilmente reconocerlos y buscarlos. Por último, un exceso de información ecdótica y/o diplomática, en notas o incluso dentro del documento por parte del editor que autorizó el empleo informático de sus materiales, dificulta enormemente las búsquedas y la interpretación misma de su resultado. El número de documentos que a la fecha de escribir este artículo ha pasado por este proceso de adecuación para su aprovechamiento informático supera, como ya dijimos, los 1700.

La etapa actualmente en curso culminará, como ya señalamos, en noviembre de 2014 con la apertura de *CORDIAM* a la comunidad académica en ocasión del Congreso de Academias de la Lengua. Además están previstas tres nuevas etapas. La primera se dedicará a la lematización manual de las palabras que no pueden ser lematizadas automáticamente; en el estado de desarrollo informático actual de CORDIAM-WEB, uno de los componentes de este programa lematiza automáticamente un alto porcentaje de las palabras de los documentos, como se explica en García Córdova, pero dada la alta diversidad gráfica de los documentos más la existencia de paradigmas morfológicos supletivos, más otros aspectos de variación morfológica no existentes ya en el español actual – y por tanto de difícil o imposible programación– habrá que hacer una no pequeña labor de lematización manual en un futuro. La segunda etapa, posiblemente simultánea a la anterior, es la inclusión de facsimilares, en aquellos casos en que los investigadores-colaboradores nos los han procurado. La tercera etapa, aún no decidida institucionalmente, supone un cambio cualitativo y consistirá en la inclusión en *CORDIAM* de textos periodísticos, aquellos publicados desde la aparición desde los primeros periódicos americanos, a inicios del siglo XVIII, aproximadamente; no sabemos aún si esta etapa, aunque deseable y enriquecedora para CORDIAM, será viable, porque entraña retos de otra naturaleza.

Un problema general en la construcción de cualquier corpus es la representatividad lingüística de los datos –que los usuarios-investigadores transformarán luego en información

filológica y/o lingüística–. Una correcta representatividad garantiza la fiabilidad del corpus en cuestión. Si bien, a primera vista, esto podría suponer un problema para *CORDIAM*, creemos que tal representatividad / fiabilidad está garantizada a través de un conjunto de recursos. Por un lado, el programa siempre arroja, junto a las concordancias resultantes de las búsquedas, información cuantitativa detallada: el número de ocurrencias en relación al número de documentos en que aparece lo buscado y también el total de palabras, el universo de palabras, contenido en los documentos en los que aparece la ocurrencia. Las dos primeras informaciones cuantitativas son usuales en cualquier búsqueda en los otros corpus diacrónicos digitales existentes para el español, pero la información cuantitativa global del universo de palabras en que se ha realizado determinada búsqueda es, hasta donde sabemos, una innovación de CORDIAM-WEB. Por ejemplo, ante el pedido de que muestre formas terminadas en *-ado* y con cualquier manifestación del lema *haber*, arriba de las concordancias se presenta la información cuantitativa, resaltada en negritas rojas, que puede apreciarse en la Figura 1 a continuación:

Figura 1: Pantalla de búsqueda en la que se ve el universo de palabras y textos de los que se extrae la concordancia

Por otro lado, el conjunto de metadatos de CORDIAM-WEB permite al investigador delimitar su propio corpus dentro del corpus total, ya que puede seleccionar los documentos acorde con ciertas variables. Si bien es posible trabajar con la totalidad del corpus, también es posible generar un subcorpus. Un usuario-investigador puede querer ver sólo textos escritos por mujeres, o sólo escritos en el actual Perú, o sólo escritos en el Virreinato de Nueva Granada o sólo pertenecientes a un determinado tipo textual. En cualquiera de estos casos, podrá restringir su búsqueda mediante la selección de alguno o algunos de los metadatos, que han sido establecidos como variables de búsqueda. No todos los metadatos estarán activos para ser utilizables como criterios de restricción de las búsquedas, porque de haberlo hecho así, los

resultados de la investigación podrían quedar atomizados al generar subcorpus muy escasos en cantidad de palabras y por ello de baja utilidad para hacer generalizaciones lingüísticas, diacrónicas o dialectales, que son dos de los objetivos centrales de *CORDIAM*. Por ello, los metadatos que hemos considerado variables restrictivas para la generación de subcorpus son solamente 7, las más importantes, en nuestra opinión, para lograr los objetivos con que fue creado *CORDIAM*: 1. Siglo, 2. Año, 3. Intervalo de años, 4. Autor (étnico y sexo), 5. País actual, 6. Adscripción histórica y 7. Tipo textual.

El sentido de los metadatos no es solamente poder realizar búsquedas restringidas sino también contextualizar los datos obtenidos por inducción para poder trascenderla y transformar los datos en información. El uso de metadatos como variables permite crear categorías intermedias[4] entre el conjunto del corpus y las ocurrencias, evitando un corpus monolítico. Como ha señalado Kabatek:

> Si a mediados del siglo XX, el objetivo era modificar la exageración estructuralista, hoy en día se trata de modificar, de nuevo, un monolitismo que parte del supuesto de la existencia de una —y solo una— gramática representativa de cada lengua y cada época, monolitismo reanimado por modelos actuales y por una lingüística de corpus en la que se supone que la variación textual no es más que un problema de cantidad y que, a partir de un cierto tamaño de la muestra, la variación se esfuma en la nada del "ruido" estadísticamente irrelevante.
>
> (Kabatek 2008: 12)

Un tipo de metadatos relevante es en qué tipo textual se da una ocurrencia particular. Por esta razón, en la construcción de *CORDIAM* incluimos una tipología textual cuyos criterios de conformación desarrollamos en la sección siguiente.

