Scriptum digital
Vol. 9 (2020), pp. 167-189

# *ORALIA DIACRÓNICA DEL ESPAÑOL*: UN NUEVO CORPUS DE LA EDAD MODERNA[1]

MIGUEL CALDERÓN CAMPOS (*Universidad de Granada*)
calderon@ugr.es
ORCID-iD: http://orcid.org/0000-0002-0656-3643
GAEL VAAMONDE DOS SANTOS (*Universidad de Granada*)
gaelvaamonde@ugr.es
ORCID-iD: http://orcid.org/0000-0001-8360-2805

**RESUMEN**
Transcurridas casi dos décadas desde la aparición de los grandes corpus históricos, la posibilidad de acceder a grandes bancos de datos de una forma rápida y sencilla se ha asumido ya como metodología habitual. No obstante, en los últimos años se ha ido imponiendo una mirada más exigente que ha derivado en la construcción de corpus históricos especializados de tamaño más reducido (Enrique-Arias 2009; Kabatek 2016). En este artículo se presenta el corpus *Oralia diacrónica del español* (*ODE*), compuesto por inventarios de bienes, declaraciones de testigos y certificaciones de cirujanos. Respecto a la metodología, las transcripciones de los manuscritos inéditos se procesan en la plataforma *TEITOK* (Janssen 2016), especialmente diseñada para tokenizar, normalizar y anotar textos en formato *XML-TEI*. El resultado final es un producto que combina la edición digital, la imagen facsimilar y el corpus lingüísticamente anotado, convirtiendo así a *ODE* en un recurso electrónico de utilidad para paleógrafos, filólogos, lingüistas e historiadores.
**PALABRAS CLAVE**: diseño de corpus electrónicos, siglos XVI-XIX, anotación de corpus, dialectología histórica, *XML-TEI*.

## *ORALIA DIACRÓNICA DEL ESPAÑOL*: A NEW CORPUS OF THE MODERN AGE

**ABSTRACT**
Almost two decades after the appearance of big historical corpora, accessing large databases in a quick and easy way has already become part of the standard methodology. However, in recent years a more demanding approach has been imposed, which has led to the construction of smaller specialized historical corpora (Enrique-Arias 2009; Kabatek 2016). This article presents the corpus *Oralia diacrónica del español* (*ODE*), composed of inventories of goods, witness statements and surgeons' certifications. Regarding the methodology, transcriptions of unpublished manuscripts have been processed on the *TEITOK* platform (Janssen 2016), which is specially designed to tokenize, standardize and annotate texts in the *XML-TEI* format. The final result is a product that combines the digital edition, the facsimile image and the linguistically annotated corpus, making *ODE*, as such, a useful electronic resource for palaeographers, philologists, linguists and historians.
**KEY WORDS**: Electronic corpus design, 16th-19th centuries, corpus annotation, historical dialectology, *XML-TEI*.

## 1. INTRODUCCIÓN

El corpus *Oralia diacrónica del español* (*ODE*) recoge documentación manuscrita inédita del periodo comprendido entre 1492 y 1900, centrada en tres tipos textuales: inventarios de bienes, declaraciones de testigos en juicios criminales y certificaciones periciales de cirujanos vinculadas con esos juicios. Es continuación del *Corpus diacrónico del español del reino de Granada*, 1492-1833, *CORDEREGRA* (Calderón Campos 2015), del que se diferencia en tres aspectos fundamentales:

[1] Este trabajo se inscribe en el proyecto *«Hispanae Testium Depositiones». Las declaraciones de testigo en la historia de la lengua española, 1492-1833* (*HISPATESD*, FFI2017-83400-P, MINECO/AEI/FEDER/UE, 2018-2021).

Fecha de recepción: 03/04/2020
Fecha de aceptación: 20/05/2020
ISSN: 2014-640X

En primer lugar, se modifica la extensión geográfica del corpus. *CORDEREGRA* ofrecía documentación únicamente de las actuales provincias de Granada, Málaga y Almería, que durante el periodo comprendido entre 1492 y 1833 habían constituido la entidad administrativa denominada «reino de Granada», prolongación del antiguo reino nazarí.

La ampliación territorial se hace con objeto de añadir subcorpus de control respecto de los cuales la documentación granadina adquiere mayor significado, tanto desde el punto de vista dialectal como diacrónico. La información del primitivo corpus base no podía contrastarse, al no disponer de datos comparables procedentes de los mismos tipos textuales de otras regiones o épocas.

Para solventar este problema se decidió ir creando paulatinamente, en sucesivas etapas vinculadas a nuevos proyectos de investigación[1], dos subcorpus de referencia: uno de la región centro-norte peninsular y otro de la zona occidental de Andalucía (Cádiz, Huelva, Sevilla). Ambos corpus estarán constituidos con los mismos tipos textuales de *CORDEREGRA*.

En segundo lugar, se amplía el lapso cronológico. Originariamente, *CORDEREGRA* se limitaba a la duración del reino de Granada, por lo que se cerraba en 1833, año en el que el ministro de Fomento Javier de Burgos redactó el decreto por el cual España quedaba dividida en 49 provincias. En concreto, en lo referente a Andalucía, desaparecieron a partir de entonces los antiguos reinos de Córdoba, Granada, Jaén y Sevilla, para dar paso a las ocho provincias actuales.

Dado que *ODE* ya no se centra solo en el reino de Granada, es más lógico, en consonancia con la práctica de otros corpus históricos (*CODEA*+ 2020)[2], extender el arco temporal a todo el siglo XIX. De esta forma, recogerá documentación de toda la edad moderna y principios de la contemporánea.

Por último, se renuevan completamente las características tecnológicas del corpus: en *CORDEREGRA* se ponía el foco en aspectos filológicos como la selección documental representativa o la transcripción rigurosa de los manuscritos, que se hacía en formato *Microsoft Word* y se publicaba en línea en formato *PDF*[3]. En *ODE*, por el contrario, se suma al rigor filológico un decidido compromiso de renovación tecnológica, que implica profundas novedades en la metodología empleada en el tratamiento de los datos. Los documentos de *ODE* se transcriben en *XML* adoptando los estándares de codificación propuestos por el consorcio *TEI* (*Text Encoding Initiative*) para la edición de textos en formato digital. Una vez trascrita, la documentación se procesa computacionalmente en la plataforma *TEITOK* (Janssen 2014), una herramienta especialmente diseñada para crear,

[1] En el proyecto *Atlas lingüístico y etnográfico de Andalucía. Siglo XVIII. Patrimonio documental y humanidades digitales*. *ALEA-XVIII* (P18-FR-695, Junta de Andalucía. Proyectos I+D+i Modalidad: Generación de conocimiento "Frontera"; *PAIDI* 2020; convocatoria de 2018), vinculado con *HISPATESD*, se abordará la elaboración de dos corpus de inventarios del siglo XVIII: uno de las provincias andaluzas occidentales (Huelva, Sevilla y Cádiz) y otro de la Comunidad de Madrid. Ambos se incluirán en *ODE* como corpus de referencia de la documentación del reino de Granada.

[2] El corpus *CODEA* ha ido ampliando su arco cronológico en sucesivas ediciones: empezó siendo un *Corpus de documentos españoles anteriores a 1700*, para abarcar posteriormente hasta 1800 (*CODEA*+ 2015) y finalmente hasta 1900 (*CODEA*+ 2020).

[3] En el proyecto *Cíbola* (Jerry R. Craddock, UC Berkeley), que recoge documentación en español de los siglos XVI a XVIII del suroeste de los Estados Unidos, tanto las transcripciones semipaleográficas como la imagen facsímil de los documentos se presentan en línea en formato PDF.

mantener y publicar en línea ediciones digitales y simultáneamente corpus con anotación lingüística.

De esta forma, *ODE* aúna intereses de carácter filológico con otros propios de la lingüística computacional. En el primer caso, el foco está en la edición digital rigurosa de los documentos (respeto de las grafías originales, información sobre cambios de escribano, tachaduras, fragmentos de lectura dudosa, etc.); en el segundo, la atención se focaliza en la anotación morfosintáctica de los textos (*POS tags*), en la lematización y normalización de las variantes morfológicas y ortográficas y en la elaboración de buscadores potentes, capaces de satisfacer todas las necesidades del análisis histórico-lingüístico.

## 2. ORALIDAD Y TIPOS TEXTUALES DEL CORPUS

Sin lugar a dudas, una de las grandes limitaciones de los corpus diacrónicos es la escasa representación de la lengua hablada (Rodríguez Puente 2018: 90). Por este motivo, desde los años noventa se intenta incluir en los estudios histórico-lingüísticos tipos textuales próximos a lo oral y cotidiano, y a las distintas esferas de actividad de una comunidad: cartas privadas (Fontanella de Weinberg 1992; Raumolin-Brunberg y Nevalainen 2007[4]; Fernández Alcaide 2009; Arias Álvarez y Hernández Mendoza 2013[5]; Vaamonde 2015, 2018a, 2018b), crónicas de soldados (Di Tullio y Resnik 2019)[6], quejas (Octavio de Toledo y Huerta, y Pons Rodríguez 2017), autobiografías (Rivadeneira en prensa), peticiones de ayuda a la beneficencia (Sánchez-Prieto Borja y Vázquez Balonga 2017, 2019), diálogos (Culpeper y Kytö 2010)[7], declaraciones de testigo (Calderón Campos 2015, *Old Bailey Corpus*[8]) e inventarios de bienes (Morala 2012, 2018).

