Dialectologia 31 *(2023), 105-126.*
ISSN: 2013-2247

Received 11 October 2021.

Accepted 3 January 2022.

Published 30 July 2023.

DOI: 10.1344/DIALECTOLOGIA2023.31.5

CORPUS DE TEXTOS NOTARIALES EXTREMEÑOS (CORTENEX S. XVII). LA EDICIÓN DE UN CORPUS HISTÓRICO-LINGÜÍSTICO EN EL ÁMBITO DE LAS HUMANIDADES DIGITALES[1]

Inmaculada GONZÁLEZ SOPEÑA[*]

Universidad de Granada

isopena@ugr.es

ORCID: 0000-0002-0439-8787

Resumen

El presente artículo se centra en la metodología seguida en la elaboración de un corpus de documentación notarial extremeña del siglo XVII (CORTENEX) siguiendo las propuestas del consorcio TEI en lo referente a la codificación y etiquetado de documentos históricos. Todo ello se basa en el uso de lenguaje marcado XML en las transcripciones y en el procesamiento lingüístico de los textos en la plataforma digital TEITOK a través de cuatro fases fundamentales: tokenización, normalización, lematización y anotado morfosintáctico. Este corpus se conforma con un subcorpus de *Oralia diacrónica del español* (ODE). Actualmente, CORTENEX ya cuenta con documentación accesible y, por el tipo de textos que incluye, su interés fundamental reside en analizar la variación léxica del español desarrollado en el territorio que se corresponde con la Comunidad Autónoma de Extremadura. Esta variedad carece prácticamente de estudios de corte diacrónico que permitan analizar la lengua de aquella región en perspectiva histórica.

Palabras clave: documentación notarial, lingüística de corpus, XML, TEITOK, historia del léxico español

[1] El presente artículo ha sido realizado dentro del marco del proyecto "Atlas Lingüístico y Etnográfico de Andalucía, S. XVIII. Patrimonio documental y humanidades digitales" (Proyectos I+D+i Junta de Andalucía-FEDER, P18-FR-695).

[*] Campus de la Cartuja Universidad de Granada, Calle del Prof. Clavera, s/n, Departamento de Lengua Española, Facultad de Filosofía y Letras de la Universidad de Granada.

CORPUS DE TEXTOS NOTARIALS EXTREMENYS (CORTENEX S. XVII). L'EDICIÓ D'UN CORPUS HISTORICOLINGÜÍSTIC A EN ÀMBIT DE LES HUMANITATS DIGITALS

Resum

Aquest article se centra en la metodologia utilitzada en l'elaboració d'un corpus de documentació notarial extremenya del segle XVII (CORTENEX) que segueix les propostes del consorci TEI pel que fa a la codificació i l'etiquetatge de documents històrics. Es basa en l'ús de llenguatge marcat XML en les transcripcions i en el processament lingüístic dels textos en la plataforma digital TEITOK a través de quatre fases fonamentals: tokenització, normalització, lematització i anotació morfosintàctica. Aquest corpus es conforma amb un subcorpus d'*Oralia diacrònica del español* (ODE). Actualment, CORTENEX ja compta amb documentació accessible i, pel tipus de textos que inclou, el seu interès fonamental radica a analitzar la variació lèxica de l'espanyol usat en el territori que es correspon amb la Comunitat Autònoma d'Extremadura. Aquesta varietat no té pràcticament estudis de tall diacrònic que permetin analitzar la llengua d'aquella regió des d'una perspectiva històrica.

Paraules clau: documentació notarial, lingüística de corpus, XML, TEITOK, història del lèxic espanyol

CORPORA OF NOTARIAL TEXTS FROM EXTREMADURA (CORTENEX S. XVII): EDITING HISTORICAL LINGUISTIC CORPUS IN THE FIELD OF DIGITAL HUMANITIES

Abstract

This article focuses on the methodology followed in the preparation of a corpus of notarial documentation from Extremadura during the seventeenth century (CORTENEX), thanks to the proposals of the TEI consortium regarding the coding and labeling of historical documents. This is based on the use of XML markup language in the transcriptions and in the linguistic processing of the texts in the TEITOK digital platform through four fundamental phases: tokenization, normalization, stemming and morphosyntactic annotation. CORTENEX is a subcorpus of *Oralia diacrónica del español* (ODE). Currently, CORTENEX already has accessible documentation, and, due to the type of texts that it includes, its fundamental interest lies in analyzing the lexical variation of the Spanish developed in the territory that corresponds to Extremadura. This variety practically lacks diachronic studies that allow analyze the language of that region from a historical perspective.

Palabras clave: notarial documentation, corpus linguistics, XML, TEITOK, history of Spanish lexicon

1. Introducción

Este trabajo se centra en el análisis de las directrices metodológicas seguidas en la elaboración de un corpus diacrónico de documentos notariales extremeños digitalizado, CORTENEX (S. XVII). Los documentos notariales seleccionados se

concretan en inventarios de bienes, cartas de dote y testamentos de la primera mitad del siglo XVII procedentes del Archivo Histórico Provincial de Badajoz[2] y de Cáceres.[3]

La elaboración de corpus documentales basados en textos de archivo es una práctica cada vez más extendida en el estudio de las lenguas. En lengua española son numerosos los corpus diacrónicos que recopilan documentación archivística de distinto tipo (cartas, textos jurídicos, religiosos, notariales, administrativos, etc.) y de diferentes épocas y territorios desde los orígenes del español. Algunos ejemplos de ello son: CHARTA, CORDIAM, *Cíbola*, ODE, *Post Scriptum*, entre muchos otros. Este tipo de documentación resulta muy fiable a la hora de plantear estudios de variación diatópica y diacrónica. Cabe destacar aquí, por las similitudes que presenta con respecto a CORTENEX, el *Corpus Léxico de Inventarios* (Morala Rodríguez 2014),[4] dado que es un corpus confeccionado a partir de inventarios de bienes del siglo XVII de España y América. Dichas fuentes documentales permiten atestiguar voces fuera del registro general, términos no documentados, variantes diatópicas, préstamos, entre otras cuestiones (Morala Rodríguez 2014). Estos textos son relaciones de bienes que incluyen un caudal léxico valioso vinculado a diferentes ámbitos de la vida cotidiana de gran valor para la historia del léxico español.

