RLA. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 57 (2), II Sem. 2019, pp. 41-64.

CL ISSN 0033 - 698X

LOS CORPUS DEL ESPAÑOL CLÁSICO Y MODERNO: ENTRE LA FILOLOGÍA Y LA LINGÜÍSTICA COMPUTACIONAL[1]

CLASSIC AND MODERN SPANISH CORPORA: BETWEEN PHILOLOGY AND COMPUTATIONAL LINGUISTICS

MIGUEL CALDERÓN CAMPOS
Universidad de Granada
calderon@ugr.es

RESUMEN

En este artículo se analiza la práctica habitual de elaboración de corpus del español, tanto europeo como americano, del periodo comprendido entre finales del siglo XV y finales del XIX. Se prestará especial atención al modelo seguido por seis corpus diacrónicos: CHARTA, CODEA 2015, CORDIAM, *CorLexIn*, *Post Scriptum* y *Cíbola*, con el objeto de extraer conclusiones metodológicas aplicables a trabajos futuros o iniciales, como el corpus *Oralia diacrónica del español* (ODE), actualmente en fase de elaboración en la Universidad de Granada. El análisis efectuado revela que, si bien no se aprecian grandes diferencias en el rigor y los criterios de transcripción documental, no parece haber acuerdo en lo referente a la manera de procesar y estructurar la información, tanto textual como metatextual. En este trabajo se defiende la conveniencia de adoptar un modelo estandarizado basado en el uso de XML, siguiendo las directrices del consorcio TEI para la codificación y etiquetado de corpus históricos. Este modelo permitirá la integración de diferentes corpus y, sobre todo, un más fácil acceso a la información por parte de los usuarios.

Palabras clave: Historia de la lengua española, corpus diacrónicos, lingüística de corpus, XML, oralidad en los textos escritos.

ABSTRACT

This article analyses the standard practice when compiling and producing European and American Spanish corpora for the period spanning from the end of the 15th century to the late 19th century. Special attention will be given to the model used for six diachronic

[1] Este estudio se ha llevado a cabo dentro del proyecto "*Hispanae Testium Depositiones*: las declaraciones de testigo en la historia del español. HISPATESD", de referencia FFI2017-83400-P (MINECO/AEI/FEDER, UE).

corpora: CHARTA, CODEA 2015, CORDIAM, *CorLexIn*, *Post Scriptum* and *Cíbola*, in order to reach methodological conclusions applicable to any future or incipient projects - such as the *Oralia diacrónica del español* (ODE) corpus, currently being prepared at the University of Granada. The analysis shows that while there are no appreciable differences in the rigor and criteria applied to document transcription, there does not seem to be any agreement as to the way to process and structure the information - textual as well as metatextual. This paper will argue for the usefulness of adopting a standardized model based on the XML markup language, following the TEI consortium guidelines for the codification and labelling of historical corpora. This model will make it possible to integrate the different corpora and, more importantly, to provide easier user access to the information.

Keywords: History of the Spanish language, diachronic corpora, corpus linguistics, XML, orality in written texts.

Recibido: 23/05/2019. Aceptado: 15/11/2019.

1. INTRODUCCIÓN

Entre 2010 y 2014 se elaboró en la Universidad de Granada el *Corpus diacrónico del español del reino de Granada, 1492-1833* (CORDEREGRA)[2]. Se trataba de realizar un corpus de "pequeño dominio" del antiguo reino nazarí, castellanizado a partir de 1492. El territorio comprendía *grosso modo* las actuales provincias de Granada, Málaga y Almería. El proyecto supuso el primer intento de compilar un corpus sistemático a partir de documentación manuscrita inédita de esta zona de Andalucía que tiene, además, el atractivo de haberse castellanizado al mismo tiempo que el territorio americano. Se seleccionaron principalmente dos tipos textuales: declaraciones de testigos en juicios criminales e inventarios de bienes, por su cercanía a la oralidad[3].

Durante los cuatro años del proyecto, la mayor parte del esfuerzo realizado se centró en la selección documental y en la transcripción paleográfica de los manuscritos. Una vez concluida la fase puramente filológica empezaba una nueva etapa, de carácter técnico, que planteaba retos difíciles de resolver desde la perspectiva tradicional. El mayor de todos era ofrecer un corpus en línea que permitiera realizar búsquedas complejas y cruzar la información léxica con metadatos cronológicos, geográficos o tipológicos.

Este artículo surge precisamente del intento de resolver los problemas informáticos que se interponen entre la labor filológica y la constitución de un auténtico corpus digital en línea, que no sea una mera presentación en PDF de las transcrip-

[2] Actualmente el corpus CORDEREGRA no está disponible en red, puesto que se ha incluido en *Oralia diacrónica del español* (ODE, http://corpora.ugr.es/ode).

[3] Una antología de esta documentación puede verse en Calderón Campos (2015).

ciones realizadas. Parece un buen punto de partida analizar cómo se han elaborado otros corpus similares, que puedan servir de modelo para el trabajo futuro. Por este motivo, en el apartado 3 se analizan seis corpus diacrónicos, ya disponibles en línea, que incluyen documentación manuscrita del mismo periodo que el antiguo CORDEREGRA, actualmente en fase de integración en un corpus de mayor amplitud geográfica y temporal: *Oralia diacrónica del español (ODE)*. Los corpus que se han tomado como modelo son CHARTA, CODEA 2015, CORDIAM, *CorLexIn*, *Post Scriptum* y *Cíbola*[4]. Previamente, en el apartado 2 se presentan los criterios (Torruella, 2017) que se tendrán en cuenta para realizar el estudio comparativo.

2. CRITERIOS PARA LA CLASIFICACIÓN DE CORPUS DIACRÓNICOS

Los seis corpus seleccionados recogen documentación manuscrita del periodo comprendido entre 1492 y 1900, aunque difieren notablemente en las características técnicas y en la amplitud geográfica, tipológica o cuantitativa. Para proceder a su descripción se tendrán en cuenta siete parámetros: cronología, extensión geográfica, tipología textual, tamaño del corpus, tipo y número de ediciones, anotación lingüística y gestión informática. Todos ellos se usarán en el apartado 3 para analizar los seis corpus textuales.

Los cuatro primeros criterios servirán para describir de manera general cada uno de los corpus seleccionados. Los tres últimos se refieren a cuestiones técnicas sobre las que merece la pena detenerse (epígrafes 2.1-2.3) antes de realizar el estudio comparativo.

2.1. Tipo y número de ediciones. La normalización ortográfica del corpus

Al realizar la edición digital de un corpus, los compiladores pueden optar por presentar una sola versión de cada documento (una edición paleográfica, por ejemplo) o por ofrecer distintas ediciones de los manuscritos: transcripción paleográfica, presentación crítica, edición normalizada, reproducción fotográfica o facsimilar, etc.

La transcripción paleográfica respeta los usos gráficos del manuscrito (*io*, *qvarto*, *aRoba*, *seys*, *sseñor*, *marauedys*, *thenjente*, *nottario*, *tress*, *ffue*, etc.) y, en las versio-

[4] CHARTA y CODEA 2015 incluyen cualquier tipología textual, desde los textos más formales, como los legislativos, hasta los más humildes, como las notas de entrega de niños en la inclusa; *CorLexIn* se centra exclusivamente en inventarios de bienes, principalmente del siglo XVII; *Post Scriptum* recoge únicamente cartas privadas; por último, en *Cíbola* se incluye toda la documentación conservada, de carácter administrativo, privado o cronístico, relacionada con la conquista y colonización del suroeste de los Estados Unidos.

nes más estrictas, mantiene la separación de palabras del original (*dela*, *enlos*, *amodo*, *gelo*, *con duzentes*, *haviendo lo*, *primera mente*, *quales quier*, etc.). Por su parte, la edición crítica reduce la variabilidad gráfica sin valor fonético (CHARTA, 2013). Por ejemplo, la forma paleográfica *hefeto* puede simplificarse en la crítica *efeto*, eliminando una "h" antietimológica sin trascendencia fonética. Paralelamente, las formas originales *hagora*, *jurediçion*, *algund*, *julljo*, *mill*, *parrochial*, *defuncto*, etc. se editan críticamente como *agora*, *juredición*, *algún*, *julio*, *mil*, *parroquial* y *defunto*, sin alterar la interpretación fonética de esas palabras. Lo mismo cabría decir de casos del tipo *iusto* / *justo*, *maior* / *mayor*, *uisto* / *visto*, *vna* / *una*, etc.