3. TIPOLOGÍA TEXTUAL PARA *CORDIAM*

Es un hecho aceptado en la investigación en lingüística histórica de los últimos quince o veinte años que el género textual puede ser una de las variables en la difusión y/ o activación del cambio lingüístico (Company 2008). Por ello, entendimos necesario elaborar una tipología textual de *CORDIAM* acorde con las características y objetivos de este corpus.

Los documentos que integran *CORDIAM* requirieron una clasificación textual *ad hoc* porque con ella intentamos integrar y respetar varios hechos complementarios que den cuenta, por una parte, de la diversidad y complejidad textual de la documentación americana y, por otra, de la especificidad lingüística de grupos de documentos. Los criterios tenidos en cuenta responden, básicamente a la pregunta ¿qué buscaría un usuario en una tipología de un corpus documental americano no literario y no periodístico en la red? Los criterios *a grosso modo* fueron: *a*) el modo de circulación del documento en cuestión en la América virreinal y/o colonial, por ejemplo circulación privada vs. circulación pública o administrativa; *b*) el curso administrativo que siguió el documento, es decir, cómo o por qué llegó a un determinado archivo, por ejemplo, si fue un juicio o fue un informe al rey, y *c*) diferencias recurrentes estructurales lingüísticas, tales como recurrencias en *usus scribendi*, gramaticales, léxicas y semántico–pragmáticas. Lo esperado es que haya diferencias en estos tres criterios según grupos o tipos textuales. Evitamos

[4] En 1996, EAGLES realizó las primeras recomendaciones consensuadas sobre las condiciones mínimas que tiene que cumplir un corpus: *a*) el corpus debe ser tan grande como sea posible de acuerdo con la tecnología de la época; *b*) la gama de material debe ser amplia con el fin de lograr algún tipo de representatividad; *c*) **es recomendable que haya una clasificación intermedia entre el corpus mayor y las ocurrencias individuales**, y *d*) el corpus debe tener su origen explicitado (EAGLES 1996a, traducción y resaltado nuestros).

a toda costa la atomización casi natural que la compleja administración americana refleja en una primera ojeada, como muestra el listado de documentos en (2) más abajo. Veamos con cierto detenimiento estos planteamientos generales que guiaron la tipología textual de *CORDIAM*.

3.1. Lectores y usuarios

Una cuestión por demás obvia pero definitoria es que CORDIAM no tendrá lectores. Ningún corpus computarizado los tiene. Solo tendrá usuarios, ya que, a pesar de estar constituidos por textos los corpus computarizados no son la suma de ellos sino objetos de naturaleza distinta: son siempre diversas posibilidades de concordancias, son el lugar en donde encontrar evidencia que permita sustentar, descartar, formular hipótesis de análisis, así como realizar o ampliar una investigación.

3.2. Leer textos y «leer» concordancias

Asociada a la distinción anterior, vale la pena recordar la distinción entre la forma en que se lee un texto y aquella en que se «lee» un corpus, realizada por Tognini-Bonelli (2001: 3-4) que presentamos esquematizada y traducida aquí abajo:

Texto	**Corpus**
se lee como unidad	se “lee” fragmentariamente
se lee horizontalmente	se “lee” verticalmente
se conectan unidades, oraciones, párrafos	se focaliza en las concordancias obtenidas y, a partir de ellas, se va al cotexto
se lee en tanto que evento único	se “lee” en tanto que evento repetido
se lee como producto de un acto individual y voluntario	se “lee” la concordancia como parte de una práctica social
es una instancia de *parole*	ilumina sobre la *langue*
es un evento comunicativo coherente	no es un evento comunicativo coherente

3.3. Universalidad, generalidad o particularidad

La tipología que hemos creado no tiene pretensión universal alguna, ya que el dominio sometido a tipologización es el de los textos que son la base de *CORDIAM*. Los ejes o parámetros que son utilizados para la construcción de nuestra tipología, como se verá en el apartado 3.7, son generales, cuando tienen que ver con un corpus lingüístico-histórico computarizado, o son particulares, cuando surgen de las características de los tipos de documentos incluidos en *CORDIAM*. Ninguno de los ejes considerados, creemos, podría considerarse universal. En el mejor de los casos, se trata de una tipología generalizable a textos no impresos del mundo iberorromance, posiblemente románico occidental, entre los siglos XV y XIX.

3.4. Proporcionalidad

CORDIAM es, en cierto sentido, un corpus cuantitativamente pequeño, aunque vaya a tener en su momento de salida casi cinco millones de palabras. Aunque es esperable que en el futuro se amplíe, su representatividad / fiabilidad difícilmente se basará en sus muchos millones de palabras, tal como sucede, por ejemplo, en el *CORDE* o en el *CE*.