Desde el punto de vista lingüístico, el estudio de la variedad de español que se desarrolla en los territorios de la actual Extremadura en perspectiva diacrónica sigue siendo una tarea pendiente (Ariza 1985). A este vacío, se suma la falta de un atlas lingüístico completo de Extremadura,[5] que solo se ve suplido por un centenar de monografías y estudios de corte dialectal y sincrónico sobre el habla de diversas localidades. Se carece, por tanto, de más estudios histórico-lingüísticos apoyados en bases documentales de textos archivísticos de diferente tipo y de diferentes épocas

[2] Este archivo cuenta con un fondo de Protocolos Notariales de 5253 legajos datados desde 1523 hasta 1917. En línea en:
<http://archivosextremadura.gobex.es/WAREX/live/SistemaArchivistico/JuntaExtremaduraSA/ArchivosHistoricoProvincialesSA/ArchivosHistoricoProvincialesBA.html>

[3] El fondo documental de Protocolos Notariales de Cáceres cuenta con un total de 5891 legajos cuyas fechas oscilan entre 1514 y 1988. En línea en:
<http://archivosextremadura.gobex.es/WAREX/live/SistemaArchivistico/JuntaExtremaduraSA/ArchivosHistoricoProvincialesSA/ArchivosHistoricoProvincialesCC/Fondosdocumentaless/fondosisadg.html>

[4] <https://webfrl.rae.es/CORLEXIN.html>

[5] Puede consultarse la *Cartografía Extremeña* en <http://www.geolectos.com/>

que den cuenta de la evolución de fenómenos lingüísticos característicos de la variedad de español que se desarrolla en estos territorios: rastrear la presencia de léxico dialectal, presencia de confusión entre sibilantes, presencia de leonesismos, de portuguesismos, de castellanismos, de andalucismos o realizar estudios comparativos con otros corpus.

Desde el punto de vista tecnológico, la confección de CORTENEX S. XVII se cimenta en la metodología ensayada con éxito por el proyecto *Post Scriptum*[6] para la elaboración de un corpus de cartas privadas españolas y portuguesas de la Edad Moderna (Vaamonde 2015). El procesamiento lingüístico de todo ese volumen documental fue llevado a cabo a través de nuevas tecnologías en el ámbito de las humanidades digitales. De tal modo, la transcripción de los documentos en CORTENEX se hace en lenguaje marcado XML (*eXtensible Markup Language*) siguiendo los estándares de codificación lingüística propuestos por el consorcio TEI (*Text Encoding Initiative*),[7] para su presentación digital en la plataforma TEITOK (Janssen 2016) como subcorpus de *Oralia diacrónica del español* (ODE).

La plataforma TEITOK permite procesar los datos lingüísticos previamente marcados en las transcripciones a través de cuatros fases: tokenización, normalización, lematización y anotación morfosintáctica. Estas cuatro fases son el punto principal que se va a desarrollar en los sucesivos apartados aplicados a CORTENEX. El resultado se traduce en un recurso de acceso libre que permite visualizar una edición semipaleográfica de los textos con sus facsímiles y una edición crítica digitalizada de ellos. Al mismo tiempo, el corpus generado cuenta con un potente motor de búsquedas (*CQL Protocol*) para su explotación filológica.

En este trabajo, se parte de una revisión sobre la bibliografía disponible para el estudio del extremeño. Seguidamente, tras constatar las lagunas que existen en el estudio diacrónico de esa variedad, se describe paso a paso toda la metodología seguida en la elaboración del corpus CORTENEX (S. XVII) a través del uso de nuevas herramientas digitales. Finalmente, se establecen las conclusiones principales que se derivan de todo el estudio.

[6] Disponible en: <http://ps.clul.ul.pt/es/index.php?action=cqp&act=advanced>
[7] TEI incluye un estándar específico para la codificación de manuscritos.

2. Estudios lingüísticos sobre la variedad de español en Extremadura

El interés por la variedad de español que se ha desarrollado en la actual Extremadura se remonta a finales del siglo XIX (Fernández de Molina 2014: 6) con el nacimiento de la revista *El Folklore Frexnenses* (1882-1889) y de la *Revista de Extremadura* (1899). Desde entonces, las publicaciones y estudios sobre ella se centran, sobre todo, en el habla y en aspectos etnolingüísticos y culturales. A lo largo del siglo XX han visto la luz algunas monografías centradas en el habla de diversas localidades, como la de Zamora Vicente (1943), Lorenzo Criado (1948), Velo Nieto (1956), Cummins (1974) o Barros García (1974), quienes han tratado aspectos sobre el habla de Mérida, las Hurdes, Coria o el Arroyo de San Serván, entre otros municipios.

La monografía colectiva de Viudas, Ariza & Plans (1987) o la de Montero Curiel (2006), así como el capítulo que García Mouton (1996) dedica al habla extremeña, analizan la historia lingüística de Extremadura desde una perspectiva generalista. En ellas se establecen las fuentes lingüísticas sobre las que esta variedad se asienta desde el punto de vista léxico (leonesismos, arcaísmos, portuguesismos) junto con algunas notas sobre su fonética, su morfología y su sintaxis. Asimismo, se ha prestado atención al estudio de la conciencia lingüística tan negativa que tienen los extremeños (Palacios Martín 1988, González Salgado 2009). Estudios adicionales han puesto el foco en la franja fronteriza entre Extremadura y Portugal, donde se constata la pervivencia de una antigua lengua romance conocida como *fala*, localizada en San Martín de Trevejo, Eljas y Valverde del Fresno (Gargallo Gil 1999).