La presentación crítica moderniza también la separación de palabras, el uso de mayúsculas y minúsculas, la acentuación, y añade puntuación a los textos. Podría decirse que la transcripción paleográfica está especialmente indicada para quienes se planteen estudios directamente relacionados con la historia de la ortografía; la edición crítica facilita la lectura de los textos y se orienta a los estudios morfológicos, sintácticos o léxicos (CHARTA, 2013), aunque no está reñida con los análisis fonéticos pues, como se ha visto, no modifica grafías fonéticamente pertinentes[5].

El último paso en la modernización del texto se da en la llamada "edición normalizada" o estandarizada (Vaamonde, 2015), entendiendo como tal aquella en la que el manuscrito original se adapta por completo a los criterios ortográficos actuales. Así, a las formas paleográficas *hagora*, *jurediçion* y *defunto* les corresponden las modernas *ahora*, *jurisdicción* y *difunto*. Este tipo de ediciones están destinadas a quienes se acercan al corpus con intereses distintos de los lingüísticos, de tipo cultural, histórico o sociológico. Además, facilitan labores computacionales posteriores, como la lematización y el etiquetado morfosintáctico.

Tabla I. Tipos de edición.

Paleográfica	Crítica	Normalizada
hefeto	efeto	efecto
hagora	agora	ahora
defuncto	defunto	difunto
jurediçion	juredición	jurisdicción
algund	algún	algún
uisto	visto	visto
julljo	julio	julio
heçebto	ecebto	excepto

[5] En la edición crítica, obviamente, no se modernizarían usos seseantes (*dise*, *lisensidado*, etc.) o yeístas (*llo*, *yevaba*, *balloneta*, *fayesido*, etc.). Tampoco se estandarizarían formas propias de la época como *muncho*, *ansí*, *naide*, *trujo*, *recebí*, etc.

El reto desde el punto de vista computacional está en poder vincular la edición normalizada con la paleográfica y la crítica, de manera que se puedan encontrar en una sola consulta todas las variantes ortográficas asociadas a una forma estándar. Cuando esto ocurre, se dice que el corpus está normalizado. Este aspecto es especialmente útil en lingüística histórica, dada la enorme vacilación ortográfica de los textos antiguos. Por ejemplo, la forma moderna *recibe* puede escribirse en los siglos XVI y XVII de muchas maneras distintas: *reçiue*, *reziue*, *rezive*, *recibe*, *rescibe*, *rrezive*, etc. Lo mismo cabría decir de las variantes *hefeto*, *efeto* y *efecto*, mencionadas arriba. Si el corpus está normalizado, la búsqueda del estándar actual (*recibe*, *efecto*) arrojará todas las formas ortográficas vinculadas en una sola consulta (*reçiue*, *rescibe*, *hefeto*, *efeto*, etc.).

Para suplir la carencia de normalización, algunos corpus permiten el uso de expresiones regulares en las que se sustituyen cadenas de caracteres por comodines[6] (?, *). Aun así, la normalización es un recurso mucho más potente para el análisis léxico, no solo por la rapidez de las consultas, sino también por su alcance; piénsese que el empleo de comodines puede no ser suficiente para localizar variantes poco previsibles, como *estrébedes*, *extrébedes* o *trebes* ('trébedes'), o la forma *aijada* ('aguijada'), por ejemplo.

2.2. Anotación lingüística: lematización y etiquetado morfosintáctico

Otro aspecto relevante de los parámetros clasificatorios es el de la anotación lingüística. Las formas ortográficas o *tokens* de un corpus pueden etiquetarse con información sobre el lema con el que se relacionan (lematización) o sobre la categoría gramatical a la que pertenecen (etiquetado morfosintáctico).

El lema es la palabra que encabeza los artículos de un diccionario, algo así como el "representante" de las variantes morfológicas de una palabra: el infinitivo, para las formas verbales, el masculino singular para los sustantivos. Algunos corpus están lematizados, es decir, han asignado un lema a cada *token*. Este proceso simplifica enormemente las búsquedas, puesto que permite obtener de una sola vez todas las formas vinculadas con el lema. Así, la búsqueda lematizada de *recibir*, por seguir con el ejemplo anterior, proporcionaría cualquier forma de la conjugación de este verbo: *recibo*, *recibimos*, *recibiría*, *recibiendo*, etc.

Por último, etiquetar un corpus consiste en asignar a cada *token* una marca informativa sobre su categoría gramatical (sustantivo, adjetivo, adverbio, verbo,

[6] Los signos ? y * se pueden usar como comodines para hacer consultas amplias: el signo de interrogación (?) sustituye a un carácter cualquiera en una posición determinada. Así, la búsqueda de "a?ul" podrá arrojar los resultados *azul*, *açul* o *asul*. Por su parte, el asterisco (*) equivale a cualquier número de caracteres, desde cero hasta infinito. De esta forma, si se busca "azul*" se obtendrá como resultados posibles *azul*, *azules*, *azulgrana*, *azulejo*, *azulón*, etc.

etc.). Esta marcación proporciona a los corpus un nivel más abstracto de búsquedas, dado que permite encontrar combinaciones enormemente variadas, del tipo "adjetivo + sustantivo", "preposición *en* + gerundio", "artículo + nombre propio", etc.

2.3. Gestión informática del corpus

Se trata de determinar el grado de autonomía que tienen los creadores del corpus (es decir, los filólogos encargados de la selección y trascripción documental) en su gestión informática. En algunos corpus están muy separadas la fase filológica de la computacional. Terminada la transcripción de los documentos, los filólogos acaban su cometido y dejan el resto del trabajo a expertos informáticos, que se encargan de todas las tareas propiamente de lingüística de corpus (visualización de la documentación en la web, normalización, lematización, etiquetado, etc.).

Esto quiere decir que los transcriptores carecen de independencia para modificar los textos que se visualizan en línea, y que necesitan del apoyo de personal técnico para introducir cambios en el corpus publicado. En algunos casos, este modelo tradicional puede ralentizar el proceso de configuración del corpus, puesto que sus auténticos creadores no tienen capacidad para actualizar la información disponible, es decir, no pueden subir nuevos documentos o modificar los existentes, en los que pueden haber detectado errores.

Según este criterio se dividirán los corpus entre corpus de gestión autónoma o interna, cuando los compiladores gozan de independencia para gestionar la mayor parte de las tareas computacionales, y corpus de gestión externa, si los filólogos terminan su tarea en la fase de transcripción documental.

3. ESTUDIO COMPARATIVO DE SEIS CORPUS DIGITALES

En este apartado se analizan seis corpus en línea que incluyen documentación manuscrita de los siglos XVI a XIX, con el objetivo de aprender de la experiencia acumulada en los últimos años y aplicarla a proyectos futuros de características similares.