Los tres tipos textuales representados en *ODE* (declaraciones de testigos, inventarios y certificaciones médicas) suponen el traslado al papel de las declaraciones orales de testigos, tasadores y cirujanos, por lo que no es de extrañar que se cuelen en los documentos rasgos lingüísticos del vernáculo de los declarantes o del escribano. En los tres

[4] El *Corpus of Early English Correspondence* (*CEEC*, 1410-1681) es un corpus diacrónico de correspondencia personal realizado por el grupo de investigación *VARIENG* (*Research Unit for Variation and Change in English*) de la Universidad de Helsinki. El corpus inicial (1998) se ha ido complementando en sucesivas fases, en las que se han incorporado mejoras relacionadas con la anotación morfosintáctica, la extensión cronológica (desde 1402 hasta 1800) o la representatividad regional.

[5] En el *Corpus electrónico del español colonial mexicano* (*COREECOM*) se recogen todas las variedades textuales posibles, lo que significa que en el nivel diafásico se distinguen tres tipos de textos: informales (cartas de amor, cartas a familiares y amigos, recados), semiformales (denuncias, autodenuncias, cartas de relación y defensas) y formales (juicios, testamentos, cédulas, facturas y actas).

[6] Analizan lingüísticamente el *Diario de un soldado*, un texto anónimo de comienzos del XIX en Buenos Aires, redactado por un autor que representa los rasgos prototípicos de llamado "scripteur maladroit" (Blanche-Benveniste 1994) o "mão inábil" (Marquilhas 1998).

[7] El *CED* (*A Corpus of English Dialogues 1560-1760*, 2006, Universidades de Uppsala y de Lancaster) consta de 1,2 millones de palabras, pertenecientes a cinco tipos textuales: procesos judiciales, declaraciones de testigos, obras teatrales, diálogos didácticos y prosa de ficción. Puede consultarse en la *CQP-web* de la Universidad de Lancaster: https://cqpweb.lancs.ac.uk/.

[8] La última versión del *Old Bailey Corpus* (versión 2.0) consta de 24,4 millones de palabras. Contiene 637 actas judiciales procedentes de Old Bailey, el antiguo tribunal de justicia de Londres. Recoge declaraciones de testigos del periodo comprendido entre 1720 y 1913. El corpus, completamente anotado, puede consultarse en la *CQP-web (CLARIN)* de la Universidad de Saarland: http://corpora.clarin-d.uni-saarland.de/cqpweb/.

casos la información se transmite de forma muy similar: el declarante habla y el escribano anota lo que oye, con una finalidad puramente práctica (registrar la verdad de los hechos), y sin una excesiva preocupación ni control por la forma, con el filtro normativo de su formación escrituraria, en pugna con sus propios hábitos dialectales.

Este particular proceso de transmisión informativa convierte a las declaraciones de testigos y cirujanos y a los inventarios de bienes en documentos válidos para, por un lado, reconstruir diacrónicamente la oralidad[9] y, por otro, para estudiar la variación histórico-dialectal (Morala 2018), puesto que todos los documentos se generan en un lugar determinado identificado inequívocamente en los manuscritos y con intervención de informantes y escribanos que proceden, por lo general, de la región donde se han producido los hechos documentados.

Los tres tipos textuales de *ODE* presentan las características descritas en los siguientes apartados.

### 2.1. Declaraciones de testigo en juicios penales

La mayoría de los juicios transcritos tienen relación con delitos de robos, asesinatos, peleas, riñas de vecinos, injurias, estafas, violaciones o abusos de poder, procesos que se han seleccionado por reproducir situaciones en las que es fácil que los testigos narren acontecimientos con fuerte implicación emocional; además, con frecuencia recurren al estilo directo para dar mayor veracidad a sus testimonios (ejemplo (2)).

En *ODE* se recogen dos subtipos textuales relacionados con los juicios criminales: las preguntas del interrogatorio (ejemplo (1)) y las declaraciones de testigo propiamente dichas (ejemplo (2)):

> (1) Por las preguntas siguientes se an de examinar los testigos que fueren presentados por parte de el licenciado Xpoual Sanchez de Escouar, venefiziado de la uilla de Guadahortuna, y de Ysauel de Escouar, su sobrina, en el pleito con Francisco de la Cueua, alcalde de la hermandad de la dicha uilla: 1 Primeramente si conozen a las partes y tienen notizia de este pleyto. 2 Yten si sauen que siendo la dicha Ysauel de Escouar donçella de grande onestidad y rrecojimiento y criada y recojida en casa de el dicho su tio, la a solicitado el dicho Francisco de la Cueua de dos años a esta parte paseando por la calle y acudiendo a la iglessia donde la susodicha iba y haziendole seña y procurandole hablar con color de cassarse con ella; digan ettcetera (*ODE*, GR1620D9068).

> (2) este testigo una noche paseando con el dicho Françisco de la Cueva, llegaron a tratar cossas de mugeres, y le dijo a este testigo: “No saueis como esta noche e estado con una donçella y tratado carnalmente con ella”, y este testigo le dixo que le dijese quien hera, el qual no lo quisso decir. Pero este testigo, como sauia los amores que trataua con la dicha Ysauel de Escouar, creyo y tubo por çierto que hera ella la persona con quien auia estado (*ODE*, GR1620D9068).

[9] Si el proceso de trasmisión, que es común a los tres tipos textuales, aproxima estos textos a la dimensión de lo oral, la naturaleza del contenido, que es diferente en cada caso, permite establecer grados a lo largo de esta dimensión: las expresiones en estilo directo incluidas con frecuencia en las declaraciones de testigos conforman seguramente el subcorpus de *ODE* más próximo al discurso hablado (V. la Figura 10).

### 2.2. Inventarios de bienes

El segundo tipo textual está representado por los inventarios de bienes (ejemplo (4)), entendiendo como tal «cualquier texto hecho con la finalidad de enumerar, de la forma más minuciosa posible, los bienes de una persona o una institución» (Morala 2012: 200). En la propuesta de tipología documental de la red *CHARTA* (2014) los inventarios se incluyen en el grupo 5, integrado por textos «de sintaxis poco elaborada y habitualmente de estructuras repetitivas», entre los que se encuentran inventarios *post mortem*, listados, almonedas, testamentos, codicilos, mandas testamentarias, cuentas, deslindes, amojonamientos, registros de navíos, repartos de herencia, etc. A estos tipos textuales deben sumarse las cartas de dote, incluidas en el grupo 8 «certificaciones», es decir, «documentos de carácter semipúblico emitidos por notarios, escribanos o personas autorizadas». Deben incluirse también las donaciones y los intercambios de bienes (grupo 2, «cartas de compraventa y contratos»). Por último, pueden añadirse los embargos judiciales (ejemplo (3)) y los recibos.

(3) Dilixencia buscando a Phelipe Ruiz y enbargo de bienes. E luego incontinenti, en el dicho dia beinte y quatro de abril de dicho año, para mas justificacion desta causa Pedro Ximenez, alcalde, asistido de mi el escrivano, se paso a casas de Phelipe Ruiz, que son las de Xpl Ruiz, su padre. Y, preguntandole por dicho su hixo, respondio no aber parezido desde anoche, por cuia razon, y por la culpa que puede resultar, y por quenta de lo que le puede tocar de su lexitima, se enbargaron los siguientes: vna artesa con dos zedazos; su tabla, tendido, baretas y tabla; vna caldera; vna mesa pequeña de pino; vna tarima de pino de quatro tablas; vn cofre con su zerradura y llaue; vna almirez con su mano; vna sarten (*ODE*, GR1713I9014).

(4) Digo yo Franco Gimenez v$^{o}$ deste lugar de Turon que hallandome falto de vista para no poderme gobernar determino yrme con mi hijo Andres a su casa bajo las condiciones siguientes: me ha de mantener y cuidar segun sus proporciones, no he de ser hobligado haunque mejore de vista ha ir al canpo pues ya mi edad no lo permite; en la casa hayudare en lo que pueda y no mas. Yo llevo cama y ropa de mi vso y ademas le entrego los cortos bienes muebles y efectos que tengo en mi casa, ha saver: dos fanegas y nuebe zelemines de trigo garvanzos, tres cuartillos, sal, celemin y medio y medio quartillo, vna sarten remendada, otra pequeña mediada, vn caciquillo chico, vna almirez con su mano, vnas treves mas grandes y otras chicas, vnas tenazas, una rasera que no tiene pala, vn bail, cuatro espuertas de ha fanega, tres pequenas, cuatro paneras, vn seron terrero, vn porron, dos zalonas (*ODE*, GR1829I2005).

Los inventarios contribuyen al conocimiento de las condiciones de vida del pasado (como se observa en el ejemplo (4)) y permiten profundizar en la realidad cotidiana de las sociedades antiguas. Al mismo tiempo, son una fuente complementaria de primer nivel para el estudio histórico del léxico, puesto que permiten documentar vocabulario dialectal[10] no habitual en corpus generales (Morala 2012: 200). Las principales ventajas de un corpus de estas características son las siguientes:

[10] V. los casos de *treves* 'trébedes', *bail* 'badil', *rasera*, *serón*, *zalona*, etc. del ejemplo 4, o *molledo* y *almadraqueja* en el ejemplo (5).

– La perfecta e inequívoca datación y localización geográfica de los documentos.
– La abundancia de esta documentación en los archivos de protocolos e histórico-provinciales.
–La homogeneidad y comparabilidad del tipo textual, similar a las encuestas con las que se confeccionan atlas etnográficos y dialectales (Morala 2012: 202).

### 2.3. Certificaciones de cirujanos y sangradores

Los pleitos y probanzas criminales pueden incluir declaraciones médico-legales llevadas a cabo por cirujanos, barberos o sangradores, a veces denominadas «declaraciones de esencia» (Calderón Campos 2018). Son certificaciones médicas en las que estos facultativos dictaban al escribano de turno una descripción minuciosa de las heridas sufridas por las víctimas de la agresión que estaba siendo investigada. Se trata de un tipo textual interesantísimo para estudiar el léxico médico que se empleaba en el día a día de la práctica terapéutica, porque recoge tanto las voces populares (V. *molledo* 'bíceps' en el ejemplo (5)) como los tecnicismos médicos que servían para nombrar o describir las partes del cuerpo o las dolencias de las víctimas.