A principios del siglo XXI, se publica la *Cartografía Extremeña* (González Salgado 2003) que intenta imitar al *Atlas Lingüísticos y Etnográfico de Andalucía*, si bien de forma incompleta. Como puede observarse, el interés por la variedad extremeña de español se ha centrado en aspectos dialectales, etnológicos y de habla. A partir de los noventa se comenzó a insistir en la importancia de realizar estudios diacrónicos sobre esta variedad. Ejemplos de ello se encuentran en Flores Manzano (1988) o en el estudio realizado por Marcos Álvarez (1992) donde se analiza el léxico de la

indumentaria en el Badajoz de los siglos XVI y XVII a través de inventarios de bienes[8]. Este último autor llama la atención sobre la falta de estudios histórico-lingüísticos para la variedad extremeña de español: "Estas observaciones locales, aún inéditas en la historiografía extremeña, reclaman un tratamiento interdisciplinar riguroso donde la lingüística diacrónica tendría que aportar, sin duda, un cúmulo de observaciones aprovechables" (Marcos Álvarez 1992: 1162). A estos estudios lingüísticos mínimos apoyados en documentación antigua, cabe añadir el volumen de textos notariales que el *CorLexIn* incluye para la comunidad extremeña.

Tras este breve panorama esbozado, ha sido la escasez de estudios histórico-lingüísticos, así como la carencia de un corpus específico con documentación archivística geográficamente adscrita a Extremadura, la que ha propiciado la elaboración de CORTENEX (S. XVII) a partir de documentación inédita de corte notarial vinculada a las provincias Badajoz y Cáceres. Este tipo de documentación es rica en la descripción del léxico de la vida cotidiana del siglo XVII y contribuye a la recopilación de datos lingüísticos sobre los que cimentar estudios histórico-lingüísticos de corte eminentemente léxico sobre la historia del español en Extremadura en diferentes ámbitos de la vida cotidiana (vestimenta, joyas, enseres de cocina, aperos de labranza, fauna, toponimia).

3. CORTENEX y ODE. Antecedentes

CORTENEX (S. XVII) se erige como un subcorpus integrado en el corpus *Oralia diacrónica del español* (ODE),[9] desarrollado en la Universidad de Granada. La elaboración del corpus ODE se fundamenta en el uso de la herramienta tecnológica TEITOK, la cual permite ofrecer un corpus totalmente digital y anotado. Este corpus tuvo origen en el proyecto *Corpus diacrónico del español del Reino de Granada, 1492-1833*, CORDEREGRA (Calderón Campos 2015), y está compuesto por documentación

[8] A este estudio, se suma la reciente tesis doctoral de Sánchez Sierra (2019), quien recopila un volumen importante de textos notariales de diversos puntos de la comunidad de Extremadura.
[9] <http://corpora.ugr.es/ode>.

notarial y jurídica de las provincias andaluzas de Málaga, Granada y Almería desde finales del siglo XV hasta el siglo XIX.

Actualmente, el corpus ODE ya cuenta con un volumen importante de textos que pueden ser consultados. Además, los proyectos de investigación HISPATESD[10] y ALEA-XVIII[11] han permitido continuar con la digitalización y estudio de los documentos de CORDEREGRA, así como impulsar la ampliación del corpus con documentación notarial de otras provincias andaluzas (Huelva, Sevilla, Cádiz) con el objetivo de analizar la variación diatópica y diacrónica del español que allí se desarrolla. Adicionalmente, ODE integra diversos subcorpus a modo de corpus de control [12] para establecer comparaciones y búsquedas cruzadas que permitan analizar las diferencias y similitudes entre distintas variedades lingüísticas en perspectiva diacrónica. Esos subcorpus presentan textos de los mismos tipos documentales que incluye ODE. Con ello, CORTENEX se establece como otro subcorpus de control de documentación notarial extremeña del siglo XVII.

Por el momento, CORTENEX solo cuenta con textos notariales de la provincia de Badajoz del siglo XVII, si bien va aumentado de forma progresiva y, posteriormente, se incluirá documentación notarial de la provincia de Cáceres. Con respecto a la provincia de Badajoz, están disponibles para su consultan un total de 41 documentos totalmente digitalizados en ODE.

4. El corpus CORTENEX (s. XVII)

Entre los primeros pasos en la elaboración de CORTENEX S. XVII, se realizó un proceso de búsqueda y selección de textos notariales a través de criterios histórico-lingüísticos en el Archivo Histórico Provincial de Badajoz y de Cáceres. Se procuró

[10] "Hispanae Testium Depositiones: las declaraciones de testigo en la historia del español" FFI2017-83400-P (MINECO/AEI/FEDER, UE).
[11] "Atlas Lingüístico y Etnográfico de Andalucía, S. XVIII. Patrimonio documental y humanidades digitales" (Proyectos I+D+i Junta de Andalucía-FEDER, P18-FR-695).
[12] Por ejemplo, ODE ya cuenta con documentación notarial procedente de Madrid (Arrabal Rodríguez 2020).

seleccionar aquellos documentos que se correspondiesen con cartas de dote, testamentos e inventarios de bienes en el mejor estado posible. Este tipo de textos, por sus características, listan una variedad enorme de objetos, muebles y otros enseres domésticos con gran nivel de detalle en su descripción y valor. Desde el punto de vista léxico, estos textos se erigen como especialmente adecuados a la hora de plantear estudios de léxico histórico, de variación diatópica, de documentación de préstamos y voces regionales, así como de voces procedentes de otras variedades del español.