3.1. Los corpus CHARTA y CODEA 2015

El *Corpus hispánico y americano en la red: textos antiguos* (CHARTA) es el proyecto de más largo alcance de los seis analizados, pues se concibe como un macrocorpus

hispánico, europeo y americano, del periodo comprendido entre 1200 y 1800 (Sánchez, 2012). Recoge documentación manuscrita de muy diversa tipología, desde documentos oficiales cancillerescos hasta notas breves y cartas privadas. En el momento de la consulta (abril de 2019) cuenta con un total de 1.346.094 palabras, distribuidas de manera desigual, tanto cronológica como geográfica o tipológicamente. El siglo más representado es el XIII, con el 41,18% de la documentación total, mayoritariamente peninsular. Desde el punto de vista tipológico, predominan los textos de alta codificación, legislativos (38,17%) y notariales (cartas de compraventa y contratos, 39,34%). Por tanto, hasta la fecha, CHARTA es un corpus mayoritariamente medieval (71,3 % del total)[7], peninsular (93,5%) y formal (más de un 90%).

Los textos de CHARTA no están normalizados ortográficamente, lematizados ni etiquetados. En compensación, se pueden usar comodines (? y *) para agilizar y hacer más productivas las consultas. Los investigadores, según sus intereses, optan por dos opciones de visualización de los datos: la presentación clásica en forma de concordancia, con la palabra clave en contexto (PCEC o KWIC), o la triple presentación, uno de los aspectos más destacados del corpus, que hace posible el análisis pormenorizado de cada documento particular, del que se ofrece en la web la reproducción fotográfica[8] acompañada de la transcripción paleográfica y de la edición crítica.

Un aspecto muy destacable de CHARTA es que todos los documentos han sido transcritos expresamente para el corpus, siguiendo unos criterios homogéneos de edición, consensuados en 2013 por los equipos que integran la red internacional (CHARTA, 2013). Tanto el rigor filológico de las transcripciones como la visualización en triple formato hacen de CHARTA un referente fundamental en la compilación actual de corpus históricos del español.

El *Corpus de documentos españoles anteriores a 1800* (CODEA + 2015) está estrechamente vinculado con CHARTA. Ha sido creado por el grupo de investigación GITHE (*Grupo de investigación de textos para la historia del español*, Universidad de Alcalá de Henares), dirigido por Pedro Sánchez-Prieto Borja (Miguel Franco y Sánchez, 2016). Toda la documentación se ha transcrito expresamente para CODEA, siguiendo los criterios de la red CHARTA.

En abril de 2019, CODEA contaba con 1.445.162 palabras, distribuidas en 2491 documentos del español peninsular de los siglos XII a XVIII (véase Gráfico 1).

[7] El 21,7% de la documentación pertenece a los siglos XVI y XVII; solo el 7% data de los siglos XVIII y XIX.

[8] Lo cual permite comprobar la fiabilidad de las transcripciones.

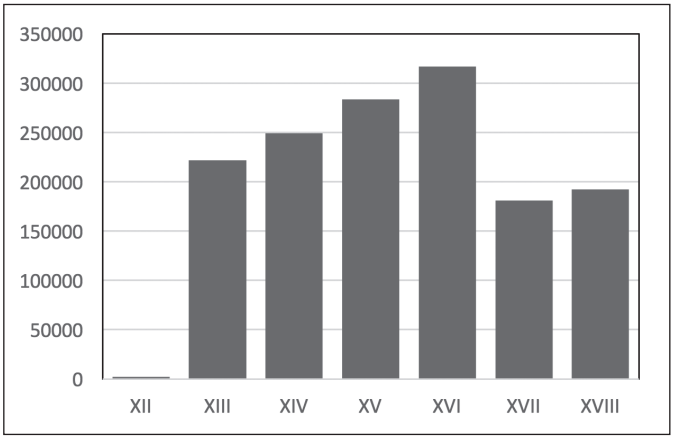

Gráfico 1. Número de palabras por siglo en CODEA 2015.

El tipo documental (véase Gráfico 2) más representado es el de "cartas de compraventa y contratos", que supone el 33% del total de la documentación[9]. A este le siguen los textos legislativos (21% del total)[10] y las actas y declaraciones (17%). Con un volumen similar, próximo a las 100.000 palabras (aproximadamente el 7% del total, cada uno de ellos) están representados los testamentos e inventarios de bienes, las cartas privadas[11] y los certificados. Se completa el corpus con textos narrativos ("Informes y relaciones"), estatutarios y notas breves, la mayoría de abandono de niños en la inclusa (Sánchez y Vázquez, 2017).

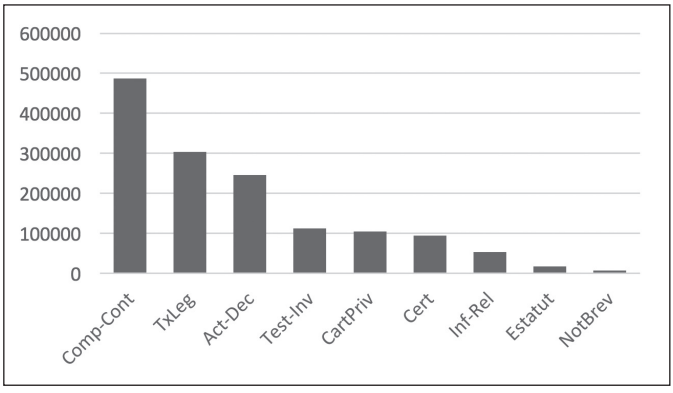

Gráfico 2. Número de palabras por tipo textual en CODEA 2015.

[9] De este 33%, el 85% es documentación anterior al año 1500. Es decir, la mayor parte de las cartas de compraventa y contratos pertenecen a la Edad Media.

[10] De este 21%, el 61% lo constituyen textos medievales. El 80% de los textos anteriores a 1500 está constituido por cartas de compraventa y contratos y textos legislativos.

[11] Todas ellas de los siglos XVI, XVII y XVIII.

Desde el punto de vista técnico, CODEA se construye en la misma plataforma que CHARTA. Aunque actualmente los textos no están normalizados, lematizados ni etiquetados, todas estas tareas se acometerán en la nueva edición (CODEA 2020), en la que además está prevista la ampliación cronológica hasta el año 1900 y la incorporación de tecnología de web semántica, que permitirá recorrer, en ambos sentidos, el camino de la forma paleográfica al más abstracto de la familia léxica. Tanto CODEA como CHARTA reciben una gestión informática externa.

Otra novedad, ya disponible en la edición de 2015, es la posibilidad de ver los resultados de las búsquedas en mapas dialectales, lo que convierte el corpus en una herramienta dinámica muy valiosa para los estudios de dialectología histórica (Sánchez, 2018).

3.2. El *Corpus diacrónico y diatópico del español de América* (CORDIAM)

CORDIAM es el corpus más grande de los analizados (Bertolotti y Company Company, 2014). Cuenta en el momento de la consulta (marzo de 2019) con un total de 7.341.245 palabras. Se divide en tres subcorpus: *CORDIAM-Documentos*, de algo más de cuatro millones de palabras; *CORDIAM-Prensa* (siglos XVIII y XIX), casi 1.700.000; y *CORDIAM-Literatura*, ligeramente inferior al millón y medio. Recoge documentación de todos los países americanos: los más representados son México (2.517.880 palabras), Perú (1.250.867) y Uruguay (974.446), seguidos de Colombia y Chile (en torno al medio millón de palabras), y de Argentina, Estados Unidos, Bolivia y Venezuela (alrededor de 300.000 palabras cada uno). El resto está por debajo de las 100.000 palabras (véase Gráfico 3).