(5) Y le mandaron le reconoçiese sus heridas, el qual hauiendo llegado a el lo allo bibo y quejandose y, hauiendolo desnudado en vna almadraqueja donde estaua acostado, bio tenia en el braço y costado derecho vna herida de forma que tenia pasado el braço por el molledo, y correspondiente a esta entraba en el costado, dadas a el parecer con ynstrumento de fuego y de bala redonda y de escopeta, que le penetraua el costado (*ODE*, GR1700C9001).

Cada tipo textual facilita el análisis de algún nivel lingüístico: los inventarios de bienes sirven preferentemente para estudiar el léxico de la vida cotidiana, el que se suele encontrar en los atlas lingüísticos y etnográficos; las certificaciones médicas complementan a los inventarios proporcionando designaciones anatómicas populares y su contraste con la terminología técnica incipiente; por último, las declaraciones de testigo permiten analizar en mayor profundidad que los textos anteriores la evolución de las estructuras morfosintácticas del español y de sus formas de tratamiento.

Pero además, de manera secundaria, los inventarios pueden utilizarse en dialectología histórica para estudiar el cambio fonético[11] o morfológico[12]; y lo mismo puede decirse de las declaraciones de testigo, en las que rasgos fonéticos, léxicos o morfológicos vernaculares afloran ocasionalmente (Calderón Campos 2015: 37-74).

## 3. LIMITACIONES TECNOLÓGICAS DEL *CORDEREGRA*

Cuando se terminó de transcribir la documentación del *CORDEREGRA* descubrimos que habíamos llegado a un punto muerto desde el que no se podía avanzar tecnológicamente. Las transcripciones realizadas, en formato *Microsoft Word*, seguían la estructura que se muestra en (6):

(6)

1620
Guadahortuna (Granada)
ARChGr 9582/41
Título: Denuncia contra Francisco de la Cueva, por estupro
Resumen: Información sumaria a petición del licenciado Cristóbal Sánchez de Escobar y su sobrina, Isabel de Escobar, vecinos de Guadahortuna, contra Francisco de la Cueva, alcalde de la hermandad de dicha villa, al que denuncia por estupro.

{h 8v} {1} con uiolenzia y fuerza que le hizo la estrupo y co{2}nozio carnalm<en>te y vbo su virginidad debaxo {3} de que le auia dado la d<ic>ha palabra de casam<ien>to {4} digan et<cé>t<er>a.

[11] Un estudio sobre el seseo en el siglo XVII (Calderón Campos 2019: 116-119), realizado con datos del *Corpus Léxico de Inventarios* (*CorLexIn*), revela que la frecuencia mayor de confusiones gráficas indicadoras de seseo (*asul* por azul, *haser* por hacer, etc.) se da en Andalucía (Huelva, Cádiz, Málaga, Sevilla y Córdoba), en Tenerife y en América (Puerto Rico, El Salvador, Bolivia, Colombia, México). En el resto de las provincias españolas, los escribanos no "confunden" estas grafías. Con los datos obtenidos de este estudio puede afirmarse que las pretendidas confusiones gráficas de Andalucía, Canarias y América deben interpretarse como pruebas de seseo y no como meros errores ortográficos, pues no tendría sentido pensar, en caso de que no tuvieran un respaldo fonético, que solo cometieran esos errores los escribanos de Cádiz, por ejemplo, y nunca los de Valladolid. Se confirma, además, que aun sin conocer el lugar de nacimiento de los escribanos, estos en su mayoría tienen que ser originarios de la zona donde trabajaban, pues su conducta lingüística revela, en lo referente al seseo, un patrón lingüístico compatible con el de la procedencia geográfica de los manuscritos. Los mismos inventarios de bienes del *CorLexIn* han servido para trazar la isoglosa de la aspiración de F- en el siglo XVII (Morala y Perdiguero 2019) o para demostrar, en el caso del vocalismo tónico (Pérez Toral 2017), que determinadas variantes gráficas permiten documentar el proceso de monoptongación de algunos diptongos (*fleco* < *flueco*, *frente* < *fruente*), o el rechazo de los hiatos, con distintos procedimientos (*tualla*, *almuada*, *rial*, *toballa*, *ral* 'real', etc.).

[12] Egido (2019: 128), en un estudio sobre el neutro de materia, reivindica el uso de los inventarios de bienes como fuente para estudiar la variación geográfica de los cambios gramaticales, es decir, para profundizar en aspectos poco tratados en Dialectología Histórica (v. también Egido en este volumen). Para un estudio de los diminutivos en *-ico* a partir de inventarios, puede verse Calderón Campos (2019: 119-122) y la tesis doctoral en curso de Arrabal Rodríguez (en prep.), a partir de datos de *ODE*. V. Egido (en este volumen: §4.2.4) para un estudio de este diminutivo en asturleonés y aragonés con el *CorLexIn*.

{5} [margen izquierdo: 5] Yten si sauen que al tiempo y quando entro en la {6} casa del d<ic>ho liz<encia>do escouar y estrupo a la d<ic>ha su {7} sobrina, estaua el sussod<ic>ho en la ciu<da>d de Gra<na>da {8} curandose de vna graue enfermedad que tenia {9} digan et<cé>t<er>a.
{10} [margen izquierdo: 6] Yten si sauen que despues del d<ic>ho estrupo a contin[ua]{11}do el d<ic>ho fran<cis>co de la cueua [sic: el] amistad con la {12} d<ic>ha Ysauel de escouar por algunos dias haziendos[e] {13} Regalos el uno al otro con nota escandalo y mur{14}muracion de los veçinos de la d<ic>ha uilla digan {15} et<cé>t<er>a

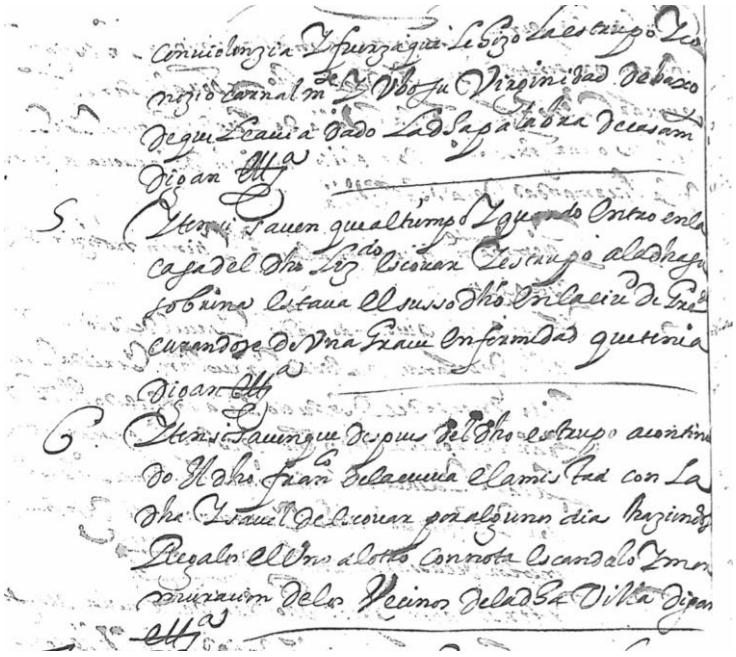

Figura 1. Facsímil del documento de Guadahortuna, 1620.

A simple vista, para la mente humana, los documentos de *CORDEREGRA* seguían una estructura comprensible y racional: en primer lugar, se visualizaban por separado los metadatos y la transcripción propiamente dicha. Cada línea de la cabecera se reservaba para un dato sobre el manuscrito: fecha, lugar de redacción, signatura, título y resumen; de la misma forma, cualquier lector podía interpretar intuitivamente el significado de algunas convenciones empleadas para ofrecer información paratextual (es decir, relacionada con los elementos visuales que rodean al manuscrito) o información sobre intervenciones del editor (expansiones de abreviaturas, conjeturas, lecturas dudosas, aclaraciones, etc.):

| Información paratextual | Símbolo | Ejemplo |
|---|---|---|
| Inicio de página | {h nºr/v} | {h 8v} |
| Inicio de línea | {nº} | {3} |
| Adición al margen | [margen: texto añadido al margen] | [margen izquierdo: 5] |
| Expansión de abreviaturas | <expansión de la abreviatura> | et<cé>t<er>a |
| Conjetura del editor | [texto conjeturado] | contin[ua]{11}do |
| Texto incorrecto o inexacto | [sic: texto incorrecto o inexacto] | [sic: el] |

Tabla 1. Algunas convenciones editoriales de *CORDEREGRA*.

Sin embargo, la estructuración de los documentos no se hacía a nivel computacional: informáticamente, los ordenadores se encontraban con una cadena de caracteres no estructurados, lo que dificulta, cuando no impide, una recuperación eficaz de la información del corpus.

Por ejemplo, si estamos interesados en un estudio histórico de la palabra *estupro*[13], el buscador de *Microsoft Word* nos daría dos resultados para el documento de 1620 de Guadahortuna, uno procedente del título («Denuncia contra Francisco de la Cueva, por estupro») y otro del resumen («al que denuncia por estupro»). Con este resultado obtenemos una visión sesgada de la realidad lingüística del siglo XVII, puesto que ambos ejemplos son achacables a los editores del corpus, no al escribano de la Real Chancillería de Granada, donde se guarda el documento. El problema, obviamente, procede de la incapacidad de restringir la búsqueda a la parte textual correspondiente a la transcripción, dado que, como se ha visto, los datos no están estructurados.