Esto se explica por dos razones. Una primera es que los corpus que constituyen la base documental de *CORDIAM* –o de cualquier corpus para la lingüística diacrónica– no se pueden construir sumando cualquiera de los documentos que están en un archivo sino que se construyen por una selección mediante criterios muy estrictos que combinan filología, ecdótica y lingüística histórica. Por ello, en general, sólo un porcentaje bajo de los textos contenidos en los archivos cumple con las condiciones necesarias para ser integrados a un corpus creado con fines lingüístico-históricos. Una segunda razón es que, una vez seleccionado un conjunto de «buenos» documentos, el investigador comienza el proceso de reproducción del original, de transcripción y de corrección de la transcripción, lo cual lleva, en el más fácil de los casos, unas cinco horas por folio; es decir, es cosa sabida que no es una tarea fácil construir un corpus computarizado con varios miles de documentos procedentes de archivo. Por tanto, dado que el corpus es y será relativamente pequeño, la cantidad de posibilidades por variable o metadatos debería ser también pequeña, ya que, de otra manera, se corre el riesgo de atomización ya comentado, y se reduce la representatividad. Esas pocas categorías textuales, debían ser, además, lo suficientemente abarcadoras como para dar cuenta del largo listado de medio centenar de clases de textos ya identificados por los investigadores-colaboradores de CORDIAM, que listamos en el próximo apartado.

3.5. Tipos, clases de textos y géneros discursivos

La finalidad de realizar una tipología de un corpus como *CORDIAM* es eminentemente operativa. La revisión de los tipos de textos propuestos por los investigadores-colaboradores a partir de sus corpus fue una larga lista de clases de textos que recogemos parcialmente, en orden alfabético, en (2):

2. Actas de bautismo, actas de cabildo, actas fundacionales, autos por juicios de residencia, bandos, capitulaciones, cartas de oficiales, cartas de particulares dirigidas a instancias oficiales, cartas entre particulares, decretos, denuncias, descripciones geográficas, informes, inventarios de barcos, inventarios de bienes de difuntos, juicios de residencia, civiles o criminales, juicios de memoriales o relaciones de méritos, nombramientos de cargos diversos, notitas, recibos privados y pagarés de diversa naturaleza, ordenanzas, padrones, peticiones de embarque, peticiones de mercedes, probanzas de méritos y de limpieza de sangre, procesos judiciales, querellas civiles y criminales, relaciones de expediciones o sucesos, sentencias, testamentos, testimonios en juicios.

Tres cuestiones surgen de la lectura de (2): el número de etiquetas es muy alto, no se trata de una tipología sino una enumeración de clases y la lista podría ser todavía aun más extensa. Desde un punto de vista práctico, el conservar esta clasificación no hubiera resultado en un aporte para el usuario sino en un problema y hubiera reducido la utilidad de *CORDIAM*. La búsqueda de determinado dato lingüístico limitada a uno de esas clases de documentos hubiera reducido el número de concordancias significativamente, dando resultados muy atomizados y cuantitativamente muy pobres. Existe, sin duda, la posibilidad informática de ofrecer al usuario más de una opción de búsqueda; un investigador podría elegir, entonces, hacer una búsqueda en actas de bautismo, actas de cabildo y padrones, y otro investigador en cartas de particulares, cartas oficiales e informes. Si bien eso es posible, sería poco recomendable en cuanto a la

amigabilidad del programa, ya que para realizar cada búsqueda el usuario debería revisar una lista superior a las cincuenta etiquetas para seleccionar las pertinentes para una determinada investigación. Además, esta posibilidad informática debilitaría los objetivos esenciales de *CORDIAM*, que, recordemos una vez más, es posibilitar la investigación en la gramática y la dialectología históricas del español americano. Tomando en cuenta todo lo anterior, decidimos definir unos pocos tipos a partir de unos pocos ejes, parámetros o criterios. Examinemos antes de mostrar nuestra propuesta, algunas tipologías de géneros discursivos ya existentes.

3.6 Algunas tipologías textuales ya existentes

Para realizar nuestra propia tipología textual para *CORDIAM*, consultamos algunos trabajos previos y evaluamos su pertinencia para una tipología como la necesaria para *CORDIAM*.

EAGLES, por ejemplo, hace una propuesta de criterios que deben ser tomados en cuenta para tipologizar; aparece resumida y traducida a continuación:

> Género literario (poesía, narrativa, (auto)biografía, novela, *nouvelle*, novela histórica, ciencia ficción, humor, dramaturgia), tópico, medio (libros, cartas/correspondencia, periódicos, folletos, volantes), ficción o no ficción, estilo (distancia, distendido o solemne, especializado o técnico), otros (manuales y libros de texto).
>
> (EAGLES 1996b)

Si hubiéramos tomado en cuenta estos criterios de tipologización, hubiéramos tenido que distribuir los documentos de *CORDIAM* entre tipos creados cruzando, por un lado, lo no literario, no impreso, no ficción y, por otro, el continuum de estilo y la variopinta clase de tópico[5]. De haber seguido esta clasificación, hubiéramos sumado a los problemas operativos de trabajar con el continuum que presenta la documentación virreinal y/o colonial americana (de mayor a menor distancia comunicativa entre escribiente y destinatario, o de más distensión a más solemnidad) los derivados de trabajar con una categoría muy atomizada –y mucho más apegada al mundo que a la lengua– como es la de tópico.

Las tipologías textuales realizadas en el ámbito de la lingüística también fueron revisadas. Si bien ya es sabido, sobre todo a partir de los trabajos de Adam (1992), que tipologizar con criterios intralingüísticos es una tarea infructuosa, buscamos una asociación entre algunas rutinizaciones lingüísticas contenidas en los textos y algunas características externas reconocibles fácilmente. La tipología propuesta es una clasificación-guía de lo lingüísticamente «esperable» en el campo del léxico, de algunos fenómenos morfológicos, de ciertos aspectos de la sintaxis; de cuestiones propiamente textuales y de bastantes regularidades semántico-pragmáticas.