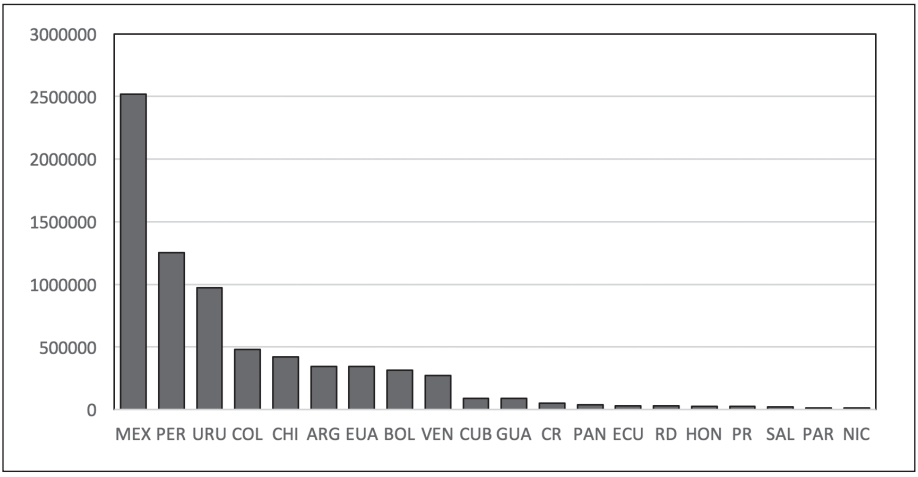

Gráfico 3. Tamaño de CORDIAM por países.

Presenta documentación desde 1494 hasta 1905, distribuida bastante homogéneamente entre las cuatro centurias representadas (véase Gráfico 4).

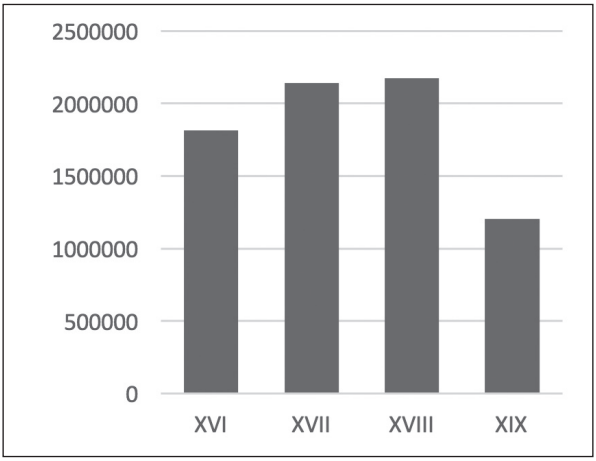

Gráfico 4. CORDIAM: distribución por siglos.

Por último, tienen cabida en CORDIAM todas las tipologías textuales, como se aprecia en el Gráfico 5. Predominan los documentos jurídicos (1.600.249) y administrativos (1.370.839 palabras) y los textos literarios cronísticos (1.315.906), seguidos de la prensa informativa (937.930), la documentación entre particulares (712.533), la prensa de comentarios (665.857) y los documentos cronísticos (562.460). Menos representación tienen los documentos publicitarios y la literatura en prosa y verso (los tres por debajo de las 100.000 palabras).

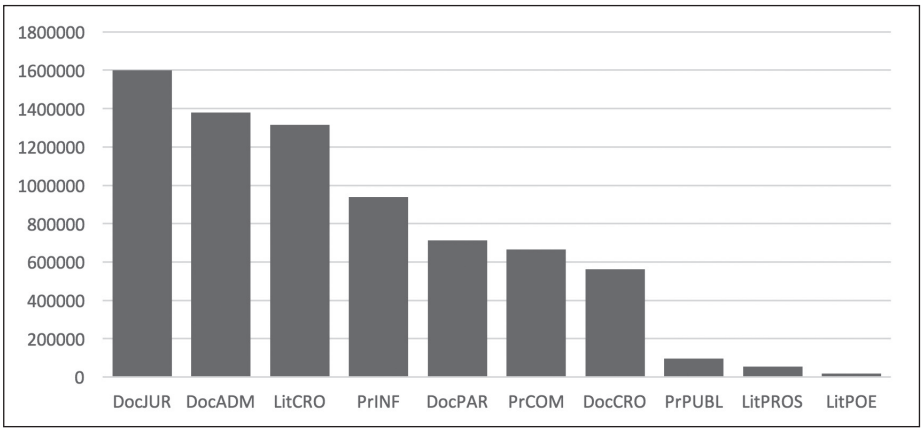

Gráfico 5. CORDIAM: tipos textuales.

CORDIAM no está normalizado, lematizado ni etiquetado, aunque permite el uso de comodines (?, *). En el diseño del corpus se ha tenido un gran interés en la posibilidad de cruzar la búsqueda léxica con metadatos cronológicos, geográficos, textuales y sociales.

Es el único de los corpus analizados que proporciona datos cuantitativos del total de palabras sobre el que se ha realizado una búsqueda, lo cual es fundamental para conocer la frecuencia relativa y la dispersión o distribución de una determinada palabra en el corpus (Brezina, 2018). Por ejemplo, del sustantivo *bastimento(s)* se obtienen 484 casos, distribuidos en 282 documentos (de un total de 10.667), lo que significa que tiene una frecuencia por millón de 66 y un rango de dispersión del 2,6%[12].

La atención en los detalles cuantitativos hace de CORDIAM una herramienta muy potente para el análisis estadístico. Por ejemplo, se pueden obtener datos diacrónicos muy precisos del decrecimiento paulatino (en frecuencia por millón) de la variante *dende* ('desde') entre los siglos XVI y XVIII: del XVI al XVII su uso disminuye en casi un 50%, y del XVII al XVIII se reduce un 98,12%, hasta casi la desaparición.

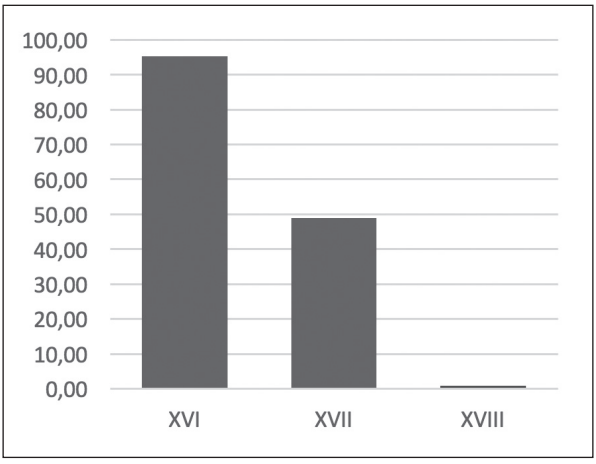

Gráfico 6. Frecuencia de uso de *dende* en CORDIAM.

Otro aspecto destacable de CORDIAM es la posibilidad de ver y descargar en PDF el documento asociado con cada búsqueda, lo que permite ampliar el contexto reducido de la concordancia. Los documentos de CORDIAM no han sido transcritos para el corpus, sino que se han recopilado a partir de trabajos previos

[12] Es decir, está presente en el 2,6% de los textos del corpus.

de diversa procedencia[13]: tesis doctorales y colecciones documentales de distintos países realizadas por historiadores de la lengua. Ofrece una sola edición de cada documento, sin acceso al facsímil. Por último, la gestión informática es de tipo externo.

3.3. El *Corpus Léxico de Inventarios* (*CorLexIn*)

Se trata de un corpus de inventarios de bienes del siglo XVII, donde están representadas la mayor parte de las provincias españolas y algunas regiones americanas[14], con subcorpus de unas 20.000 palabras. En el momento de la consulta, el número total supera las 700.000. La tipología textual tan cerrada tiene la ventaja de la comparabilidad de los datos, tanto interna, es decir, entre las distintas regiones representadas en el corpus, como externa, con otros corpus similares que puedan elaborarse. Además, una posible ampliación del corpus a otros siglos permitiría realizar estudios diacrónicos muy precisos de la historia del léxico cotidiano en español.

El proyecto está diseñado con criterios próximos a los de la geografía lingüística, puesto que, salvando las distancias, "los escribanos de las notarías serán... los dialectólogos encuestadores, los declarantes [y tasadores] harán las veces de informantes y los inventarios conservados en los archivos se convertirán en los cuadernos de encuestas de cada uno de los puntos que componen este particular atlas lingüístico del Siglo de Oro" (Morala, 2012, p. 202).