Se estableció una selección de manuscritos de diversos notarios desde finales del siglo XVI hasta mediados del siglo XVII. En la Tabla 1 se observa la relación de notarios seleccionados para la provincia de Badajoz, donde se ordenan los legajos por fecha y se indica el número total de textos notariales que se han seleccionado de cada uno de ellos:

Relación de notarios y legajos	**Número de textos notariales seleccionados**
Diego López (1597-98; 1601-1608)	11
Marcos de Herrera (1601; 1622-23)	29
Diego Sánchez (1609)	5
Baltasar Suárez (1614; 1624)	10
Diego Martín Sequera (1615; 1628-1630; 1645)	29
Joan Gómez (1614)	3
Pedro de Tovar (1629)	19
Pedro Sánchez Ardilla (1631)	11
Manuel de León (1635, 1640)	17
Diego Martín Gamo (1651)	8

Tabla 1. Relación de notarios y número de documentos seleccionados

Como se observa, se ha procurado seleccionar textos en intervalos de tiempo de entre 5 y 10 años.[13] Los documentos disponibles para su consulta en la actualidad se

[13] Debido al estado de algunos legajos no siempre se ha podido seguir con rigurosidad este criterio de selección documental en intervalos temporales.

corresponden los de Diego Martín Sequera,[14] Diego López y Baltasar Suárez. Por el momento, la selección documental integrada a CORTENEX queda de la siguiente forma:

Diego Martín Sequera (1615-1645)	Inventarios de bienes	7
	Cartas de dote	11
	Testamentos	11
Diego López (1601)	Inventarios de bienes	1
	Cartas de dote	10
Baltasar Suárez (1624)	Inventarios de bienes	1
Total	41	

Tabla 2. Documentación consultable de CORTENEX

4.1. Transcripción en XML-TEI

Una vez seleccionados los documentos, la siguiente fase en el proceso de elaboración de este corpus consistió en transcribirlos en lenguaje marcado XML, atendiendo al estándar del consorcio TEI (http://www.tei-c.org/index.xml) a la hora de codificar corpus digitales.[15] La estructura que presentan los documentos en formato XML-TEI permite etiquetar información lingüística de los manuscritos de tal forma que su procesamiento informático es fácil de interpretar computacionalmente (Calderón Campos 2019: 24), lo cual facilita las posibilidades de estudios comparados con otros corpus regidos por la misma metodología. Todos los documentos XML-TEI se organizan en dos bloques de contenido: la cabecera y el cuerpo. Los datos que se introducen en la cabecera son de tipo metatextual y permiten clasificar los documentos, así como aportar otra información adicional sobre el proyecto, el título del texto, el archivo del

[14] Existen protocolos asociados a este notario desde 1598 hasta 1657. Es posible que hubiera dos escribanos con este nombre, seguramente, padre e hijo (Guerra Guerra 1977: 65).
[15] Este modelo de transcripción documental cuenta con gran uso en lingüística, pero aún es limitado en su aplicación a corpus diacrónicos (Calderón Campos 2019: 23).

que procede, la signatura del protocolo, la data o los folios. El cuerpo se traduce en la transcripción propiamente dicha, haciendo uso de las etiquetas establecidas y consensuadas en ODE.

Los criterios de transcripción documental se basan, fundamentalmente, en: 1) por un lado, normalizar la puntuación de los documentos acorde a las normas actuales de la Real Academia Española, así como adaptar el uso de las mayúsculas y la separación de palabras (Arrabal Rodríguez 2020: 69); y 2) de otro lado, mantener las grafías originales de los manuscritos por el interés filológico que muestran y por presentar una edición paleográfica lo más estrecha posible de los textos.

El empleo del lenguaje XML-TEI ha permitido marcar con etiquetas aspectos de los manuscritos relacionados con el contenido y la disposición que este presenta, esto es, información textual y paratextual: la numeración, la disposición de las líneas, las tachaduras, conjeturas, etc. Obsérvese la Figura 1:

```
<pb n="132v" facs="IMG_1827.JPG"/>
<p>
    <lb/> Una escudilla p<add place="above">a</add>||para linpiar dos reales
    <lb/> Media dosena de platos y media de es<lb/>cudillas blancos en quatro reales
    <lb/> Un manto de anascote nuebo y otro raydo
    <lb/> tasados en çien reales
    <lb/> Una basquiña de bayeta aforrada
    <lb/> de bocasi azul <unclear>con una banda de</unclear> terçiope<lb/>lo tasada en quatro ducados
    <lb/> Otra basquiña de <gap/> blanco y negro
    <lb/> tasada en tres ducados
    <lb/> Otra basquiña <gap/> cabellado
    <lb/> aforrada en <gap/> en dos
    <lb/> ducados
    <lb/> Un faldellin de jerguilla azul tasado
    <lb/> en çinquenta reales
    <lb/> Un jubonsillo de tafetan negro doble
    <lb/> aforrado en fustan tasado en seis duca<lb/>dos
    <lb/> Otro jubonsillo de sarga negra tasado en
    <lb/> tres ducados
    <lb/> Un corpiño de tramoya de seda leonado
    <lb/> y negro tasado en dos ducados
    <lb/> Una esterilla de junco blanca en seis
    <lb/> reales
    <lb/> Un arca pequeña de madera con su se<lb/>rradura y llabe, tasada en ocho reales
    <lb/> Dos candiles de hierro en quatro reales
    <lb/> Dos sedasos y una ajuera en ocho reales
    <lb/> Una caldera mediana de cobre tasada en
```

Figura 1. Ejemplo de transcripción semipaleográfica de carta de dote en XML

Como se deduce, en lenguaje XML hay etiquetas de uso recurrente, como el marcado de inicio de línea (<lb/>), de párrafo ( ) o de página (<pb/>). Otras

etiquetas frecuentes son: <add></add> para indicar la disposición de algunas palabras (por encima o por debajo de la línea) o añadidos al margen, <del></del> para indicar una tachadura, <supplied></supplied> o <unclear></unclear> para conjeturas más o menos claras del editor, entre otras.[16]