Además, desde el punto de vista lingüístico, la búsqueda de la variante estándar *estupro*, que es la que cualquier usuario actual haría, no obtendría la forma popular *estrupo*, que sí utilizó el escribano («despues del d<ic>ho estrupo»). Si extendemos la consulta a todo el corpus *CORDEREGRA*, en su versión de 2015, se obtienen 17 ejemplos, contabilizando los casos del sustantivo (*estupro*/*estrupo*) y del verbo (*estuprar*/*estrupar*), repartidos de la siguiente forma: 11 casos textuales, en los que predomina la variante popular con metátesis (*estrup*-), de la que se registran 7 ejemplos, por 4 de *estupr*-. En los metadatos se registra siempre, como era de esperar, la forma normativa.

Los pocos ejemplos registrados en otros corpus de documentación manuscrita son también de la variante vulgar, que parece que se imponía a la forma etimológica procedente de STUPRUM:

(7) En consequencia, siendo mi cuñada a quien ha *estrupado*, y se halla en seis meses de su preñez, debe de hacerse la justicia con mas rigor que lo que dha ley previene (*P.S. Post Scriptum*, PSCR5753, 1820, Carta de José de Escalante, comerciante, para Santos Ruiz Marqué, escribiente).

(8) ... que rindan las armas y sean despoxados de ellas y se pegue fuego a las estufas que son receptaculos del domonio, cassas de ydolatria, *estrupos* y obsenidades (*Cibola Project*, Antonio de Otermín's Attempted Reconquest of New Mexico Winter 1681-82, transcripción de Jerry R. Craddock).

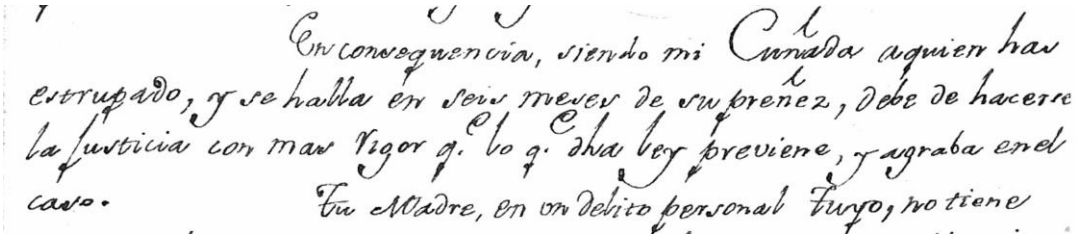

Figura 2. Facsímil correspondiente al ejemplo (7).

[13] Esta denominación ha desaparecido del vigente Código Penal. Históricamente, se ha empleado para designar el delito consistente en tener relaciones sexuales con un menor, mediante engaño o aprovechándose de alguna situación de superioridad, o bien, sin consentimiento de la víctima y causándole algún daño moral (*DEJ* 2016: s.v. *estupro*).

Así pues, los escribanos de la época parecían preferir la variante con metátesis, pero esta información se nos escapa si el buscador es incapaz de separar los metadatos de la información textual.

Otro problema que nos planteaba *CORDEREGRA* se relaciona con la enorme variabilidad ortográfica de los documentos antiguos, lo que equivale a tener que resolver la cuestión de cómo encontrar todas las formas ortográficas de una misma palabra, especialmente aquellas poco previsibles, como podría ser el caso de *bail* 'badil', *treves* 'trébedes' o *haunque* del ejemplo (4). Por ejemplo, en *CORDEREGRA* la palabra *vecino* aparece transcrita de ocho formas distintas, sin contar posibles abreviaturas: *veçino*, *ueçino*, *beçino*, *vezino*, *uezino*, *bezino*, *vesino* y *besino*. El problema es similar con palabras como *Isabel* (*Ysabel*, *Isauel*, *Ysavel*, etc.), *hubo* (*ubo*, *vbo*, *uvo*, etc.), *violencia* (*uiolenzia*, *violençia*), etc. La cuestión, con el modelo adoptado, no tenía otra solución que multiplicar las búsquedas por el número de variantes posibles.

Por otra parte, el verbo *estuprar* presenta en *CORDEREGRA* dos variantes morfológicas (participio y pretérito perfecto simple) y tres ortográficas: *estrupado*, *estuprado* y *estupro* ('estupró'), imposibles de recuperar en una sola consulta, al no disponer de un sistema de lematización compatible con el modelo de edición empleado.

Además, si de nuevo volvemos al documento de Guadahortuna (1620), encontramos transcripciones en las que una palabra se corta con códigos paratextuales que informan del inicio de línea (co{2}nozio), de una conjetura y un inicio de línea (contin[ua]{11}do), o de una expansión de la abreviatura (carnalm<en>te, Gra<na>da). La superposición de elementos textuales y paratextuales provoca pérdida de información, puesto que las palabras marcadas con este procedimiento no son recuperables en búsquedas posteriores.

Por último, en *CORDEREGRA*, siguiendo el modelo de la red internacional *CHARTA*, se quería hacer una triple presentación del documento: dos ediciones del manuscrito (paleográfica y normalizada) acompañadas del facsímil. La única solución tecnológica a nuestro alcance era presentar en un mismo PDF las tres visualizaciones, con una maquetación manual muy complicada y estática.

En definitiva, los límites tecnológicos de *CORDEREGA* en 2015 pueden resumirse en las siguientes preguntas, a las que se responde en el apartado 4:

1. ¿Cómo limitar las búsquedas de palabras al contenido textual, separando los metadatos de las transcripciones? Aplicado al ejemplo anterior, se trataría de encontrar los ejemplos textuales de *estupro*, sin mezclarlos con los metatextuales.

2. ¿Cómo hacer una edición filológica digital, en la que se incluya información sobre aspectos visuales del documento (por ejemplo, cambios de línea) o adiciones del editor (por ejemplo, expansiones de abreviaturas o conjeturas) sin provocar pérdidas en la recuperación de la información textual? En nuestro ejemplo, esta pregunta equivale a poder recuperar formas como *contin[ua]{11}do* o *carnalm<en>te*.

3. ¿Cómo recuperar, sin multiplicar las consultas, todas las variantes ortográficas de una palabra (*uvo*, *uuo*, *huvo*, etc.), incluso las menos previsibles, como *treves* o *bail*?

4. ¿Cómo lematizar y etiquetar morfosintácticamente el corpus, para recuperar todas las formas del paradigma de un verbo, que además presenta variación ortográfica? ¿Y por otro lado, cómo combinar la búsqueda de un lema con una determinada etiqueta morfosintáctica, lo que en nuestro ejemplo equivaldría a poder obtener todos los casos del pretérito perfecto simple del verbo *estuprar*, teniendo en cuenta que puede presentar distintas variantes ortográficas (*estupro*, *estupró*, *estrupo*, *estrupó*), o del verbo *haber*, donde la variabilidad ortográfica es aún mayor (*hubo*, *uuo*, *huuo*, *ubo*, etc.)?

5. ¿Cómo presentar en línea distintas ediciones (semipaleográfica y normalizada) de un mismo documento, incluida su imagen facsimilar?

6. ¿Cómo hacer todo lo anterior sin necesidad de delegar la gestión tecnológica del corpus en expertos informáticos externos al proyecto de investigación?

## 4. RESOLUCIÓN DE LAS LIMITACIONES TECNOLÓGICAS DEL *CORDEREGRA*

Como se ha explicado en la introducción, *ODE* se diferencia de *CORDEREGRA* en tres aspectos fundamentales: ampliación de la representación geográfica, ampliación del arco cronológico y renovación tecnológica del corpus. El presente apartado se centra en este último aspecto, incidiendo en las mejoras tecnológicas que permiten dar una respuesta satisfactoria a las preguntas anteriores y que, en definitiva, han supuesto un salto cualitativo de *ODE* respecto a su predecesor.

Aunque son muchas y variadas las funcionalidades que ofrece *ODE*, todas ellas parten en último término de la adopción de dos estrategias básicas que son, además, complementarias: la marcación de datos en lenguaje *XML-TEI*, por un lado, y la utilización de la plataforma *TEITOK*, por otro lado.

Las directrices *TEI*, basadas en el lenguaje de marcado *XML*, ofrecen un estándar para la representación de textos en formato digital ampliamente difundido en el ámbito de las Humanidades Digitales. Su adopción en *ODE* no solo permite contar con datos estructurados para facilitar su posterior recuperación, sino que garantiza su preservación, su reutilización o su posible integración en repositorios digitales, entre otras ventajas conocidas (Burnard 2014).

La plataforma *TEITOK* ofrece un sistema web para publicar, explotar y editar corpus basados en lenguaje *XML-TEI*. Su funcionalidad es doble: por un lado, permite al usuario navegar a través de una edición digital y, simultáneamente, crear un corpus anotado para realizar búsquedas; por otro lado, permite al administrador editar los datos en *XML-TEI* y aplicar herramientas de procesamiento automático de textos, desde la tokenización hasta la anotación lingüística del corpus. En palabras de su creador:

> *TEITOK* is a web-based system for viewing, creating, and editing corpora with both rich textual mark-up and linguistic annotation. For visitors, the system provides a graphical user interface in which the annotated document can be visualized in a number of different ways,

depending on what the user is interested in. And for administrators of the corpus, TEITOK uses the same interface to allow easy editing of the underlying XML document, meaning administrators can correct their corpus while they are consulting it (Janssen 2014).

La utilidad de *TEITOK* ha sido demostrada en numerosos proyectos de investigación y en la creación de corpus de diferente naturaleza y tamaño[14]. Desde el punto de vista tecnológico, ambos métodos de trabajo –marcación en *XML-TEI* y uso de la plataforma *TEITOK*– constituyen la piedra angular de *ODE* para resolver los problemas que cierran el apartado anterior, como se explica a continuación.