El corpus está parcialmente normalizado[15] y lematizado. Para hacer búsquedas de formas de palabras debe desactivarse la pestaña de "consulta lematizada", que se presenta por defecto. Todos los documentos han sido transcritos en edición semipaleográfica por el equipo de filólogos del grupo de investigación del *CorLexIn*. No se tiene acceso al facsímil. La gestión del corpus es externa. Los resultados de las consultas no se visualizan en forma de concordancias, sino que se accede a la página completa, en PDF, donde se localizan los ejemplos. Los documentos pueden descargarse página a página.

[13] Una lista completa de la procedencia de los documentos del corpus se encuentra en el apartado "Colaboradores y Referencias" de la página de inicio.

[14] El corpus recoge inventarios de nueve países americanos: Bolivia, Chile, Colombia, Guatemala, México, Panamá, Perú, Puerto Rico y El Salvador.

[15] Por ejemplo, la búsqueda de la forma estándar *lienzo* ofrece los resultados de *lienzo*, *lienso* y *lienço*, pero no así la de *viga*, que arroja tan solo dos ejemplos con "v" y no los de "b".

3.4. *Post Scriptum*: un corpus de cartas privadas de la Edad Moderna

El corpus *Post Scriptum: Archivo digital de escritura cotidiana en Portugal y España en la Edad Moderna* (Rita Marquilhas, CLUL) se compone de dos subcorpus de cartas privadas escritas en la península ibérica, uno portugués y otro español. Cada uno consta de unas 1.500 cartas, aproximadamente un millón de palabras. Abarca toda la edad moderna y el inicio de la contemporánea, es decir, desde el siglo XVI hasta el primer tercio del XIX.

Permite que los documentos se lean en edición paleográfica, crítica y normalizada, además de la comparación con la fotografía del manuscrito. Lo más novedoso de *Post Scriptum* es el hecho de haber realizado las transcripciones en el lenguaje de marcación XML (eXtensible Markup Language), siguiendo el estándar propuesto por el consorcio TEI (http://www.tei-c.org/index.xml) para la codificación de corpus digitales. Esta decisión de partida, muy poco frecuente todavía en el mundo de la lingüística histórica del español (Contreras Seitz, 2009, Marttila, 2014), tiene la ventaja de permitir aprovechar los avances en lingüística computacional que, en su mayoría, requieren el uso de XML. En el caso concreto de *Post Scriptum*, ha hecho posible gestionar el corpus en la plataforma TEITOK (Janssen, 2016).

Los documentos de *Post Scriptum* han sido normalizados, lematizados y etiquetados morfosintácticamente. Esto proporciona enormes posibilidades de recuperación de la información, en la que se pueden cruzar metadatos (lengua, año, lugar de origen, tipo de carta, autor, estatus social, etc.) con consultas léxicas muy complejas. En este aspecto, el buscador permite partir de las siguientes opciones:

1. Grafía original de la palabra (*uino*, *haora*, *reçeui*)
2. Grafía normalizada (*vino*, *ahora*, *recibí*)
3. Lema (*venir*, *ahora*, *recibir*)
4. Etiqueta de anotación morfosintáctica (VMIS3S0, RG, VMIS1S[16]).

La interfaz de consulta, que sigue el protocolo CQP (Corpus Query Protocol) permite búsquedas de palabras simples, pero también combinar estas búsquedas tradicionales con lemas y etiquetas morfosintácticas (POS): por ejemplo, encontrar todos los casos en los que una forma ortográfica (*mujer*) está seguida por cualquier adjetivo ([form = "mujer"] [pos = "A.*"]), o ejemplos de cualquier forma del verbo *haber* seguida de participio ([lemma = "haber"] [pos = "VMP.*"]). Todos los documentos *Post Scriptum* pueden descargarse tanto en formato txt como XML.

[16] TEITOK utiliza una versión del etiquetario EAGLES. En este caso, VMIS3S0 significa 'verbo principal ("main"), indicativo, pasado (S), 3ª persona, singular'; RG designa a los adverbios (R), generales.

3.5. El proyecto *Cíbola*

Cíbola es el nombre de una región mítica situada en el suroeste de los Estados Unidos, de la que se decía estar formada por ciudades bañadas en oro. La leyenda hizo que el virrey Antonio de Mendoza enviara en 1539 una expedición comandada por el fraile franciscano Marcos de Niza a explorar el territorio del actual estado de Nuevo México. El proyecto *Cíbola*, dirigido por Jerry R. Craddock (Berkeley, Universidad de California) tiene como objetivo transcribir, editar y publicar en red toda la documentación relacionada con la conquista y colonización del suroeste de los Estados Unidos, desde el momento inicial hasta el siglo XVIII (Craddock, 2013).

Realmente *Cíbola* no es un corpus digital en sentido estricto, puesto que los documentos se suben al repositorio de la Universidad de Berkeley uno a uno y solo se pueden consultar en PDF e individualmente. No obstante, el volumen documental disponible en la actualidad es de tal envergadura (supera los dos millones de palabras), que merece citarse como uno de los corpus documentales más valiosos del español americano.

Los documentos se transcriben paleográficamente para el proyecto e incluyen la reproducción facsímil de los manuscritos, anexados después de la transcripción. Hasta la fecha se han publicado unos 110 documentos, entre los que se incluyen crónicas y relaciones, documentación diversa de carácter administrativo y juicios.

Los textos cronísticos son narraciones y descripciones de la conquista de los nuevos territorios, de los problemas de la evangelización y las revueltas de indios, o diarios de campaña escritos o dictados por soldados españoles. La documentación administrativa recoge la correspondencia oficial entre gobernadores y reyes, solicitando permisos o ayuda económica y material para hacer frente a los retos de la conquista. Se incluyen también nombramientos, certificaciones, contratos, inspecciones de tropas, equipamiento y armas, juramentos de obediencia y vasallaje, instrucciones de gobierno, etc.

Por último, se han transcrito juicios y declaraciones de testigos sobre delitos diversos: procesos inquisitoriales por supersticiones, conductas judaizantes o blasfemias; juicios contra indios y soldados, por huida, rebelión o asesinato, o contra gobernantes y clérigos.

3.6. Análisis comparativo de los seis corpus

Desde el punto de vista de la tipología textual, cuatro de los corpus estudiados tienen vocación abarcadora: CHARTA, CODEA, CORDIAM y *Cíbola* recogen documentación temáticamente variada. Esta opción permite ofrecer un panorama muy amplio de la lengua, pero exige tener que equilibrar la representación de los

distintos tipos textuales (Torruella, 2016). Por el contrario, *Post Scriptum* y *CorLexIn* se centran en un único tipo textual, lo que sin duda da una visión parcial del conjunto de usos lingüísticos, pero facilita la comparabilidad de los textos, vinculados con la escritura cotidiana (*Post Scriptum*) o con el léxico de la vida material (*CorLexIn*).

Respecto de la cronología, CHARTA y CODEA se plantean como corpus de toda la diacronía del español, desde los orígenes hasta el siglo XIX[17]. CORDIAM, *Post Scriptum* y *Cíbola* se concentran en el español clásico y moderno. El siglo XIX solo es abordado íntegramente en CORDIAM. *CorLexIn* se limita al siglo XVII[18], con algunas incursiones en el XVI y XVIII.

El único corpus panhispánico es CHARTA. CODEA, *Post Scriptum* y *CorLexIn* se circunscriben al español peninsular, aunque este último está incorporando documentación americana. CORDIAM y *Cíbola* son exclusivamente americanos.