Las valiosas reflexiones y propuestas de Koch y Oesterreicher (1990) y Oesterreicher (1996) nos llevaban a agrupamientos demasiado grandes o demasiado inespecíficos para *CORDIAM*. Si tomamos, por ejemplo, el concepto de inmediatez, no-inmediatez comunicativa propuesto por estos autores como un categorizador, entonces, íbamos a tener que la mayor parte del corpus caería dentro de esa categoría, ya que su influyente propuesta guió en buena medida la construcción de muchos de los corpus que integran *CORDIAM*. No se nos escapa que, en la realidad, los textos podrían ser todos ubicados en esta línea de tipologización, en sendos continuos que fueran desde lo más privado a lo más público, desde lo menos planificado a lo

[5] Los tópicos o temas definidos a partir de corpus existentes son los siguientes, listados en el orden en que son propuestos: religión, tecnología, legal, deportes, arte, política, historia, medicina, filosofía, economía, educación, psicología, ciencia, ocio, civilización, física, biología, matemática, hogar, comunicación, lenguaje y lengua, literatura, arquitectura, moda/indumentaria, computación, agricultura, geografía, ecología/medioambiente, tráfico/ transporte, química, finanzas (EAGLES 1996c).

más planificado, desde lo concebido más oralmente a lo concebido más escrituralmente, etc. Sin embargo, esta forma de pensar no por adecuada es operativa para un corpus informatizado.

Los trabajos de Biber (1985, 1986) y Biber y Conrad (2009), que establecen un conjunto de rasgos lingüísticos, proponen tres parámetros que dan cuenta de la variación textual: interactivos-editados; descontextualizados-situados y de estilo diferido-de estilo inmediato. Si tomamos por ejemplo los rasgos «interactivo-editado» resultaría que muchos de los textos caerían dentro de las dos categorías: un juicio tiene partes fuertemente descriptivas y también partes dialógicas. Por tanto, adscribirlo a una de las dos clases sería fuertemente inespecífico y por lo tanto una mala guía para el usuario de *CORDIAM*. Si, como proponen Biber y Conrad (2009), los tres parámetros de definición que determinan un continuum se entrecruzaran, ganaríamos en especificidad, pero, por otra parte, no tendríamos tipos sin asociaciones probables con conjuntos de rasgos que puedan graduarse. Esto no sería útil para el usuario o sería de muy escasa utilidad real para entender mejor la documentación antigua americana. Además, el hecho de tener que analizar rasgos y combinatoria de rasgos por texto, desplazaría el trabajo del equipo de *CORDIAM* de la construcción del corpus al de analistas y caracterizadores de textos, y, de nuevo, se perdería el objetivo esencial de la creación de *CORDIAM*.

Revisamos, igualmente, tipologías basadas en documentación americana (Barbosa y Lopes 2003; Bertolotti, Coll y Polakof 2010; Guzmán com. pers., Virkel 2010). Coincidimos con las dos primeras en tomar como parámetro la forma de circulación de los textos. Si bien el análisis de todas ellas nos ha resultado iluminador, en la medida en que atienden la especificidad de sus propios corpus, no fue posible retomarlas en su totalidad. La complejidad documental, heterogeneidad geográfica y profundidad histórica de *CORDIAM* requería una tipología hasta cierto punto simple pero fidedigna de los cursos administrativos e históricos que siguió la documentación en la América virreinal y/ o colonial y fidedigna, hasta donde ello fuera posible, de los hablantes-escribientes que produjeron esa documentación. La pregunta que nos volvíamos a formular era qué hace de una tipología una buena tipología.

3.7. Los ejes taxonómicos que estructuran *CORDIAM*

Luego de realizar las reflexiones y revisiones presentadas arriba, decidimos realizar la tipología textual preguntándonos no por los textos sino por los usuarios de *CORDIAM*. Resolvimos preguntarnos qué buscaría un usuario en una tipología de un corpus en red con las características del nuestro[6]. Consideramos que, para orientar las búsquedas y facilitar la interpretación de las concordancias arrojadas por el programa, los usuarios de *CORDIAM* pondrían en juego sus conocimientos o intuiciones sobre algunas regularidades posibles asociadas a textos agrupables, subsumibles en un tipo textual.

Un parámetro relevante, y sin duda primario, para cualquier lingüista o filólogo es que el texto haya sido escrito para la circulación pública o no. Por tanto, el primer eje taxonómico, fue la circulación del documento y las razones para su conservación en un archivo. Esto nos permite distinguir un primer grupo de textos que son los que hemos agrupado bajo la etiqueta *Documentos entre particulares: cartas y otros*. Este tipo de documentos no está escrito para trascender sino que son producto de la necesidad de comunicación entre dos particulares, lo que condiciona unas determinadas regularidades en secuencias textuales, léxicas, sintácticas y morfológicas, además de (orto)gráficas. Suelen tocar tópicos personales y, muy importante,

[6] Trabajamos inicialmente con una propuesta de ocho tipos que redujimos a cinco y más tarde a cuatro, luego de discutirla en un Seminario realizado en el Instituto de Lingüística de la Universidad de la República (Uruguay) con Magdalena Coll, Marisa Malcuori, Cristina Píppolo y Ana Clara Polakof, a quienes agradecemos la disponibilidad para oír los problemas y planteamientos y la generosidad de sus aportes.

llegan a los archivos, en general, por azar, aunque no siempre sea así. Es decir, este tipo de documentos fue escrito para ser leído por otro particular y, al momento de ser escrito, no se pensó –puede haber excepciones– en que el documento podría llegar alguna vez a una instancia oficial. Son sin duda, por esa conciencia de privacidad, la «joya de la corona» para la historia de la lengua, tanto porque suelen mostrar una elevada inmediatez comunicativa como porque son el único tipo de documento que se atreve a escribir el que no «sabe» escribir.