### 4.1. Texto y metatexto

Volviendo al ejemplo recogido en (6), es obvio que para el ojo humano resulta fácil distinguir, a partir de una rápida lectura, qué datos constituyen el contenido del texto y qué datos describen o complementan el texto con información de diferente tipo (año, lugar, título, etc.). Repárese, no obstante, en que se trata de una distinción implícita que la mente humana establece a partir de conocimientos previos sobre lo que es un texto o sobre el carácter metatextual de un título, un año, un lugar o una referencia archivística. Para que el ordenador «comprenda» esta distinción, esto es, para poder procesar computacionalmente datos de diferente naturaleza, es necesario que esa distinción –o cualquier otra que se considere relevante– se haga explícita: es aquí donde entra en juego el lenguaje *XML-TEI*.

```
<TEI>
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>Denuncia contra Francisco de la Cueva, por estupro</title>
                ...
            </titleStmt>
            ...
            <sourceDesc>
                <msDesc>
                    <msIdentifier>
                        ...
                        <idno>ARCHGR 9582/41</idno>
                    </msIdentifier>
                    <msContents>
                        <summary>Información sumaria a petición del licenciado Cristóbal
                        Sánchez de Escobar y su sobrina, Isabel de Escobar, vecinos de
                        Guadahortuna, contra Francisco de la Cueva, alcalde de la
                        hermandad de dicha villa, al que denuncia por estupro.</summary>
                        ...
                    </msContents>
                    ...
                </msDesc>
            </sourceDesc>
        </fileDesc>
        ...
        <profileDesc>
        ...
            <settingDesc>
                <setting>
                    <name>España, Granada, Guadahortuna</name>
                    <date>1620</date>
                </setting>
            </settingDesc>
        </profileDesc>
        ...
    </teiHeader>
    <text>
        <pb n="8v"/>
        <lb n="1"/> con uiolenzia y fuerza que le hizo la estrupo y co<lb n="2"/>nozio
        carnalm<ex>en</ex>te y vbo su virginidad debaxo <lb n="3"/> de que le auia dado
        la d<ex>ic</ex>ha palabra de casam<ex>ien</ex>to
        ...
    </text>
</TEI>
```

Figura 3. Documento *XML-TEI* correspondiente al ejemplo (6).

[14] V. el apartado «Projects» en Janssen (2014) para una lista no exhaustiva.

Scriptum digital
Vol. 9 (2020), pp. 167-189

La estructura general de un documento en *XML-TEI* consta de dos elementos básicos que están pensados precisamente para marcar la distinción que venimos comentando: un <teiHeader>, que contiene el conjunto de metadatos que describen el texto, y un <text>, que contiene el texto propiamente dicho. Ambos pueden contener a su vez diferentes elementos, tantos como sean necesarios hasta obtener unos datos estructurados dentro del documento. Por tanto, la información del ejemplo (6) se organizaría en *ODE* conforme a la estructura que se recoge en la Figura 3:

Las ventajas de contar con datos así estructurados son evidentes. Con esta metodología, resulta factible recuperar las ocurrencias de *estupro* en el texto y solo en el texto, es decir, las ocurrencias registradas dentro del elemento <text>; y resulta igualmente posible consultar las ocurrencias de *estupro* en el título (<title>) o en el resumen (<summary>), o en documentos de Granada datados en 1620. La distinción entre texto y metatexto se ha hecho explícita y, por tanto, se puede procesar eficazmente en función de la información que sea requerida en la consulta.

### 4.2. Tokenización de la edición digital

La marcación *XML-TEI* que se recoge en la Figura 3 se ha simplificado por razones de claridad. En realidad, en *ODE* se establece una estructura sobre los metadatos algo más completa que la expuesta en dicha figura y, además, se hace uso de un conjunto de marcas sobre los datos textuales para dejar constancia de diversa información paratextual (inicios de página y línea, adiciones, cancelaciones, expansiones de abreviaturas, conjeturas, etc.). En cierto modo, este tipo de información ya se marcaba explícitamente en *CORDEREGRA* mediante las convenciones editoriales propuestas por la red *CHARTA* (V. la Tabla 1), aunque su finalidad era primordialmente editorial, no computacional.

Para este último propósito, las marcas tipográficas basadas en los criterios CHARTA fueron convertidas a elementos basados en las directrices *TEI* y, en consecuencia, los archivos en formato *Microsoft Word* de *CORDEREGRA* dieron lugar a archivos *XML* en *ODE*. La Tabla 2 recoge, a modo de ejemplo, algunas correspondencias entre ambos estilos de marcas:

| Información paratextual | *CORDEREGRA* (*CHARTA*) | *ODE* (*XML-TEI*) |
|---|---|---|
| Inicio de página | {h 8v} | <pb n="8v"/> |
| Inicio de línea | {3} | <lb n="3"/> |
| Adición al margen | [margen izquierdo: 5] | <add place="left">5</add> |
| Expansión de abreviaturas | et<cé>t<er>a | et<ex>cé</ex>t<ex>er</ex>a |
| Conjetura del editor | contin[ua]{11}do | contin<supplied>ua</supplied><lb n="11"/>o |
| Texto incorrecto o inexacto | [sic: el] | <sic>el</sic> |

Tabla 2. Algunas correspondencias entre marcas de *CORDEREGRA* y *ODE*.

Siguiendo con el ejemplo que nos ocupa y aplicando las directrices *TEI* a la información paratextual de dicho ejemplo, el resultado sería la transcripción en lenguaje *XML-TEI* que se recoge en la Figura 4.

Nótese que las marcas *XML-TEI* señalan cuestiones que puedan afectar a palabras enteras (<sic>el</sic>) o a conjuntos de caracteres dentro de una misma palabra

Scriptum digital
Vol. 9 (2020), pp. 167-189

(contin<supplied>ua</supplied><lb n="11"/>o). Casos como este último constituyen un problema en términos de recuperación de información, puesto que las marcas *XML-TEI* impiden establecer una correspondencia entre la palabra buscada (*continuado*) y el conjunto de caracteres realmente transcrito (contin<supplied>ua</supplied><lb n="11"/>o). Este problema, que ya existía en *CORDEREGRA* como consecuencia de la aplicación de convenciones editoriales *CHARTA* dentro de una misma palabra (*contin[ua]{11}do*), se resuelve en *ODE* durante el proceso de tokenización de los datos en *XML-TEI* a través del sistema *TEITOK*.

```
<text>
    <pb n="[8v]"/>
    <lb n="1"/> con uiolenzia y fuerza que le hizo la estrupo y co<lb n="2"/>nozio
    carnalm<ex>en</ex>te y vbo su virginidad debaxo <lb n="3"/> de que le auia dado la
    d<ex>ic</ex>ha palabra de casam<ex>ien</ex>to; <lb n="4"/> digan
    et<ex>cé</ex>t<ex>er</ex>a. <lb n="5"/> <add place="left">5</add> Yten si sauen que
    al tiempo y quando entro en la <lb n="6"/> casa del d<ex>ic</ex>ho liz<ex>encia</ex>do
    Escouar y estrupo a la d<ex>ic</ex>ha su <lb n="7"/> sobrina, estaua el
    sussod<ex>ic</ex>ho en la ciu<ex>da</ex>d de Gra<ex>na</ex>da <lb n="8"/> curandose de
    vna graue enfermedad que tenia; <lb n="9"/> digan et<ex>cé</ex>t<ex>er</ex>a.
    <lb n="10"/> <add place="left">6</add> Yten si sauen que despues del d<ex>ic</ex>ho
    estrupo a contin<supplied>ua</supplied><lb n="11"/>do el d<ex>ic</ex>ho
    Fran<ex>cis</ex>co de la Cueua <sic>el</sic> amistad con la <lb n="12"/>
    d<ex>ic</ex>ha Ysauel de Escouar por algunos dias, haziendose <lb n="13"/> regalos el
    uno al otro con nota y escandalo y mur<lb n="14"/>muracion de los veçinos de la
    d<ex>ic</ex>ha uilla; digan <lb n="15"/> et<ex>cé</ex>t<ex>er</ex>a.
</text>
```

Figura 4. Transcripción XML-TEI del texto correspondiente al ejemplo (6).

La tokenización consiste en la identificación y marcación de tokens, esto es, palabras y signos de puntuación. En *TEITOK*, este proceso se realiza automáticamente: el sistema añade un elemento <tok> y un identificador único a cada token dentro del archivo *XML*. Por ejemplo, los cuatro primeros tokens del fragmento de la Figura 4 se marcarían como se recoge en (9):

(9) <tok id="w-257">con</tok> <tok id="w-258">uiolenzia</tok> <tok id="w-259">y</tok>
<tok id="w-260">fuerza</tok>

Si el token en cuestión incluye elementos *XML* en su interior, el sistema extrae la palabra correspondiente, esto es, el conjunto de caracteres excluyendo la marcación *XML-TEI*, y lo almacena en un atributo específico (@form) dentro del elemento <tok>. Por ejemplo, la palabra *continuado* se marcaría como se recoge en (10):

(10) <tok id="w-342" form="continuado">contin<supplied>ua</supplied><lb n="11"/>o</tok>

Finalmente, en caso de que el token incluya información aportada por el editor relativa al desarrollo de una abreviatura, el sistema extrae tanto la forma original que aparece en el manuscrito (@form) como la forma expandida (@fform, *full form*) ya sin marcación *XML*. Por ejemplo, el resultado de la tokenización sobre la palabra *carnalmente*, transcrita con expansión sobre la abreviatura *carnalmte*, sería el que se ofrece en el ejemplo (11):

(11) <tok id="w-268" form="carnalmte" fform="carnalmente">carnalm<ex>en</ex>te</tok>

El corpus *ODE* se genera a partir del contenido almacenado en el conjunto de archivos *XML*, esto es, a partir de las transcripciones textuales –incluyendo metadatos– realizadas en lenguaje *XML-TEI*. La tokenización constituye un paso esencial en este proceso de creación del corpus: el contenido del elemento <tok> incluye siempre la transcripción *XML-TEI* de cada token, mientras que sucesivos atributos almacenan, si fuese necesario, la información correspondiente a ese token sin marcas *XML* para su correcta recuperación en el corpus.