Desde el punto de vista computacional hay que destacar que un solo corpus, *Post Scriptum*, se mueve en el entorno de XML-TEI. Gracias a ello y a la combinación con la plataforma TEITOK, consigue que los textos estén totalmente normalizados, lematizados y etiquetados morfológicamente, lo que refuerza y flexibiliza al máximo las posibilidades de recuperación de datos. Todos los otros parten de transcripciones con procesadores de textos en los que no se emplea el lenguaje de marcación XML. Este punto de partida no impide, pero dificulta, la implementación de algunos avances de la lingüística computacional. Como resultado, solo *CorLexIn* está lematizado y parcialmente normalizado y otros (CODEA 2020, CHARTA) se encuentran en fase de renovación técnica, anunciada para próximas versiones.

El empeño de la moderna filología por ofrecer ediciones rigurosas y fiables está presente en los seis corpus y en CHARTA, CODEA, *Post Scriptum* y *Cíbola* encuentra su reflejo en la posibilidad de confrontar las transcripciones con el facsímil disponible en línea. Por otra parte, la documentación está transcrita expresamente para cada uno de los corpus, siguiendo criterios de transcripción paleográfica bastante rigurosos. Solo CORDIAM utiliza documentación previa, aunque procedente de trabajos realizados por historiadores de la lengua.

Por último, únicamente *Post Scriptum* permite que los filólogos controlen todas las fases del proceso de compilación, desde la selección y transcripción documental hasta la lematización y etiquetado, incluyendo la visualización de los documentos y su revisión instantánea en la web. En todos los demás casos, la gestión computacional del corpus recae en expertos externos. En el caso de *Cíbola* esta cuestión no es analizable, puesto que su objetivo no es compilar un corpus *stricto sensu*, sino ofrecer en línea documentos independientes en PDF.

[17] CODEA 2020 tiene previsto compilar documentación hasta 1900.

[18] El grueso de la documentación de *Cíbola* es también de finales del XVI y sobre todo del XVII.

Tabla II. Comparación entre los seis corpus.

	CHARTA	CODEA 2015	CORDIAM	CorLexIn	Post Scriptum	Cíbola
Cronología	1200-1800	1200-1800	1494-1905	1600-1700	1517-1833	1539-1748
Geografía	Panhispánica	España	América	España. Partes de América	España Portugal	Suroeste EE. UU.
Tipología	Variada	Variada	Variada	Inventarios	Cartas privadas	Variada
XML	No	No	No	No	Sí	No
Normalización	No	No	No	Parcial	Sí	No
Lematización	No	No	No	Sí	Sí	No
Etiquetado (POS)	No	No	No	No	Sí	No
Facsímil	Sí	Sí	No	No	Sí	Sí
Varias ediciones	Sí	Sí	No	No	Sí	No
Guardar y ver doc. completo	Sí	Sí	Sí	Sí	Sí	Sí
Gestión informática	Externa	Externa	Externa	Externa	Autónoma	-
Palabras	1.346.094	1.445.162	7.341.245	738.160	2.000.000	2.000.000

4. VENTAJAS DE XML-TEI: SU APLICACIÓN EN *ORALIA DIACRÓNICA DEL ESPAÑOL* (ODE)

El análisis de estos seis corpus ha servido para tomar decisiones sobre cómo mejorar el antiguo CORDEREGRA, que había llegado a un punto muerto en el que los usuarios debían conformarse con las escuetas opciones de un buscador básico de texto en PDF. Parecía claro que el nuevo corpus (ODE) debía mantener los criterios de edición de la red CHARTA, con visibilidad del manuscrito original; que debía poseer una interfaz de consulta de fácil manejo, como la de CORDIAM, en la que además se informara del total de palabras sobre el que se hacía la búsqueda; parecía también importante limitar la tipología textual (en este caso a inventarios de bienes y declaraciones de testigos) para facilitar la comparabilidad y la representatividad estadística de los datos; y por último, era necesario disponer de un buscador tan avanzado como el de *Post Scriptum*, con la opción, además, de controlar todas las fases del proceso, desde la transcripción hasta la visualización final.

Para conseguir todo esto, el modelo tecnológico más completo lo proporciona *Post Scriptum*, que parte de XML-TEI en el entorno de la plataforma TEITOK. Esto significa, como punto de partida, tener que convertir todas las transcripciones originales de CORDEREGRA (que se habían hecho en *Microsoft-Word*) a las actuales de ODE en XML.

La reticencia a utilizar XML-TEI en la compilación de corpus históricos no se da solo en el caso del español, donde solo uno de los seis corpus analizados emplea este estándar de etiquetado, sino que parece ser una constante general observable en otras lenguas (Marttila, 2014). Sin embargo, la capacidad de encontrar información en un corpus depende muy estrechamente de cómo haya sido anotado (Kytö, 2011). Y en este aspecto, XML-TEI se revela como la mejor estrategia de partida, además de la más estandarizada. Su uso permite estructurar los textos y etiquetarlos con instrucciones de procesamiento informático interpretables unívocamente por cualquier ordenador.

4.1. Estructurar textos: información textual y metatextual

Los textos de CORDEREGRA no están estructurados desde el punto de vista computacional, lo que ocasiona problemas a la hora de realizar consultas. El siguiente fragmento reproduce algunas convenciones originales de las transcripciones del corpus:

> 1554, Málaga
> ARCHGR 1511/4
> Petición para que el pescado se venda al peso y no a ojo y para que se pesen las asaduras y **turmas** de los carneros
>
> Así mysmo suplicamos a v*uest*ra señoría que porque el balor / de las carnes es grande y que no se puede conpadesçer, mande / que se pese las asaduras y **turmas** de los carneros y las / asaduras de las otras carnes eçebto de la baca. / Porque pesandose asaduras y **turmas** del carnero / baxaria por libra quatro m*aravedís* e asi nos ofresçemos que abra quien / lo abaxe del presçio questubiere; y este beneffiçio es general / y que todos partiçipan del y de uenderse las asaduras y **turmas** / como agora se benden no pueden gozar ni gozan todos, sino sola/mente las personas poderosas y priminentes, y los pobres / no las pueden auer, y aunque se vbiese de dar a la gente pobre / no los pueden aver todos porque son pocas y de la baxa de la / d*ic*ha carne gozarian todos ygualmente. / Y si se dixese que, pesandose las asaduras, se les daria la / mayor parte a los pobres, en tal caso v*uest*ra señoria puede / mandar debaxo de pena a los cortadores que la repartan // [6v, 5940] ygualmente y, aunque al pobre se le diese algo dellas no se le haze **a/grauio** porques carne sin gueso y le compliria mas que la otra carne.

A simple vista se distinguen los metadatos de la cabecera, por un lado, de la transcripción propiamente dicha del manuscrito, por otro. En esta segunda parte no es difícil identificar algunas marcas utilizadas en CORDEREGRA: la barra oblicua (/) para separar líneas, la doble barra (//) para indicar cambio de página, la numeración del nuevo folio [6v], la indicación del número de fotografía [5940] con el que se vincula la página, el empleo de la cursiva para indicar expansiones de abreviaturas (v*uest*ra, m*aravedí*s), etc.

El problema surge cuando se quiere realizar en un texto etiquetado de esta forma algunas de las operaciones siguientes:

1. Localizar todos los ejemplos textuales de *turmas* 'testículos'.
2. Localizar todos los ejemplos de *turmas* en documentos escritos en la segunda mitad del siglo XVI.
3. Encontrar todos los ejemplos de la forma ortográfica *agrauio*.
4. Encontrar todas variantes de la forma normalizada *agravio*.

Ninguna de las convenciones adoptadas, intuitivas para la mente humana, son interpretables para un ordenador. Por consiguiente, cuando se buscan en PDF los casos de *turmas*, la consulta ofrece cuatro resultados (destacados en negrita), al no poder separarse los datos de la cabecera (achacables al editor) de los estrictamente textuales (achacables al autor del texto).

Por otra parte, tampoco se pueden aislar los casos de *turmas* procedentes de documentos escritos entre determinadas fechas o en ciertas localidades, puesto que tanto la fecha (*1554*) como el lugar de escritura (*Málaga*) son simples cadenas de caracteres indiferenciadas del resto.