Los documentos que fueron concebidos teniendo en cuenta su circulación pública son de índole muy diversa y fueron ordenados de acuerdo con su funcionalidad social, tomando en cuenta, por cierto, el funcionamiento cultural e institucional durante la Conquista, la Colonia y las primeras décadas de las independencias americanas. Distinguimos dentro de los documentos de circulación pública tres grandes tipos: *Documentos cronísticos*, *Documentos jurídicos* y *Documentos administrativos.*

En una caracterización muy somera, es posible afirmar que los *documentos cronísticos* describen paisajes, relatan sucesos raros o curiosos para el «cronista», describen acciones propias de ciertos grupos humanos, creencias, costumbres, festividades y comportamientos de esos grupos y tienen, muchas veces, aunque no necesariamente, una ordenación temporal. Se acercan a un texto literario sin tener, sin embargo, una finalidad estética. Predominan en ellos descripciones y narraciones.

Los *documentos administrativos* ordenan, registran, disponen y regulan la vida cotidiana, con descripciones y lineamientos sumamente detallados, las más de las veces; desde cómo deben transitar hombres y animales por las antiguas calles hasta cómo comportarse en un espectáculo público; dan también cuenta de bienes materiales, de vivos y de difuntos; dan testimonio de la genealogía de los individuos y de su lugar y fecha de nacimiento. Básicamente, dan cuenta del mundo en que un individuo se mueve cotidianamente y de las instancias que regulan esa cotidianeidad e interacción oficial entre personas.

Los *documentos jurídicos* se producen en el mundo del Derecho. En este sentido, se acercan bastante a los documentos administrativos. Sin embargo, y a diferencia de estos, son documentos de tipología textual muy compleja y, hasta cierto punto, heterogénea, ya que contienen tipos de documentos dentro de otros tipos (denuncias, querellas, postulación de preguntas, interrogatorio y respuestas, sentencias, segunda instancia de un juicio, traslados, etc.), a manera de cajas chinas o muñecas rusas; esto es, son textos que pueden ser identificados como una unidad no obstante que al entrar en ellos podemos encontrar otras unidades que, a su vez, al entrar en ellas pueden contener otras unidades, etc. Por esta razón, y a pesar de circular en forma similar a los administrativos, suelen tener múltiples enunciadores, y múltiples despliegues discursivos internos. Además, desde un punto de vida lingüístico, especialmente los subtipos documentales ‘interrogatorio / declaración’, suelen ser textos altamente dialógicos que nos aproximan en cierta medida a la oralidad de los declarantes.

Esta tipología nos enfrenta al problema –que evitamos pero no resolvemos– de la ontología del documento. En *CORDIAM* optamos por una definición operativa y consideramos *documento* la unidad determinada por un colaborador-investigador y que él haya entendido como tal por su validez empírica para hacer lingüística histórica o historia de la lengua[7]. Por lo tanto, nos encontramos tanto con juicios transcriptos en su totalidad como con un fragmento de un juicio

[7] La única excepción son las cartas. Dado el enorme valor de estos textos para la Lingüística Histórica, cuando se encuentran cartas incluidas en un documento mayor, con el acuerdo del investigador colaborador, se la separa del documento madre independizándola como un documento propio.

consistente en la sentencia o con un fragmento de un juicio consistente en las declaraciones de testigos.

En resumen, *CORDIAM* contiene cuatro tipos textuales o discursivos, 1. documentos entre particulares: cartas y otros, 2. documentos cronísticos, 3. documentos jurídicos y 4. documentos administrativos. Hemos intentado en esta tipología textual, como ya dijimos, evitar la atomización haciendo agrupamientos grandes que tengan cierta «homogeneidad» estructural interna; son cuatro etiquetas generales que respetan el funcionamiento de la administración americana virreinal / colonial. Los cuatro tipos textuales, *a grosso modo*, operan sobre un eje-continuum: más privado > más público, cuyos cortes internos no tienen, como es lógico, fronteras nítidas.

En la Tabla 1 distribuimos la larga lista de textos de (2) que habíamos incluido más arriba en los cuatro grandes tipos que acabamos de definir.

Tipo de texto	**Entre particulares: cartas y otros**	**Cronísticos**	**Jurídicos**	**Administrativos**
Clases de textos actualmente presentes en CORDIAM	Cartas personales Notitas Recibos Pagaré	Cartas de oficiales Descripciones geográficas Informes Relaciones de expediciones o sucesos	Actas de cabildo Autos de juicio de residencia Bandos Capitulaciones Decretos Denuncia Juicios de residencia Memoriales de méritos Probanzas de méritos Probanza de limpieza de sangre Procesos judiciales Querellas Sentencias Testimonios en juicios	Actas de bautismo Actas fundacionales Cartas de oficiales Cartas de particulares Informes Inventarios de barcos Inventarios de bienes Nombramientos Padrones Pagarés Peticiones de merced Testamentos

Tabla 1. Clases de textos distribuidos por tipos

3.8. Ejemplos de tipos textuales

A continuación mostramos un fragmento de documentos adscritos a cada uno de los tipos propuestos.