### 4.3. Normalización ortográfica del corpus

Es un hecho que los documentos antiguos suelen presentar variabilidad ortográfica. Esta variabilidad será más o menos acusada en función de diversos factores, como puede ser el tamaño del corpus, el género textual, la procedencia geográfica o el arco cronológico, pero en todo caso se trata de un rasgo consustancial a la documentación que integra los corpus históricos. Desde el punto de vista filológico, la representación gráfica de una palabra tiene un interés evidente y por eso debe ser respetada en la transcripción del manuscrito original; desde la perspectiva de la lingüística de corpus, sin embargo, esta variabilidad constituye un escollo para obtener resultados precisos en el análisis de frecuencias, en la etiquetación morfosintáctica automática o en la recuperación de información (Baron *et al.* 2009).

Para solucionar este problema, el sistema *TEITOK* incorpora un normalizador ortográfico que asocia la forma original de cada token con su correspondiente forma en grafía normalizada según el estándar actual. El resultado de este proceso se almacena dentro de un atributo @nform (*normalized form*). Por ejemplo, el token *uiolenzia* del ejemplo (9) aparece marcado como muestra el ejemplo (12) una vez aplicado el normalizador ortográfico de *TEITOK*:

(12) <tok id="w-258" nform="violencia">uiolenzia</tok>

Esta estrategia presenta varias ventajas. En primer lugar, permite recuperar en una única consulta todas las ocurrencias de una palabra, con independencia de cómo aparezca escrita en el corpus (Figura 5):

| | | |
|---|---|---|
| contexto | de Escouar, adonde \| \| con | **uiolenzia** y fuerza que le hizo |
| contexto | casa estando ardiendo \| con gran | **violenzia** ; asi vn almiar de |
| contexto | vltraxada la jussa y la | **biolençia** que le querian \| azer al |
| contexto | offio. \| Y pretendio con | **biolençia** quitarle la dha \| bara al |
| contexto | amutinada y que pretendian con | **biolençia** quitarle la bara \| al dho |
| contexto | partiendo todos juntos con mucha \| | **biolençia** para entrar en cassa de |
| contexto | , a \| açer fuerça y | **biolençia** para querer benir a entrar |
| contexto | de \| mano armada y con | **biolençia** , para cassa de su |
| contexto | por \| justiçia y no por | **biolençia** . Y a este tienpo |
| contexto | diferentes testigos sin fuerça ni | **uiolençia** alguna, y no se |
| contexto | , y con fuerza \| y | **uiolencia** la sacaron hasta q la |

Figura 5. Concordancias de la palabra *violencia* en *ODE*. Consulta en CQP: [nform="violencia"].

En segundo lugar, permite recuperar palabras cuya grafía original es difícilmente conjeturable, como son los casos de (13):

(13) a. <tok id="w-1398" nform="trébedes">trebes</tok>
b. <tok id="w-66" nform="badil">bail</tok>
c. <tok id="w-1867" nform="aunque">haunque</tok>

En tercer lugar, se puede realizar la consulta de una palabra por su forma normalizada y agrupar el resultado por orden de frecuencia de sus variantes ortográficas. Es lo que hemos hecho con las palabras *violencia*, *vecino* e *Isabel*, cuya variación ortográfica en el corpus *ODE* aparece recogida en la Tabla 3:

| Grupo de formas (*violencia*) | Nº | Grupo de formas (*vecino*)[15] | Nº | Grupo de formas (*Isabel*) | Nº |
|---|---|---|---|---|---|
| *biolençia* | 7 | *vezo* | 47 | *Ysauel* | 31 |
| *violenzia* | 1 | *vezino* | 35 | *Ysabel* | 6 |
| *uiolençia* | 1 | *vecino* | 24 | *Isavel* | 6 |
| *uiolenzia* | 1 | *vo* | 20 | *Isabel* | 2 |
| *uiolencia* | 1 | *vzo* | 10 | *Ysavel* | 1 |

Tabla 3. Variación ortográfica de las palabras *violencia*, *vecino* e *Isabel* en *ODE*.

Finalmente, contar con un nivel de normalización ortográfica tiene aún dos ventajas adicionales. Por un lado, facilita la lectura de documentos antiguos al usuario no familiarizado con este tipo de textos; por otro lado, reduce ostensiblemente los errores producidos por los etiquetadores automáticos, facilitando así al investigador la revisión manual de la anotación lingüística (Sánchez-Marco *et al.* 2012).

### 4.4. Anotación morfosintáctica y lematización

Además de un normalizador ortográfico, el sistema *TEITOK* incorpora un etiquetador morfosintáctico llamado *NeoTag* (Janssen 2012). De manera automática, *NeoTag* asocia una etiqueta morfosintáctica y un lema a cada token del corpus y, además, utiliza los propios datos etiquetados –una vez revisados manualmente– como corpus de entrenamiento, lo que significa que mejora a medida que el corpus aumenta de tamaño.

De forma análoga al proceso de normalización ortográfica, el resultado de *NeoTag* se almacena en atributos dentro del elemento <tok>. La etiqueta morfosintáctica se guarda en un atributo @pos y el lema correspondiente a cada token se almacena en un atributo @lemma. Volviendo nuevamente a la palabra *uiolenzia*, el resultado final una vez procesado este token en *TEITOK* es el que se muestra en (14):

(14) <tok id="w-258" nform="violencia" pos="NCFS000" lemma="violencia">uiolenzia</tok>

El sistema de etiquetas utilizado para anotar el corpus *ODE* está basado en la propuesta del grupo *EAGLES* para la anotación morfosintáctica de lexicones y corpus para

[15] En el caso de la forma normalizada *vecino*, se recogen solo las 5 posibilidades ortográficas más frecuentes. Actualmente, en *ODE* se registran 21 formas ortográficas diferentes para esta palabra, contando abreviaturas.

Scriptum digital
Vol. 9 (2020), pp. 167-189

todas las lenguas europeas (Leech y Wilson 1996). El conjunto de etiquetas *EAGLES* se rige por un sistema de posiciones: cada etiqueta consta de una secuencia de letras y números, donde cada letra o número representa un rasgo morfosintáctico determinado dependiendo de su posición dentro de la secuencia. Por ejemplo, la forma *violencia* lleva la etiqueta NCFS000, que representa nombre (N), común (C), femenino (F), singular (S).

Es obvio que contar con un corpus lematizado y etiquetado facilita y multiplica las opciones de búsqueda. Una única consulta en *ODE* permite recuperar todas las formas del paradigma del verbo *haber*, por ejemplo, pese a la alta variabilidad ortográfica que pueda presentar esta forma (Figura 6):

| | | | |
|---|---|---|---|
| contexto | sanidad, y escogio como | **a** | dho de entre otras \| munchas |
| contexto | Blanco, \| y que como | **avia** | suzedido el salirse auia sido |
| contexto | esta ciudad, \| los comparecientes | **han** | determinado consignarlos por \| medio |
| contexto | lugar \| \| en dro, confiesa | **ha** | recivido del expresado Josef \| Varea su |
| contexto | compete, otorga y \| confiesa | **haber** | recivido real y efectivamente por |
| contexto | hecha, otorga y confiesa \| | **haber** | recibido rl y efectivamte de |
| contexto | los quales otorga y confiesa \| | **aver** | recibido real y efectibamente de |
| contexto | dha ciud, el qual \| | **hauiendo** | jurado, ofrecio decir verdad |
| contexto | sus \| heridas al qual | **hauiendo** | llegado a el, lo |
| contexto | por virtud del cual | **habría** | de aumentarse el capital |
| contexto | Vn cajon en el qual | **habia** | diferentes yerrezillos uiejos y asimismo |
| contexto | rl sitio; la qual | **a** | trahido a su poder \| por |

Figura 6. Muestra de concordancias del lema *haber* en *ODE*. Consulta en CQP: [lema="haber"].

La información morfosintáctica combinada con el lema permite, además, realizar consultas más detalladas. Por ejemplo, es posible recuperar únicamente las formas del pretérito imperfecto de indicativo de un verbo. La Tabla 4 muestra el resultado de dicha consulta en *ODE* para los verbos *haber* y *tener* agrupados por forma transcrita y ordenados por frecuencia descendente:

| Grupo de formas | Nº | Grupo de formas | Nº |
|---|---|---|---|
| *auia* | 27 | *tenia* | 19 |
| *avia* | 14 | *tenja* | 3 |
| *abian* | 4 | *tenya* | 2 |
| *abia* | 3 | *tenyan* | 1 |
| *hauia* | 2 | *tenjan* | 1 |
| *havia* | 1 | *tenian* | 1 |
| *habia* | 1 | | |
| *auian* | 1 | | |

Tabla 4. Variación ortográfica de *haber* y *tener* en el pretérito imperfecto de indicativo.

### 4.5. Visualización de la edición digital

Todos los aspectos comentados hasta el momento tienen que ver con la creación del corpus lingüístico, su procesamiento automático y sus posibilidades de búsqueda. No obstante, como ya se ha indicado en la introducción, la mejora de estos aspectos no va en menoscabo de los intereses filológicos de *ODE*, que ya estaban presentes en el

Scriptum digital
Vol. 9 (2020), pp. 167-189

*CORDEREGRA*. Al contrario, del salto tecnológico que supone el segundo con respecto al primero resulta una optimización evidente en este sentido. Por un lado, el paso desde las transcripciones en *Microsoft Word* a las transcripciones en lenguaje *XML-TEI* permite ofrecer una edición crítica digital de los documentos sin pérdida de rigor filológico; por otro lado, el uso de la plataforma *TEITOK* facilita la visualización de los textos, en sus diferentes versiones (Figura 8), así como la del propio facsímil, que se puede consultar en paralelo al texto digital (Figura 7):

con uiolenzia y fuerza que le hizo la estrupo y co
nozio carnalmte y vbo su virginidad debaxo
de que le auia dado la dha palabra de casamto;
digan etta.
5 Yten si sauen que al tiempo y quando entro en la
casa del dho lizdo Escouar y estrupo a la dha su
sobrina, estaua el sussodho en la ciud de Grada
curandose de vna graue enfermedad que tenia;
digan etta.