Con todo, las *etiquetas* se corresponden con marcas entre paréntesis angulares. Además, estas etiquetas van ampliándose a través de distintos *atributos*, que, a su vez, pueden incluir un *valor*. Esto sucede, por ejemplo, con la etiqueta <add>. Como se observa en la Figura 1, <add> lleva un atributo *place* con un valor *above*, lo que permite visualizar la letra de la palabra que ha sido escrita originalmente por encima de la línea. El uso de todas estas etiquetas, atributos y valores permite conseguir una visualización del texto semipaleográfica en TEITOK, totalmente fiel al manuscrito, sin presencia de dichas etiquetas para facilitar la lectura (Figura 2):

Una escudilla pª linpiar dos reales
Media dosena de platos y media de es
cudillas blancos en quatro reales
Un manto de anascote nuebo y otro raydo,
tasados en çien reales
Una basquiña de bayeta aforrada
de bocasi azul con una banda de terçiope
lo tasada en quatro ducados
Otra basquiña de [...] blanco y negro
tasada en tres ducados
Otra basquiña [...] cabellado
aforrada en [...] en dos
ducados
Un faldellin de jerguilla azul tasado
en çinquenta reales
Un jubonsillo de tafetan negro doble
aforrado en fustan tasado en seis duca
dos
Otro jubonsillo de sarga negra tasado en
tres ducados

Figura 2. Presentación semipaleográfica de un fragmento de carta de dote en ODE

Por lo que respecta a los textos extremeños seleccionados, estos han sido transcritos en su totalidad, incluyendo las fórmulas estereotipadas de apertura y cierre de estos tipos textuales, así como el cuerpo del inventario, carta de dote o testamento,

[16] Las etiquetas más usadas en ODE pueden consultarse en Calderón Campos (2019: 25).

que presenta el mayor interés léxico. El objetivo consiste en que esta documentación pueda servir para estudios, no solo de carácter léxico, sino también de corte fonético y morfosintáctico, así como para otras cuestiones etnográficas o culturales.

La transcripción en XML permite tener las ediciones semipaleográfica y normalizada en un mismo documento que posteriormente se procesa en TEITOK. A pesar de que existen múltiples editores XML gratuitos, para elaborar las transcripciones en CORTENEX se ha usado el editor *Oxygen* por las ventajas que ofrece en el uso de plantillas predeterminadas y adaptas a TEI.[17]

4.2. Procesamiento lingüístico

4.2.1 Tokenización y normalización ortográfica

Una vez realizada la transcripción de los documentos, la siguiente fase en la elaboración de CORTENEX consiste en procesar lingüísticamente toda la información marcada con el objetivo de obtener un corpus en el que se puedan hacer búsquedas lingüísticas a un nivel muy avanzado. Para ello, desde la plataforma TEITOK se realizan directamente los cuatro pasos más importantes en cuanto a procesamiento lingüístico: tokenización, normalización lingüística, lematización y anotación morfosintáctica (Vaamonde 2015).

Los documentos en formato XML-TEI, junto a las imágenes de los facsímiles digitalizadas en formato JPG, se importan a la plataforma TEITOK. De forma automática, se visualiza la opción de *tokenizar* el documento, es decir, añadir a cada palabra del documento una nueva etiqueta <tok></tok>.

Tras realizar este proceso, se comienza la fase de normalización ortográfica de los textos. Gracias a la adición de la etiqueta <tok>, TEITOK ofrece una herramienta semiautomática para normalizar cada documento. De esta forma, dentro de cada token se irá añadiendo toda la información relacionada con la edición de cada palabra en forma de atributos: la forma original (*pform*), la forma expandida (*fform*) y forma

[17] https://www.oxygenxml.com/.

normalizada (*nform*). Los valores concretos de estos tres atributos se traducen en la forma que se corresponde con cada variante de la palabra. Así, será posible realizar búsquedas dependiendo de la forma ortográfica, dado que todas las versiones de una palabra quedan asociadas a un mismo token.

La tokenización de los textos en TEITOK, así como la normalización ortográfica, se puede hacer de forma automática. No obstante, siempre conviene revisar y añadir manualmente aquello que no haya sido normalizado de forma correcta en la plataforma. Una forma que presente como grafía original *çiudad*, tras realizar el proceso de tokenización y normalización, tendría el siguiente aspecto en el editor XML que incluye TEITOK:

(1) <tok id="w-3" nform= "ciudad">çiudad</tok>

Otros ejemplos de ello pueden observarse en el caso de los sustantivos *asofar* o *treodes*:

(2) <tok id="w-757" nform="azófar"> asofar</tok>

<tok id= "w-702" nform= "trébedes">treodes</tok>

En la documentación histórica son muy frecuentes las formas abreviadas de algunas palabras. Si seguimos con el ejemplo de *çiudad*, esta voz suele presentarse frecuentemente en los textos de forma abreviada *çiud*, por ello, es necesario el atributo *fform,* que añade la forma expandida sobre la base de la grafía original, de tal manera que en la forma expandida de *çiud* se mantiene la letra *ç*: <tok id= "w-3" nform= "ciudad"; fform= "çiudad">çiud</tok>. Algunos errores frecuentes en la normalización automática tienen que ver con palabras que, por su grafía original en la transcripción del manuscrito, pueden confundirse con otras. Por ejemplo, en los textos extremeños se documenta la forma *sera* en el sentido de 'cera'; al normalizar de forma automática, la herramienta interpreta que dicha forma se corresponde con la tercera

persona del singular del futuro del verbo *ser*. Este tipo de errores se corrigen de forma manual.