Documentos entre particulares: cartas y otros

3. Mi mas estimada y querida esposa de / mi corazon me alegrare que al rresibo desta / te alles con la salu que yo para mi deseo / en conpañia, de mis dos amadas sijas /[5] de mi corason y de tu familia y mia / la que yo difruto es buena para que me / mandes que lo are como me toca de obligasion / Juana esta se dirije a /[10] notisiarte …

30. Carta de Josef de Mesa a su esposa, 1803, Uruguay

Documentos cronísticos

4. Y / así diçen que los vnos salieron de qüebas, los otros de çerros, /25 y otros de fuentes, y otros de lagunas y otros de pies de árboles, / y otros desatinos desta manera; y que por auer salido y enpeçado / a muntiplicar destos lugares y auer sido de allí el prinçipio / de su linaje, hizieron guacas y adoratorios estos lugares / en memoria del primero de su linaje que de allí proçedió; /30 y así cada nación se uiste y trae el traje con que a su guaca / uestían.

Relación de las fábulas y ritos de los Ingas, de Cristóbal de Molina, ca. 1600, Perú

Documentos jurídicos

5. yo Jose Candido Baes besino de el pueblo de antimano y residente de la Ciudad de San Felipe / Ante V paresco y digo que el rreo nombrado ylario Silba es un honbre que me a sentensiado a muerte con una lansa que a sacado en mi misma casa y por no aber tenido los testigo (sic) no me presente ante V y de contra A una muJer que tengo en mi Casa a sacado un puñal para matarla en la casa de el Señor Miguel Bara por un pique que tie <inter: ne> con hella por una mujer que el tenia y llo la hise salir de el Sitio de Carapa y por Cullo motivo Cuantas beses pasa por mi casa a distintas horas de la noche se benga Con pegar un astaso a las tiJas de mi CoRedor que estan a la bista las tiJas quebradas en dicho Coredor

Contra Hilario Silva por haber amenazado con una lanza á varias personas, y por otros excesos, 1837, Venezuela

Documentos administrativos

6. Muy mag*níf*ico se*ñ*or: / El que la presente lleva es J*ua*n Freyle, que / a servido en esta haçienda de varvero para curar / los enfermos. El qual començó a servir dende /[5] quinçe de março, año de 1556 años. Sirvió hasta / quinçe de nobienbre del dicho año. Ganava a raçón / de çien p*es*os de minas cada año, que ansí estava / conçertado. Dévensele ocho meses como pareçerá / el asiento por el libro de la contaduría.

Carta escrita por Rodrigo Marín, 1557, México

3.9. Algunas observaciones finales

Cabe realizar algunas puntualizaciones que pueden ayudar al lector a comprender mejor esta tipología textual.

Nuestra tipología cumple con el criterio de buena tipologización de que los tipos no se incluyan unos a otros ni se superpongan. Es cierto que una relación de una expedición podrá ser tomada luego como un elemento en un juicio, una cartita entre amantes podrá ser utilizada como prueba en un juicio de limpieza de sangre. Pero, si están en un documento mayor, el conjunto que integran será adscripto al tipo *Documentos jurídicos*.

Las clases de textos pueden pertenecer a más de un tipo. Por ejemplo, un documento clasificado como *carta de un oficial* podrá pertenecer tanto a los documentos administrativos como a los cronísticos. Si en la *carta de un oficial*, este informa a su superior de la cantidad de dinero que necesita recibir para mantener la tropa, se lo considerará un *Documento administrativo*. Sin embargo, una *carta de un oficial* (incluso del mismo oficial) en la que relata sus avances en un territorio nuevo en donde describe la geografía física y humana a la que se va enfrentando será incluido entre los *Documentos cronísticos*.

No tendrán doble asignación los individuos que integran la clase, esto es, ningún documento estará etiquetado en dos tipos. Si dentro de un juicio hay un inventario de un barco, el documento es el juicio y por lo tanto irá a *Documentos jurídicos*. Es esperable que el usuario de *CORDIAM*, que dispondrá de un contexto más amplio que la mera concordancia, reconozca la

circunstancia, nada infrecuente, de la existencia de textos que tienen un origen y que luego se incluyen en textos con otros objetivos y otras formas de circulación, como ya mencionamos.

Las copias y traslados, de reconocido valor filológico, no son caracterizadas fuera de la clase a la que pertenece su original.

4. CONSIDERACIONES FINALES

Y hablando de tipos textuales… este artículo se aleja del artículo clásico, con una introducción, una presentación de la metodología y los datos, una discusión de los datos y una conclusión. Por lo tanto, no presentamos aquí conclusiones sino estas consideraciones a modo de cierre.

Relatamos aquí dos experiencias. En primer lugar, la experiencia de construcción colectiva, con la comunidad académica de la lingüística histórica hispánica, de una infraestructura para la investigación de la historia del español de América.

En segundo lugar, la experiencia intelectual de construir una tipología para *CORDIAM*, alternando inducción con deducción, revisando categorías construidas, pasando del lugar de lector al de usuario de corpus, pasando del lugar de investigador con corpus en papel al lugar del investigador con corpus computarizado.

De pretensión modesta, la tipología que construimos parece ser empíricamente adecuada y funcional. Su aplicación a cada vez más textos, dirá si efectivamente lo es y si puede ser utilizada para otros corpus históricos de otros ámbitos romances.

REFERENCIAS BIBLIOGRÁFICAS

Adam, Jean Michel. 1992. *Les textes: types et prototypes*, Paris, Nathan Éditions.