Figura 7. Edición semipaleográfica (izquierda) y facsímil (derecha).

con uiolenzia y fuerza que le hizo la estrupo y
conozio carnalmente y vbo su virginidad debaxo
de que le auia dado la dicha palabra de casamiento;
digan etcetera.
5 Yten si sauen que al tiempo y quando entro en la
casa del dicho lizenciado Escouar y estrupo a la dicha su
sobrina, estaua el sussodicho en la ciudad de Granada
curandose de vna graue enfermedad que tenia;
digan etcetera.

con violencia y fuerza que le hizo la estupró y
conoció carnalmente y hubo su virginidad debajo
de que le había dado la dicha palabra de casamiento;
digan etcétera.
5 Ítem si saben que al tiempo y cuando entró en la
casa del dicho licenciado Escobar y estupró a la dicha su
sobrina, estaba el susodicho en la ciudad de Granada
curándose de una grave enfermedad que tenía;
digan etcétera.

Figura 8. Edición con abreviaturas desarrolladas (izquierda) y con normalización ortográfica (derecha).

Las ventajas de *ODE* en relación con la dimensión filológica no solo se limitan a la visualización de la edición digital, sino que afectan también a la recuperación de información. Cualquier aspecto previamente marcado en lenguaje *XML-TEI* será potencialmente recuperable; compete al investigador decidir en qué particularidades del texto está especialmente interesado y en cuáles no. En *ODE* se han marcado, entre otros aspectos, las cancelaciones (<del>), las adiciones (<add>), las sustituciones (<subst>), las conjeturas editoriales (<supplied>), las intervenciones en estilo directo (<quote>) o las expresiones en otras lenguas (<foreign>). En consecuencia, todas estas cuestiones son fácilmente recuperables mediante una única consulta (Figuras 9 y 10):

Scriptum digital
Vol. 9 (2020), pp. 167-189


Figura 9. Muestra de sustituciones (izquierda) y de expresiones latinas (derecha) en *ODE*.

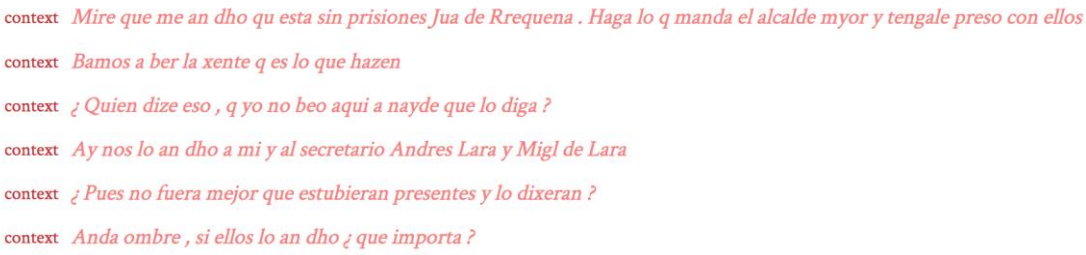


Figura 10. Muestra de expresiones en estilo directo en *ODE*.

### 4.6. Metodología y autogestión

Como se puede inferir de lo expuesto hasta ahora, el proceso de trabajo en *ODE* se basa en una metodología bien definida, que se puede resumir en los siguientes pasos sucesivos: (i) localización de los documentos en archivos históricos; (ii) transcripción del texto y de los metadatos en lenguaje *XML-TEI*; (iii) importación del archivo *XML* y del facsímil a la plataforma *TEITOK*; (iv) tokenización del texto; (v) normalización ortográfica; (vi) etiquetación morfosintáctica y lematización. El resultado final es un producto que combina la edición digital con un corpus anotado.

Conviene destacar que todas las fases de este proceso, desde la localización de las fuentes manuscritas hasta la publicación en línea, son controladas y realizadas por los propios miembros del proyecto *ODE*, esto es, sin necesidad de recurrir a servicios externos. Las tareas de corte más computacional –como son las de procesamiento lingüístico del corpus–, se realizan de forma automática, por lo que el investigador solo debe preocuparse de revisar manualmente el resultado de cada fase para continuar a la siguiente. Para esto último, la plataforma *TEITOK* cuenta con una interfaz amigable que facilita la edición de los datos sin tener que acudir directamente a los datos en lenguaje *XML-TEI*, lo que resultaría engorroso debido a la cantidad de información que va siendo almacenada en cada archivo *XML*. En cualquier caso, esta última opción siempre está disponible.

## 5. CONCLUSIONES

El corpus *ODE* (1492-1900) se está configurando como un corpus histórico regional de pequeño dominio y de carácter especializado, que sirva para complementar a los corpus históricos generales, proporcionando datos de regiones y fenómenos dialectales de baja frecuencia en la documentación literaria e histórica estándar de los corpus generales. Para conseguir este objetivo se han tomado dos decisiones estratégicas: una de carácter filológico y otra de carácter tecnológico.

Desde el punto de vista filológico, *ODE* recoge solo documentación manuscrita de tres tipos textuales (declaraciones de testigos, inventarios de bienes y certificaciones médicas), con objeto de asegurar la comparabilidad de los datos y la proximidad con la lengua oral de las tres regiones representadas: por un lado, el antiguo reino de Granada, núcleo original del corpus (*CORDEREGRA*), y por otro, Andalucía occidental y el centro-norte peninsular, que constituirán dos subcorpus de contraste imprescindibles para obtener información cuantitativa y cualitativamente significativa desde el punto de vista histórico y dialectal.

A nivel tecnológico, *ODE* implementa dos soluciones estrechamente vinculadas: la marcación de datos en *XML-TEI* y el uso de la plataforma *TEITOK*, donde se llevan a cabo la tokenización, la normalización, la lematización y el etiquetado morfosintáctico del corpus. El resultado final es una edición digital de los manuscritos que permite tres visualizaciones de los documentos, en función de los intereses de los usuarios: edición facsímil, paleográfica y normalizada. Al mismo tiempo, se obtiene un corpus anotado que permite cruzar metadatos con datos lingüísticos y recuperar de manera fácil y flexible toda la información textual, partiendo de la forma gráfica original, de la moderna, del lema o de la categoría morfosintáctica.

El modelo de *ODE* es continuador del seguido en el corpus de cartas privadas *P.S. Post Scriptum*, que abrió el camino en el ámbito iberorromance de la llamada «segunda generación» de corpus históricos (López-Couso 2016: 132), basados en el estándar *XML-TEI*, especialmente concebido para la integración con otros corpus similares y la superación de la dispersión metodológica de las etapas iniciales de la lingüística de corpus hispánica.

## REFERENCIAS BIBLIOGRÁFICAS

ARIAS ÁLVAREZ, Beatriz y Juan Antonio HERNÁNDEZ MENDOZA (2013): «Importancia de la incorporación de los parámetros diastráticos y diafásicos en la elaboración del corpus electrónico del español colonial mexicano», *Scriptum Digital*, 2, pp. 5-20. http://www.scriptumdigital.org/documents/01_Arias_y_Hernndez_DEF_Wok.pdf [Consulta: 19/10/2020].

ARRABAL RODRÍGUEZ, Pilar (en prep.): *Variación morfológica y corpus lingüístico: los diminutivos* -ico*,* -ito*,* -illo *en la provincia de Almería (ss. XVIII y XIX)*. Granada: Universidad de Granada. Tesis Doctoral dirigida por Miguel Calderón Campos.

BARON, Alistair, Paul RAYSON y Dawn ARCHER (2009): «Word frequency and key word statistics in historical corpus linguistics», *Anglistik: International Journal of English Studies*, 20, 1, pp. 41-67.

BLANCHE-BENVENISTE, Claire (1994): «The construct of oral and written language: Theoretical issues and educational implications», en Ludo Verhoeven (ed.), *Functional Literacy*. Amsterdam: John Benjamins, pp. 61-74.

BURNARD, Lou (2014): «Introduction», en *What is the Text Encoding Initiative? How to add intelligent markup to digital resources* [en línea]. Marseille: OpenEdition Press. http://books.openedition.org/oep/679 ISBN: 9782821834606. DOI: https://doi.org/10.4000/books.oep.679 [Consulta: 08/10/2020].

CALDERÓN CAMPOS, Miguel (2015): *El español del reino de Granada en sus documentos (1492-1833). Oralidad y escritura*. Berna: Peter Lang (*Fondo Hispánico de Lingüística y Filología*, 22).

CALDERÓN CAMPOS, Miguel (2018): «Las *declaraciones de esencia* del siglo XVIII: un tipo textual para el estudio de la terminología anatómica», *Dynamis*, 38, 2, pp. 427-452.

CALDERÓN CAMPOS, Miguel (2019): «La configuración de la variedad meridional en el reino de Granada», en Eugenio Bustos Gisbert y Juan P. Sánchez Méndez (eds.), *La configuración histórica de las normas del castellano*. Valencia: Tirant Humanidades (*Diachronica Hispanica*), pp. 109-134.

*CED = A Corpus of English Dialogues (1560-1760)*, 2006. Compilado bajo la supervisión de Merja Kytö (Universidad de Uppsala) y Jonathan Culpeper (Universidad de Lancaster). https://cqpweb.lancs.ac.uk/ [Consulta: 19/10/2020].

*CEEC = Corpus of Early English Correspondence (1410-1681)* https://www.helsinki.fi/en/researchgroups/varieng/corpus-of-early-english-correspondence [Consulta: 19/10/2020].

*CHARTA = Red CHARTA*: *Corpus Hispánico y Americano en la Red: Textos Antiguos*. https://www.redcharta.es/ [Consulta: 19/10/2020].