4.2.2 Lematización y anotación morfosintáctica

Puesto que el objetivo de CORTENEX no se limita únicamente a digitalizar manuscritos con la posibilidad de ofrecer varias ediciones del mismo texto, sino que también se pretende elaborar un corpus anotado, TEITOK ofrece la posibilidad de lematizar y anotar morfosintácticamente cada palabra de forma semiautomática con la herramienta Neotag (Janssen 2012). Esto es posible gracias a la normalización previa de cada texto, sobre la que esta herramienta irá asignando lemas y etiquetas morfosintácticas.

El funcionamiento de esta herramienta consiste en añadir dos nuevos atributos a cada token: 1) un atributo *lemma* y 2) un atributo POS.[18] Por ejemplo, una de las formas verbales recurrentes en los testamentos es *sepan*. Este verbo, tokenizado, normalizado, lematizado y anotado tendría el aspecto siguiente:

(3) <tok id="w-11" nform="Sepan" lemma="saber" pos="VMSP3P0">Sepan</tok>

Ambos atributos adquieren un valor concreto asignado por Neotag. Así, VMSP3P0 se corresponde con: verbo principal, subjuntivo, presente, tercera persona, plural; la forma en infinitivo *saber* será el valor del lema de cualquier tiempo verbal que presente este verbo. El etiquetario implementado en CORTENEX toma como modelo el empleado en ODE, es decir, se basa en una versión simplificada de las directrices marcadas por el grupo EAGLES.

El etiquetador POS Neotag fue diseñado originalmente para rastrear neologías, pero, por las ventajas que ofrece, se usa como etiquetador en corpus convencionales. En líneas generales, para asignar lemas y etiquetas morfosintácticas a cada palabra de un corpus, Neotag calcula las etiquetas más probables dentro de una secuencia de

[18] "Part of speech".

palabras a partir de las frecuencias de cada palabra y etiqueta de un corpus de entrenamiento (Janssen 2012), para lo que se basa en las terminaciones de cada palabra. Al mismo tiempo, realiza un proceso de suavizado léxico (*lexical smoothing*), que toma las formas ya conocidas para rebajar las irregularidades a la hora de asignar un POS final. El mayor o menor éxito dependerá del tamaño del corpus de entrenamiento.

El proceso de lematización y etiquetado morfosintáctico es simultáneo y, una vez finalizado, se mostrará el lema de cada palabra con su etiqueta morfosintáctica. La visualización final queda de la siguiente forma (Figura 3):

Figura 3. Visualización final edición normalizada, lematizada y etiquetada de una carta de dote en CORTENEX

Al igual que sucede en el proceso de normalización, es necesario hacer una revisión manual del etiquetado y de los lemas que la herramienta Neotag ha ido asignando por los posibles errores que se pueden detectar. Es común observar errores en el etiquetado de palabras ambiguas y que pertenecen a categorías gramaticales distintas (*manda* como sustantivo o como verbo).

Los documentos extremeños completamente tokenizados, normalizados, lematizados y etiquetados morfosintácticamente en TEITOK tienen el siguiente aspecto final en lenguaje XML:

Figura 4. Visión final de transcripción completa, tokenización, normalización, lema y POS

Como se observa, cada una de las palabras de un documento lleva una etiqueta <tok> asociada, dentro de la que se ha ido añadiendo diversos atributos referidos a la forma, al lema y a la categoría gramatical. De esta manera, las búsquedas que se pueden realizar oscilan desde las más simples a las combinaciones más complejas desde el punto de vista lingüístico, incluso, se pueden realizar búsquedas cruzadas para comparar la variación de un fenómeno determinado por provincias, debido a la documentación que incluye actualmente ODE.

Por ello, conviene detenerse en el motor de búsqueda CQL (*Corpus Query Language*) que incluye la plataforma TEITOK en los corpus que aloja actualmente.[19] Pueden realizarse búsquedas por lema, por forma normalizada, por grafía original, por etiqueta morfosintáctica, combinando provincias que en ODE ya cuentan con documentación, por tipología textual, etc. Para facilitar este tipo de búsquedas, la

[19] Actualmente, TEITOK da cabida a 22 proyectos de corpus lingüísticos <http://www.teitok.org/index.php?action=projects>.

plataforma cuenta con un etiquetario y un manual de búsqueda para los usuarios, además de un generador de búsquedas (*query builder*) cuya interfaz resulta sencilla de utilizar:

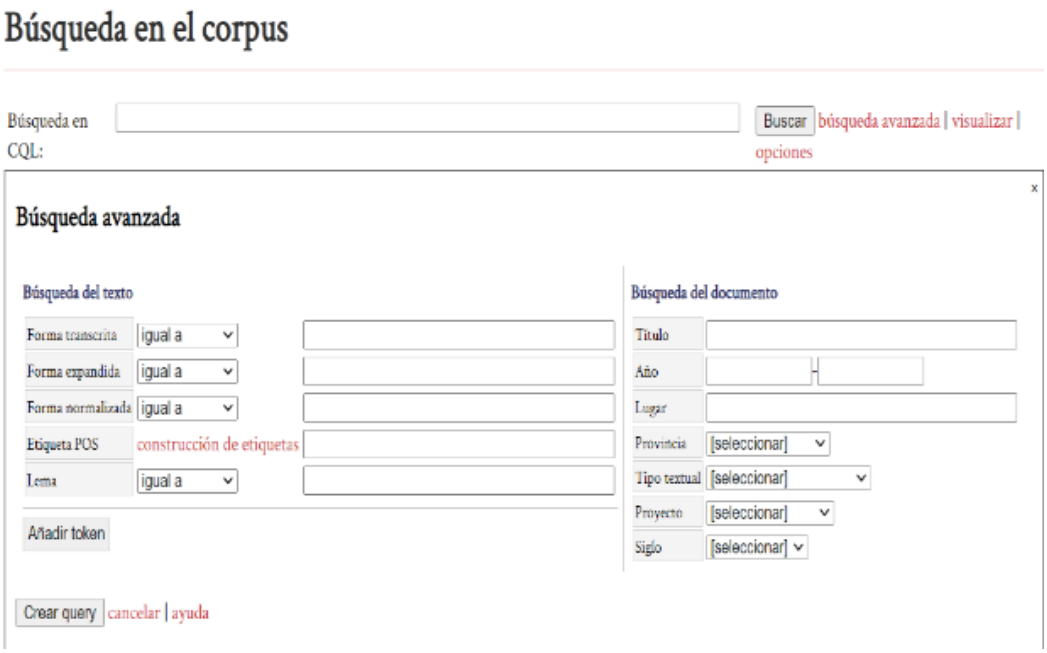

Figura 5. Generador de búsquedas en ODE

Por ejemplo, si queremos saber cuánta variación ortográfica presenta la voz *azófar* en la documentación extremeña, se realizará una búsqueda por lema a la vez que se escoge una provincia del desplegable: [lemma:“azófar”].

Figura 6: Ejemplo motor de búsqueda ODE

Si el objetivo es conocer cuántos sustantivos masculinos en plural incluyen los documentos, resulta muy útil la búsqueda por etiqueta morfosintáctica: [pos = "NCMP.*"].

Otra aplicación que ha incorporado TEITOK recientemente tiene que ver con la posibilidad de recuperar la información en un mapa, algo muy útil para los estudios de léxico histórico de corte dialectal y, a la vez, muy cómodo para usuarios no especializados o recién iniciados:

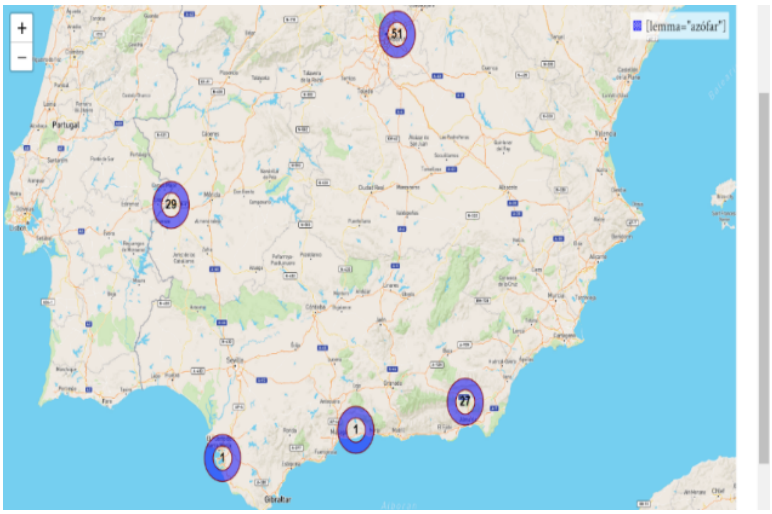

Figura 7. Mapa de *azófar* según los documentos disponibles en ODE

5. Conclusiones

La metodología llevada a cabo en la elaboración de CORTENEX S. XVII, basada en el uso del lenguaje marcado XML y de la plataforma TEITOK, es una de las más pioneras en la elaboración de corpus diacrónicos actualmente. De tal manera, este corpus se establece como una aportación más al caudal de documentos históricos de tipo notarial procesados lingüísticamente con los que cuenta la lengua española. Además, CORTENEX se encuentra integrado como subcorpus específico en ODE.

La transcripción de los documentos notariales en XML-TEI ha permitido generar dos ediciones de cada texto, gracias al hecho de poder etiquetar aquellos aspectos que resulten de interés desde el punto de vista histórico-lingüístico según las líneas de codificación que arroja el consorcio TEI. Con ello, se obtiene una edición semipaleográfica de los documentos al mismo tiempo que es posible visualizar una edición normalizada acorde a la normativa ortográfica actual. Como se ha mencionado, las posibilidades del corpus no se acaban ahí, sino que, además, es posible visualizar la lematización de cada palabra junto con su etiqueta morfosintáctica gracias al uso de la herramienta Neotag.

Debido a que CORTENEX está integrado a la plataforma TEITOK a través de ODE, es posible plantear búsquedas lingüísticas complejas gracias al motor de búsquedas CQL que se incluye. Además, al incluir ODE documentación de tipo notarial de otros territorios en el mismo marco cronológico, los estudios de variación diatópica léxica que se realicen podrán dibujar un mapa lingüístico del siglo XVII de voces de diversos ámbitos de la vida cotidiana (agricultura, ganadería, creencias, indumentaria, enseres domésticos, joyas, propiedades, etc.) de varias modalidades dialectales.

Por último, CORTENEX ofrece la posibilidad de profundizar en los fenómenos lingüísticos que caracterizan la variedad de español que se desarrolla en Extremadura desde el punto de vista diacrónico, sobre todo en el ámbito léxico. Hasta el momento, solo están disponibles documentos geográficamente adscritos a la provincia de Badajoz, aunque el proyecto cubra un territorio mucho más amplio de toda la comunidad extremeña. Este corpus se erige, por tanto, como una aportación más al

conocimiento de esta variedad en su perspectiva diacrónica y diatópica. Con ello, este corpus está orientado a diversas posibilidades de estudios de corte léxico, histórico y cultural que de él se pueden derivar y contribuye a poder establecer una mejor caracterización de los rasgos lingüísticos propios del español en Extremadura y de las áreas geográficas de convergencia de esos rasgos en áreas dialectales más extensas. Desde el punto de vista léxico, las primeras indagaciones han permitido constatar occidentalismos (*juera*), portuguesismos (*caniquí*, *tacho*), arabismos (*alioj, resma, badana, guadamecí*) y galicismos (*bombasí*) relacionados con la agricultura, los nombres de telas, de enseres domésticos, entre otros.