Barbosa, Afrânio Gonçalves; Célia Regina Lopes dos Santos. 2003. Corpora do projeto para a História do português Brasileiro de 1997 a 2003, en A. Teixeira do Castilho (Ed.) *Historiando o português brasilerio*, 139-154. Disponible en: http: //www.mundoalfal.org/Ataliba%20T.htm <17 de junio de 2013>

Bertolotti, Virginia; Magdalena Coll; Ana Clara Polakof. 2010. Presentación, en V. Bertolotti, M. Coll; A. C. Polakof. *Documentos para la historia del español en el Uruguay. Vol. 1. Cartas personales y documentos oficiales y privados del siglo XVIII*, Montevideo, Facultad de Humanidades y Ciencias de la Educación: 9-18.

Biber, Douglas. 1985. Investigating macroscopic textual variation through multifeature / multidimensional analyses, *Linguistics* 23: 337-360.

Biber, Douglas.1986. Spoken and written textual dimensions in English: Resolving contradictory findings, *Language* 62, 2: 384-414.

Biber, Douglas; Susan Conrad. 2009. *Register, genre and style*, Cambridge, Cambridge University Press.

Company Company, Concepción. 2008. Gramaticalización género discursivo y otras variables en la difusión del cambio sintáctico, en J. Kabatek (ed.), *Sintaxis histórica del español y cambio lingüístico: Nuevas perspectivas desde las Tradiciones Discursivas*, Madrid/Frankfurt: Iberoamericana/Vervuert: 17-51.

EAGLES. 1996a. Preliminary recommendations on Corpus Typology, [en línea] Disponible en: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html <17 de junio de 2013>

EAGLES. 1996b. Textual tipology, [en línea] Disponible en:
http://www.ilc.cnr.it/EAGLES96/texttyp/node36.html#SECTION00011200000000000000 <17 de junio de 2013>

EAGLES. 1996c. Topics, [en línea] Disponible en:
http://www.ilc.cnr.it/EAGLES96/texttyp/node37.html#SECTION00011300000000000000 <17 de Junio de 2013>

Kabatek, Johannes. 2008. Introducción, en Kabatek, Johannes. 2008. (ed.) *Sintaxis histórica del español y cambio lingüístico: Nuevas perspectivas desde las Tradiciones Discursivas*, Madrid/Frankfurt, Iberoamericana/Vervuert: 7-16.

Koch, Peter; Wulf Oesterreicher. 1990. *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch*, Tübingen: Max Niemeyer.

Melis3, Chantal; Agustín Rivero. 2008. *Documentos lingüísticos de la Nueva España. Golfo de México*, con la colaboración de Beatriz Arias, México, Universidad Nacional Autónoma de México.

Oesterreicher, Wulf. 1996. Lo hablado en lo escrito. Reflexiones metodológicas y aproximación a una tipología, en Kotschi, Thomas; Wulf Oesterreicher; Klaus Zimmermann. 1996. (eds.), *El español hablado y la cultura oral en España e Hispanoamérica*, Madrid/Frankfurt, Iberoamericana/Vervuert: 317-340.

Oesterreicher, Wulf; Eva Stoll; Andreas Wesch. 1998. (eds.), *Competencia escrita, tradiciones discursivas y variedades lingüísticas. Aspectos del español europeo y americano en los siglos XVI y XVII*, *ScriptOralia, 112*, Tübingen.

Sinclair, John. 1991. *Corpus, concordance, collocation*, Oxford, Oxford University Press.

Tognini-Bonelli, Elena 2001. *Corpus linguistics at work*, Amsterdam, John Benjamins.

Virkel, Ana 2010. *Documentos fundacionales de Chubut, Patagonia. Período 1865-1899.* Chubut, Universidad Nacional de la Patagonia.

CORPUS QUE ACTUALMENTE ESTÁN INTEGRADOS A *CORDIAM*

Baranowski, Edward. s/d. *Documents from the 1602-3 Sebastián Vizcaíno Expedition up the California Coast*, Cíbola Project, Research Center for Romance Studies, Berkeley, University of California Berkeley.

Bertolotti, Virginia; Magdalena Coll; Ana Clara Polakof. 2010. *Documentos para la historia del español en el Uruguay. Vol. 1. Cartas personales y documentos oficiales y privados del siglo XVIII*, Montevideo, Facultad de Humanidades y Ciencias de la Educación.

Bertolotti, Virginia; Magdalena Coll; Ana Clara Polakof. 2012. *Documentos para la historia del español en el Uruguay. Vol. 2. Cartas personales y documentos oficiales y privados del siglo XIX*, Montevideo, Facultad de Humanidades y Ciencias de la Educación.

Carrera de la Red, Micaela. s/d. (coord.) *Corpus documental de Colombia (siglo XVI)*, Transcripciones inéditas.

Company Company, Concepción. 1994. *Documentos lingüísticos de la Nueva España. Altiplano central (1525-1816)*, México, Universidad Nacional Autónoma de México.

Craddock, Jerry s/d (coord.) *Selección de documentos coloniales del sudoeste de los Estados Unidos provenientes del proyecto Cibola*, Disponible en: http://scholarship.org/uc/rcrs_ias_ucb_cibola

Díaz Collazos, Ana María; Yamileth Ortiz Vanegas. s/d. *Documentos de la Audiencia de Santa Fé de Bogotá, siglo XVI*, Transcripciones inéditas.