*Cíbola Project* = Proyecto dirigido por Jerry Craddock en que se editan documentos relacionados a la exploración y la conquista del Suroeste hispano de los Estados Unidos en la época colonial (siglos XVI-XVIII). https://escholarship.org/uc/rcrs_ias_ucb_cibola [Consulta: 19/10/2020].

*CODEA = Corpus de Documentos Españoles Anteriores a 1700*. www.textoshispanicos.es [Consulta: 19/10/2020].

*CODEA+ 2015* = GITHE (Grupo de Investigación Textos para la Historia del Español): *Corpus de documentos españoles anteriores a 1800*. http://corpuscodea.es/ [Consulta: 19/10/2020].

*CODEA+ 2020* = GITHE (Grupo de Investigación Textos para la Historia del Español):*Corpus de Documentos Españoles Anteriores a 1900*. http://textoshispanicos.es/index.php?title=Corpus_de_documentos_espa%C3%B1oles_anteriores_a_1900_(CODEA%2B_2020),_FFI1017-82770-P [Consulta: 19/10/2020].

*CORDEREGRA = Corpus diacrónico del español del Reino de Granada (1492-1833)* http://corpora.ugr.es/ode/ [Consulta: 19/10/2020].

*COREECOM = Grupo de Estudio del Español Colonial Mexicano* (*GEECOM*): Banco de datos *Corpus Electrónico del Español Colonial Mexicano*. Beatriz Arias Álvarez (coord.), México: IIFL-UNAM (Instituto de Investigaciones Filológicas, Universidad Nacional Autónoma de México) http://www.iifilologicas.unam.mx/coreecom/ DOI: 10.19130/coreecom.clh.2019 [Consulta: 19/10/2020].

*CorLexIn* = MORALA RODRÍGUEZ, José Ramón (dir.), *Corpus Léxico de Inventarios (CorLexIn)*, http://web.frl.es/CORLEXIN.html [Consulta: 16/09/2020].

CULPEPER, Jonathan y Merja KYTÖ (2010): *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge University Press.

*DEJ* (2016) = REAL ACADEMIA ESPAÑOLA (2016): *Diccionario del español jurídico*. https://dej.rae.es/ [Consulta: 26/3/2020].

DI TULLIO, Ángela y Gabriela RESNIK (2019): «Diario de un soldado: una fuente para la reconstrucción de la oralidad rioplatense del siglo XIX», comunicación presentada en *Documentos y monumentos para la historia de la lengua española. VI Congreso de la Red Internacional CHARTA*, Sevilla, 11/08/2019.

EGIDO, Mª. Cristina (2019): «La variación diatópica: un reto para el estudio de la Morfosintaxis histórica del español», en Viorica Codita y Marcela de la Torre (eds.), *Tendencias y*

*perspectivas en el estudio de la morfosintaxis histórica hispanoamerica*. Madrid/Fráncfort: Iberoamericana/Vervuert (*Lingüística Iberoamericana*, 76), pp. 127-151.

EGIDO, Mª. Cristina (en este volumen): «Variación diatópica en documentos notariales del s. XVII: asturleonés y aragonés», en Miriam Bouzouita y Antoine Primerano (eds.), *Lingüística de corpus e historias de las lenguas iberorrománicas: Nuevas propuestas y últimos desarrollos*, *Scriptum digital*, 9, pp. 15-59.

ENRIQUE-ARIAS, Andrés (coord.) (2009): *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid/Fráncfort: Iberoamericana/Vervuert.

FERNÁNDEZ ALCAIDE, Marta (2009): *Cartas de particulares en Indias del siglo XVI: edición y estudio discursivo*. Madrid/Fráncfort: Iberoamericana/Vervuert.

FONTANELLA DE WEINBERG, Beatriz (1992): «La evolución de los usos americanos de segunda persona singular», *Lingüística*, 4, pp. 7-25.

JANSSEN, Maarten (2012): «NeoTag: A POS Tagger for Grammatical Neologism Detection», en *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012) ELRA*. Estambul, Turquía, mayo de 2012, pp. 2118-2124.

JANSSEN, Maarten (2014): *TEITOK – a Tokenized TEI environment*. http://www.teitok.org/ [Consulta: 27/03/2020].

JANSSEN, Maarten (2016): «TEITOK: Text-Faithful Annotated Corpora», en *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016) ELRA*. Portoroz, Eslovenia, mayo de 2016, pp. 4037-4043.

KABATEK, Johannes (ed.) (2016): *Lingüística de corpus y lingüística histórica iberorrománica*. Berlín: De Gruyter.

LEECH, Geoffrey y Andrew WILSON (1996): *Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Document EAG-TCWG-MAC/R*, marzo de 1996. http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html [Consulta: 26/3/2020].

LÓPEZ-COUSO, María José (2016): «Corpora and online resources in English historical linguistics», en Merja Kytö y Päivi Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics*, Cambridge: Cambridge University Press, pp. 127-145.

MARQUILHAS, Rita (1998): «Mãos inábeis nos arquivos da Inquisicão. Fontes para o estudo fonológico de português do século XVII», en Dieter Kremer (ed.), *Homenaxe a Ramón Lorenzo*, III. Vigo: Galaxia, pp. 761-767.

MORALA, José Ramón (2012): «Léxico e inventarios de bienes en los Siglos de Oro», en Gloria Clavería Nadal, Margarita Freixas, Marta Prat Sabater y Joan Torruella (eds.), *Historia del léxico: perspectivas de investigación*, Madrid/Fráncfort: Iberoamericana/Vervuert, pp. 199-218.

MORALA, José Ramón (2018): «El proyecto *CorLexIn* y la variación diatópica en el léxico del Siglo de Oro», en Dolores Corbella Díaz, Alejandro Fajardo Aguirre y Jutta Langenbacher-Liebgott (eds.), *Historia del léxico español y Humanidades digitales*, Berlín: Peter Lang, pp. 397-417.

MORALA, José Ramón y Hermógenes PERDIGUERO (2019): «La isoglosa de la aspiración de /f/ en el siglo XVII», en Mónica Castillo Lluch y Elena Díez del Corral Areta (eds.), *Reescribiendo la historia de la lengua española a partir de la edición de documentos*, Berna: Peter Lang, pp. 175-199.

OCTAVIO DE TOLEDO Y HUERTA, Álvaro y Lola PONS RODRÍGUEZ (2017): *Textos para la historia del español: queja política y escritura epistolar durante la Guerra de la Independencia: documentación de la Junta Suprema Central en el AHN. Selección, edición y estudio lingüístico*. Alcalá de Henares: Universidad de Alcalá de Henares.

*ODE* = CALDERÓN CAMPOS, Miguel y María Teresa GARCÍA-GODOY (2010-2019): *Oralia Diacrónica del Español (ODE)*. http://corpora.ugr.es/ode [Consulta: 19/10/2020].

*Old Bailey Corpus* = MAGNUS HUBER, Magnus Nissel y Karin PUGA (2016). *Old Bailey Corpus 2.0.* hdl:11858/00-246C-0000-0023-8CFB-2 [Consulta: 19/10/2020].

PÉREZ TORAL, Marta (2017): «¿Escribo como hablo? Variaciones gráficas en el vocalismo tónico en documentos del XVII», *Revista Española de Lingüística*, 47, 2, pp. 49-69.

*P.S. Post Scriptum* = *CLUL* (ed.). 2014. *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. http://ps.clul.ul.pt [Consulta: 19/10/2020].

RAUMOLIN-BRUNBERG, Helena y Terttu NEVALAINEN (2007): «Historical Sociolinguistics: The *Corpus of Early English Correspondence*», en Joan C. Beal, Karen P. Corrigan y Hermann C. Moisl (eds.), *Creating and Digitizing Language Corpora: Diachronic Databases*, vol. 2, Houndmills: Palgrave, pp. 148-171.

RODRÍGUEZ PUENTE, Paula (2018): «En busca de lo hablado en lo escrito en los corpus diacrónicos del español: una comparativa con los corpus anglosajones», *E-Scripta Romanica*, 5, pp. 89-127.

SÁNCHEZ-MARCO, Cristina, Josep Maria FONTANA y Judith DOMINGO (2012): «Anotación automática de textos diacrónicos del español», en Emilio Montero Cartelle y Carmen Manzano Rovira (coords.), *Actas del VIII Congreso Internacional de Historia de la Lengua Española. Vol. 2*, Santiago de Compostela: Asociación de la Historia de la Lengua Española, pp. 1709-1720.

SÁNCHEZ-PRIETO BORJA, Pedro, y Delfina VÁZQUEZ BALONGA (2017): «Hacia un corpus de beneficencia en Madrid (siglos XVI-XIX)», *Scriptum Digital*, 6, pp. 83-103. http://www.scriptumdigital.org/documents/06_SD06_03_SanchezPrieto_VazquezBalonga.pdf [Consulta: 19/10/2020].

SÁNCHEZ-PRIETO BORJA, Pedro y Delfina VÁZQUEZ BALONGA (2019): *La beneficencia madrileña. Lengua y discurso en los documentos de los siglos XVI al XIX*. Madrid: Ediciones Complutense.

RIVADENEIRA, Marcela (en prensa), «Tratamientos nominales en la *Relación autobiográfica de Úrsula Suárez* (1666-1749)», *Rilce: Revista de Filología Hispánica*.

VAAMONDE, Gael (2015): «*P. S. Post Scriptum*. Dos corpus diacrónicos de escritura cotidiana», *Procesamiento del Lenguaje Natural*, 55, pp. 57-64.

VAAMONDE, Gael (2018a): «Escritura epistolar, edición digital y anotación de corpus», *Cuadernos del Instituto Historia de la Lengua*, 11, pp. 139-164.

VAAMONDE, Gael (2018b): «La multidisciplinariedad en la creación de corpus históricos. El caso de *Post Scriptum*», *Artnodes*, 22, pp. 118-127.