No obstante, la documentación notarial extremeña también permite plantear estudios fonéticos (p.ej. la vacilación en la grafía de las sibilantes) y morfosintácticos (p. ej. el orden de los complementos verbales, las formas de tratamiento, el uso de apreciativos, etc.); estos rasgos lingüísticos pueden ser rastreados gracias a la precisión del motor de búsqueda CQL. Toda la documentación extremeña disponible actualmente se puede consultar en http://corpora.ugr.es/ode.

Referencias bibliográficas

ACADEMIA MEXICANA DE LA LENGUA *Corpus diacrónico y diatópico del español de América* (CORDIAM). En línea: <http://www.cordiam.org/>.

ARIZA VIGUERA, Manuel (1985) “Dos estudios de la historia lingüística de Extremadura”, *Anuario de Estudios Filológicos*, 8, 7-18.

ARRABAL RODRÍGUEZ, Pilar (2020) “Edición de un corpus digital de inventarios de bienes”, *Procesamiento del Lenguaje Natural*, 65, 67-74.

BARROS GARCÍA, Pedro (1974) *El habla de Arroyo de San Serván*, Granada: Universidad de Granada.

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

CALDERÓN CAMPOS, Miguel & María Teresa GARCÍA-GODOY (2010-2019) *Oralia diacrónica del español (ODE)*. En línea: <http://corpora.ugr.es/ode>.

CALDERÓN CAMPOS, Miguel (2019) "La edición de corpus históricos en la plataforma TEITOK. El caso de *Oralia diacrónica del español*", *Chimera,* 6, 21-36.

CHARTA = *Corpus hispánico y americano en la red*. En línea: http//www.corpuscharta.es.

CorLexIn = *Corpus Léxico de Inventarios*. En línea: <http://web.frl.es/CORLEXIN.html>.

CRADDOCK, Jerry (2015) *Cíbola Project. Editing the Documents of the Hispanic Southwest in the 16th and 17th Centuries*. En línea: <http://escholarchip.org/uc/rcrs_ias_ucb_cibola>.

CUMMINS, John (1974) *El habla de Coria y sus cercanías*, Londres: Tamesis Book Limited.

FERNÁNDEZ DE MOLINA, Elena (2014) "La investigación científica en el habla de Extremadura: monografías dialectales y estudios sobre fonética y fonología extremeña", *Anuario de Estudios Filológicos,* 37, 5-20.

FLORES MANZANO, Florencio (1988) "Incidencia del factor histórico en la configuración geolingüística de Extremadura", en *Actas del I Congreso Internacional de Historia de la lengua española*, Madrid: Arco/Libros, 1449-1460.

GARCÍA MOUTON, Pilar (1996) "El extremeño", en *Lenguas y dialectos de España*, Madrid: Arco/Libros, 31-34.

GARGALLO GIL, José Enrique (1999) *Las hablas de San Martín de Trevejo, Eljas y Valverde del Fresno*, Mérida: Editora Regional de Extremadura.

GONZÁLEZ SALGADO, José Antonio (2003) *Cartografía lingüística de Extremadura*, Madrid: Universidad Complutense. En línea: http://www.geolectos.com/.

GONZÁLEZ SALGADO, José Antonio (2009) "Diez problemas de dialectología extremeña", *Revista de Estudios Extremeños*, LXV(1), 347-378.

GUERRA GUERRA, Antonio (1977) "Escribanos badajocenses del siglo XVII", *Revista de estudios extremeños*, 33(1), 5-68.

JANSSEN, Maarten (2012) "Neotag: a POS tagger for grammatical neologism detection", en *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, *LREC 2012*, Estambul.

JANSSEN, Maarten (2016) "TEITOK: Text-Faithful Annotated Corpora", en *Proceedings of the Language Resources and Evaluation Conference (LREC 2016) ELRA*. Portoroz, Eslovenia, 4037-4043.

LORENZO CRIADO, Emilio (1948) "El habla de Albalá. Contribución al estudio de la dialectología extremeña", *Revista de Estudios Extremeños*, IV, 398-407.

MARCOS ÁLVAREZ, F. (1992) "Algunas precisiones léxicas sobre indumentaria española en el siglo XVII", en *Actas del II Congreso Internacional de Historia de la lengua española*, (Madrid: Pabellón de España, 1161-1172.

MONTERO CURIEL, Pilar (2006) *El extremeño*, Madrid: Arco/Libros.

MORALA, Ramón (2014) "El *CorLexIn*, Un corpus para el estudio del léxico histórico y dialectal del Siglo de Oro", *Scriptum Digital*, 3, 5-28.

PALACIOS MARTÍN, Bonifacio (1988) "Origen de la conciencia regional extremeña: el nombre y el concepto de Extremadura", *Alcántara: Revista del Seminario de Estudios Cacereños,* 13-14, 9-22.

TEI CONSORTIUM (2007) *TEI PS: Directrices para la codificación e intercambio electrónico de texto.* En línea <http://www.tei-c.org/Guidelines/P5/>.

SÁNCHEZ SIERRA, Diego8 (2019) *Edición y estudio léxico de fuentes documentales extremeñas de los siglos XVI y XVII*, Madrid: Universidad de Alcalá de Henares.

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

VELO NIETO, Juan José (1956) "El habla de las Hurdes", *Revista de estudios extremeños*, 12 (1-4), 59-207.

VIUDAS CAMARASA, Antonio, Manuel ARIZA VIGUERA & Antonio SALVADOR PLANS (1987) *El habla en Extremadura*, Junta de Extremadura: Consejería de Educación y Cultura.

ZAMORA VICENTE, Alonso (1943) *El habla de Mérida y sus cercanías*, Madrid: Anejo XXIX de la *Revista de Filología Española*.