Egido, María Cristina. s/d. *Documentos del oriente de Bolivia (s. XVII-XVIII*). Transcripciones inéditas y versión revisada de 2III: Bolivia-b" en Rojas, E. 2008. (ed.), *Documentos para la Historia Lingüística de Hispanoamérica. Siglos XVI a XVIII*, III, Anejo LX del *Boletín de la Real Academia Española*, Madrid: 457- 483.

Elizaincín, Adolfo, Marisa Malcuori; Virginia Bertolotti. 1997. Documentos en *El español en la Banda Oriental del siglo XVIII*, Montevideo, Facultad de Humanidades y Ciencias de la Educación, Universidad de la República: 65-132.

Enguita Utrilla, José María. s/d. *Relación de las fábulas y ritos de los Ingas, de Cristóbal de Molina*, Transcripción inédita.

Fernández Alcaide, Marta. 2009. *Cartas particulares de Indias del siglo XVI. Edición y estudio discursivo*, Madrid/ Frankfurt, Iberoamericana/ Vervuert.

Fernández Lávaque, Ana María. s/d. *Diez cartas de Salta del siglo XIX*. Transcripciones inéditas.

Fontanella de Weinberg, María Beatriz. 1993. (comp.) *Documentos para la historia lingüística de Hispanoamérica. Siglos XVI al XVIII.* Madrid, Real Academia Española.

Guzmán, Martha. s/d. *Documentos del siglo XV al XVII de las Antillas*, Transcripciones inéditas.

Huamanchumo de la Cuba, Ofelia. 2011. Transcripción del «Apéndice D. Visita: 1549 Guanca» en: *Encomiendas y cristianización. Análisis pragmático de documentos jurídicos y administrativos del Perú (Siglo XVI)*, München: Mikroform Karl-Heinz Limbeck: 338-341.

Martínez Martínez, María del Carmen. 2007. *Desde la otra orilla. Cartas de Indias en el Archivo de la Real Chancillería de Valladolid (siglos XVI-XVIII) (edición, estudio, notas e índices)*, León, Universidad de León.

Masih, Mariela. 2009. *Cartas coloniales. Córdoba (Argentina). Siglos XVI-XVII*, Córdoba, Editorial Babel.

Mendoza, José G. 2000. *100 documentos para la historia lingüística de Bolivia*. *Siglos XVI-XVIII (Alto Perú)*, La Paz, Facultad de Humanidades y Ciencias de la Educación-Universidad Mayor de San Andrés.

Parodi, Claudia. s/d. (coord). *Relación de Chimalhuacán o pueblo de Sanct Andrés Apóstol (siglo XVI)*, Transcripción inédita.

Postigo de de Bedia, Ana María y Lucinda Díaz de Martínez. 2009. *Documentos del Jujuy Colonial. Aportes para el estudio histórico del español americano (siglos XVI a XIX)*, Jujuy: Universidad Nacional de Jujuy.

Ramírez Luengo, José Luis. 2011. Un corpus para la historia del español en Nicaragua: edición de documentos oficiales del siglo XVIII (1704-1756), *Moenia* 17: pp. 333-366.

Ramírez Luengo, José Luis. s/d. *Textos para la historia del español de Centroamérica. Honduras (siglos XVII-XIX)*, Transcripciones inéditas.

Ramírez Luengo, José Luis. s/d. *Textos para la historia del español de Centroamérica. El Salvador (siglos XVIII-XIX)*, Transcripciones inéditas.

Ramírez Quintana, Pedro Ángel. 2013. *Documentos lingüísticos de la Nueva España. Provincia de Campeche*, México/Campeche, Universidad Nacional Autónoma de México / Universidad Autónoma de Campeche.

Rivarola, José Luis. 2006. *Documentos lingüísticos del Perú. Siglo XVI y XVII. Edición y comentarios.* Madrid, Consejo Superior de Investigaciones Científicas.

Reyna Vázquez, Paloma Paula. 2006. *El siglo XVIII en el Altiplano Central de México. Materiales para su estudio. Edición crítica, estudio filológico, introducción y notas*, tesis de licenciatura, México, Universidad Nacional Autónoma de México, Inédita.

Rivero Franyutti, Agustín. 2000. *Aproximación al español mexicano en el siglo XVI: edición crítica y estudio filológico de un conjunto de cartas (1537-1557)*, tesis de doctorado, México, Universidad Nacional Autónoma de México, Inédita.

Rojas, Elena. 2000. *Documentos para la Historia Lingüística de Hispanoamérica, II*, Anejos del *Boletín de la Real Academia Española*, 58. Madrid, Real Academia Española.

Rojas, Elena. 2001. *Documentos para la Historia Lingüística de Hispanoamérica, III*, Tucumán, Universidad Nacional de Tucumán.

Rojas, Elena. 2008. *Documentos para la Historia Lingüística de Hispanoamérica, IV*, Anejos del Boletín de la Real Academia Española, LXI, Madrid, Real Academia Española.

Sanz-Sánchez, Israel. s/d. Transcripciones inéditas y transcripciones revisadas de Sanz-Sánchez, Israel. 2009. *The diachrony of New Mexican Spanish, 1683-1926*: *Philology, corpus linguistics and dialect change*, tesis de doctorado inédita, Berkeley, University of California, Inédita.

Stéfano, Luciana de; María Josefina Tejera. 2006. *Documentos para la Historia del Español de Venezuela siglos XVI-XVII.* Publicación en CD, Caracas, Departamento de Publicaciones. Universidad Central de Venezuela.

Zabalegui, Nerea. s/d. (coord.), *Corpus de documentos escritos en Venezuela (siglos XVIII y XIX)*, Transcripciones inéditas.