Por el contrario, los textos etiquetados en XML se estructuran como si fueran una base de datos. La estructuración más básica sirve para separar la información metatextual (o contextual: lugar, fecha, autor, archivo de procedencia, signatura, etc.) de la textual (la transcripción del documento). Para ello, todos los metadatos se incluyen entre las etiquetas de apertura y cierre de la cabecera:

<teiHeader></teiHeader>

Por su parte, el cuerpo del texto se escribe dentro de las etiquetas <text></text>. Una vez que se ha hecho esto, la orden de procesamiento se limita a localizar todos los casos de *turmas* contenidos entre las etiquetas <text></text>, como si le pidiéramos a un texto tabulado en *Excel* que buscara ejemplos solo en la columna "text". De esta forma, el resultado final de la consulta arrojaría tres ejemplos, y no los cuatro del texto sin etiquetar.

La lógica es la misma a la hora de cruzar metadatos con búsquedas textuales. La solución viene también dada mediante la estructuración que proporciona el

etiquetado XML. En el caso particular de ODE, siguiendo las directrices de TEI para corpus históricos, los metadatos de lugar y fecha de escritura se organizan como se muestra en la Figura 1:

```
<profileDesc>
    <settingDesc>
      <setting>
        <name type="place">Málaga</name>
        <date when="1554">1554</date>
      </setting>
    </settingDesc>
</profileDesc>
```

Figura 1. Cabecera XML para lugar y fecha en ODE, elaborada por Gael Vaamonde.

Desde el punto de vista técnico, esta jerarquización de las etiquetas equivale a una instrucción del tipo 'selecciona la fecha del documento que aparece en el elemento <date> que está contenido en <setting>, dentro de <settingDesc>, que a su vez pertenece a la categoría <profileDesc>'. En TEI, *profileDesc* proporciona una descripción detallada de los aspectos contextuales en que tiene lugar un determinado intercambio comunicativo, en este ejemplo concreto, el año (y lugar) donde fue escrito el documento.

4.2. La información paratextual: XML y las instrucciones de procesamiento informático

Cuando se elabora un corpus histórico no basta con diferenciar lo textual de lo contextual. En muchos casos interesa conservar información paratextual, vinculada con aspectos visuales o codicológicos de los manuscritos, como la disposición de las líneas (es decir, dónde empieza y termina cada línea en el manuscrito original), la numeración de los folios, los posibles deterioros del soporte, las intervenciones en el texto (tachaduras, añadidos entre líneas o al margen, cambios de mano), etc.

En las transcripciones de CORDEREGRA, como se ha visto, se utilizaba la barra (/) para indicar cambio de línea. Este sistema, aunque comprensible para los lectores, no es interpretable informáticamente. Además, provoca un problema añadido relacionado con la recuperación de información textual. En el caso de

querer obtener los ejemplos de *agrauio*, la marca utilizada para indicar cambio de línea impide que el buscador reconozca *a/grauio* como ejemplo.

Este problema también se puede resolver mediante el etiquetado en XML-TEI, en concreto marcando la transcripción como "a<lb break=”no”/>grauiuo", es decir con una orden de procesamiento (lb 'line begining') que se interpreta como 'considera *agrauio* como un solo *token*, pero a la hora de visualizar el texto separa la palabra en dos líneas distintas'. Como se ve, este etiquetado diferencia entre distintas capas o niveles de un texto, el de la visualización, por un lado, y el de la interpretación textual, por otro. TEI dispone de etiquetas unívocas y específicas para cada aspecto codicológico que se desee conservar. En ODE se sigue el modelo adoptado por *Post Scriptum* (Vaamonde, 2016).

4.3. Conectar formas paleográficas y normalizadas

El ejemplo de *agrauio* plantea otro de los problemas fundamentales de la lingüística de corpus históricos, relacionado con la enorme vacilación ortográfica. En el texto seleccionado la variación es reducida (*agravio*/*agrauio*), pero en otros casos puede ser bastante caótica: *escreuiendo*, *escrebiendo*, *escriviendo*, *escriujendo*, etc. Incluso, como se indicó, las variantes formales pueden llegar a ser difícilmente previsibles: *bilma* 'bizma'[19], *aijada* 'aguijada', *treves* 'trébedes', etc.

Un buen corpus es aquel que ofrece más posibilidades de recuperación de datos, de la manera más fácil, flexible y exhaustiva. Con respecto a la diversidad ortográfica, la recuperación más sencilla y completa vendrá dada por la posibilidad de que los usuarios inicien su consulta a partir de la forma normalizada, esto es, del estándar moderno que conocen y figura en el diccionario (*escribiendo*, *bizma*, *aguijada*, *trébedes*). Para conseguir esto, el corpus tiene que estar normalizado.

En CORDEREGRA no fue posible resolver este problema: la edición paleográfica (*agrauio*) y la normalizada (*agravio*) se almacenaban en archivos diferentes, no conectados entre sí, que se manejaban y recuperaban por separado. Para salir del atolladero al que lleva la desconexión entre ediciones, TEITOK recurre a la sintaxis de XML, por medio de la cual se logra que la forma ortográfica (form=“escreuiendo”), la forma normalizada (nform= “escribiendo”), la forma lematizada (lemma=“escribir”) y la etiqueta morfosintáctica (pos=“VMG0000”)[20] estén vinculadas entre sí como partes de un macroelemento <tok></tok>:

[19] Emplasto curativo.

[20] Para el funcionamiento de las etiquetas EAGLE en TEITOK, véase Vaamonde y Magro (2016).

<tok id="w-2" nform="escribiendo" lemma="escribir"
pos="VMG0000">escreuiendo</tok>

Para ello, todos los documentos de ODE pasan por cuatro fases consecutivas de codificación: tokenización, normalización, lematización y etiquetado (Vaamonde, 2015). En la primera fase se genera automáticamente un elemento *tok* para cada forma ortográfica (*token*) del corpus. Obsérvese que la forma original del manuscrito (*agrauio*, *escreuiendo*, etc.) es el contenido del elemento <tok></tok>. Ese contenido se irá relacionando con los valores de distintos atributos que se añaden en el interior de la etiqueta de apertura (<tok>). El primer atributo (id) asigna un número a cada *token*, que permite identificarlo como único dentro del corpus:

<tok id="w-12" >agrauio</tok>
<tok id="w-2">escreuiendo</tok>
<tok id="w-24">treves</tok>

Cada nueva fase genera un nuevo atributo con su valor correspondiente. Por tanto, cuando se normaliza el corpus, se crea un nuevo atributo (nform) cuyo valor es el de la forma normalizada (nform="trébedes", nform="agravio", etc.):

<tok id="w-12" nform="agravio">agrauio</tok>
<tok id="w-2" nform="escribiendo">escreuiendo</tok>
<tok id="w-24" nform="trébedes">treves</tok>

A través de este sistema quedan vinculados *agrauio* con *agravio*, *escreuiendo* con *escribiendo* y *treves* con *trébedes*. Con ello se consigue que la búsqueda de palabras pueda partir de la forma original ([form="escreuiendo"]) o de la normalizada ([nform="escribiendo"]). Después de normalizar el corpus se procede a su lematizado. Al hacer esta operación, el elemento tok se enriquece con un nuevo atributo (lemma):

<tok id="w-2" nform="escribiendo" lemma="escribir">escreuiendo</tok>

La última fase consiste en el etiquetado automático, que incorpora un nuevo atributo (pos):

<tok id="w-2" nform="escribiendo" lemma="escribir"
pos="VMG0000">escreuiendo</tok>

Todos estos procesos se realizan automáticamente, pero pueden revisarse de forma manual en la plantilla que TEITOK proporciona para cada *token*:

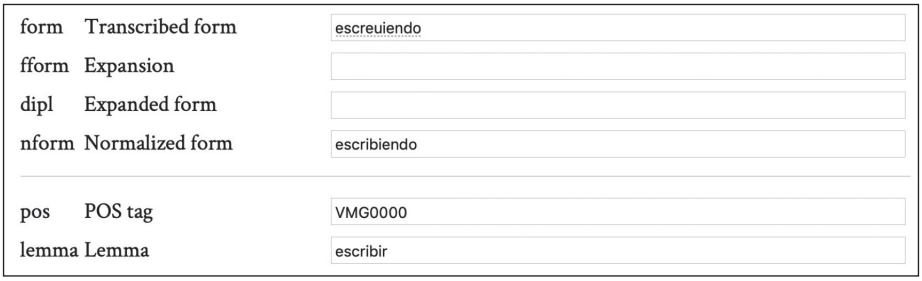

Figura 2. Plantilla de modificación de datos en TEITOK.

Cuando concluyen estas cuatro fases y la revisión manual, se dispone de un corpus normalizado, lematizado y etiquetado, que permite recuperar cualquier tipo de información lingüística, desde las formas originales de los manuscritos hasta estructuras gramaticales o léxicas más complejas y abstractas.

5. CONSIDERACIONES FINALES

El número total de palabras de los seis corpus analizados, para el periodo comprendido entre los siglos XVI y XIX, alcanza los doce millones. Esto quiere decir que en los últimos años se ha hecho un esfuerzo filológico muy notable en la constitución de una sólida infraestructura de investigación de carácter histórico-lingüístico. El valor de esta documentación se acrecienta si se tiene en cuenta que se trata de transcripciones que respetan escrupulosamente los usos gráficos de los manuscritos.

Sin embargo, sigue habiendo algunos aspectos mejorables. Uno de ellos se relaciona con la representatividad de los datos, puesto que quedan áreas geográficas (véase gráfico 3), tipologías textuales, especialmente las más próximas a la oralidad (véanse gráficos 2 y 5) y periodos (significativamente, el siglo XIX), con poca presencia en los corpus. En la medida en que se alcance un mayor equilibrio, los resultados de la comparación serán más valiosos, tanto cuantitativa como cualitativamente.

Otro aspecto donde caben mejoras es en el de las posibilidades de recuperación de la información, que dependen bastante de cómo se haya etiquetado el corpus. En este sentido, el modelo tecnológico predominante se aleja del estándar XML-TEI, seguramente como consecuencia de la separación tradicional entre estudios humanísticos y filológicos, por un lado, y enfoques técnicos y computacionales, por otro. El desarrollo de plataformas como TEITOK, utilizada con éxito en *Post Scriptum* y en vías de aplicación en ODE, puede significar un acercamiento muy productivo entre estas dos áreas.

REFERENCIAS

Academia Mexicana de la Lengua [en línea]. *Corpus diacrónico y diatópico del español de América (CORDIAM)*. Disponible en http://www.cordiam.org/. [Consulta: 3/04/2019].

Bertolotti, Virginia y Company Company, Concepción. (2014). El corpus diacrónico y diatópico del español de América (CORDIAM). Propuesta de tipología textual. *Cuadernos de la ALFAL*, 6, 130-149.

Brezina, Vaclav. (2018). *Statistics in Corpus Linguistics. A practical guide*. Cambridge, Reino Unido: Cambridge University Press.

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

Calderón Campos, Miguel y García-Godoy, María Teresa (dirs.) [en línea]. *Oralia diacrónica del español (ODE)*. Disponible en http://corpora.ugr.es/ode. [Consulta: 12/04/2019].

CHARTA. *Corpus hispánico y americano en la red: textos antiguos*. Disponible en http://www.corpuscharta.es/. [Consulta: 11/04/2019].

CHARTA. (2013). *Criterios de edición de documentos hispánicos (orígenes-siglo XIX) de la red internacional CHARTA*. Disponible en https://www.redcharta.es/criterios-de-edicion/. [Consulta: 17/04/2019].

CLUL. (2014). *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. Disponible en http://ps.clul.ul.pt. [Consulta: 10/04/2019].

Contreras Seitz, Manuel. (2009). Hacia la constitución de un corpus diacrónico del español de Chile. *RLA. Revista de Lingüística Teórica y Aplicada*. 47(2), pp. 111-134.

Craddock, Jerry. (2013). The Cibola Project: A Brief Historical Account. *Romance Philology*. 67(2), pp. 247-257.

Craddock, Jerry (dir.) [en línea]. *Cibola Project*. Disponible en https://escholarship.org/uc/rcrs_ias_ucb_cibola. [Consulta: 7/04/2019].

Grupo de Investigación Textos para la Historia del Español (GITHE). *CODEA + 2015 (Corpus de documentos españoles anteriores a 1800).* [en línea]. Disponible en http://corpuscodea.es/. [Consulta: 1/04/2019].

Janssen, Maarten (2016). TEITOK: Text-faithful annotated corpora. En *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pp. 4037-4043. Disponible en http://www.lrec-conf.org/proceedings/lrec2016/authors. html#D. [Consulta: 19/04/2019].

Kytö, Merja. (2011). Corpora and historical linguistics. *Revista Brasileira de Linguística Aplicada*. 11(2), pp. 417-457.

Marttila, Ville. (2014). Creating Digital Editions for Corpus Linguistics. The case of Potage Dyvers, a family of six Middle English recipe collections. Tesis doc-

toral. Helsinky: University of Helsinky.

Miguel Franco, Ruth y Sánchez-Prieto Borja, Pedro. (2016). CODEA: A "Primary" Corpus of Spanish Historical Documents. *Variants. The Journal of the European Society for Textual Scholarship*. 12-13, pp. 211-228.

Morala, José. (2012). Léxico e inventarios de bienes en los siglos de oro. En Clavería, Freixas, Prat y Torruella (eds.). *Historia del léxico: perspectivas de investigación*. Madrid-Frankfurt: Iberoamericana-Vervuert, pp. 199-218.

Morala, José (dir.) [en línea]. *Corpus léxico de inventarios (CorLexIn)*. Disponible en http://web.frl.es/CORLEXIN.html. [Consulta: 15/04/2019].

Sánchez, Pedro. (2012). La red CHARTA: proyecto global de edición de documentos hispánicos. En Torrens y Sánchez (eds.). *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*. Berna: Peter Lang, pp. 17-44.

Sánchez, Pedro. (2018). El corpus ALDICAM-CM. Geografía lingüística diacrónica de la Comunidad de Madrid. *CHIMERA. Romance Corpora and Linguistic Studies*. 5(1), pp. 69-75.

Sánchez, Pedro y Delfina Vázquez. (2017). Hacia un corpus de beneficencia en Madrid (siglos XVI-XIX). *Scriptum Digital*. 6, pp. 83-103.

Torruella, Joan. (2016). Tres propuestas en el ámbito de la lingüística de corpus. En Kabatek (ed.) y Moreno (col.). *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín/Boston: De Gruyter, pp. 90-113.

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

Vaamonde, Gael. (2015). P.S. Post Scriptum. Dos corpus diacrónicos de escritura cotidiana. *Procesamiento del lenguaje natural*. 55, pp. 57-64.

Vaamonde, Gael. (2016). *Guía para la edición digital de textos en P.S. Post Scriptum.* Lisboa: Centro de Linguística da Universidade de Lisboa. Disponible en http://ps.clul.ul.pt/files/Manual_PS.pdf. [Consulta: 17/04/2019].

Vaamonde, Gael y Magro, Catarina. (2016). *Manual de Edición en P.S. Post Scriptum. Edición modernizada, Anotación POS, Anotación Sintáctica*. Lisboa: Centro de Linguística da Universidade de Lisboa. Disponible en http://ps.clul.ul.pt/files/Manual_Mod_Pos_Sin.pdf. [Consulta: 17/04/2019].