Revista de Humanidades Digitales



# Diseño y explotación de un corpus histórico de textos oralizantes para el estudio del español clásico y moderno[1]

*Design and Exploitation of a Historical Corpus of Speech-Related Texts for the Study of Classical and Modern Spanish*

Gael VAAMONDE
gaelvaamonde@ugr.es
Universidad de Granada
https://orcid.org/0000-0001-8360-2805

**RESUMEN**

En este artículo presentamos Oralia Diacrónica del Español (ODE), un corpus histórico de carácter especializado diseñado para investigar el léxico de la vida cotidiana y reconstruir la oralidad y la variación dialectal del español peninsular desde el siglo XVI hasta finales del siglo XIX. El corpus está compuesto por documentación inédita relativa a tres tipos textuales de inmediatez comunicativa: inventarios de bienes, declaraciones de testigos en juicios criminales y certificaciones periciales de cirujanos sobre personas heridas o fallecidas. En este trabajo detallamos el proceso de diseño y la metodología empleada para la creación del corpus: desde la edición digital de los textos en XML-TEI hasta la aplicación de herramientas de procesamiento de lenguaje natural en la plataforma TEITOK. También explicamos, mediante la aportación de varios ejemplos, las posibilidades de explotación de este recurso digital a través de su interfaz de consulta.

**PALABRAS CLAVE**

Lingüística de corpus, edición digital, TEITOK, historia del español, Edad Moderna.

**ABSTRACT**

This article presents Oralia Diacrónica del Español (ODE), a specialized historical corpus that has been designed to study the lexicon of everyday life and to reconstruct the orality and dialectology of Peninsular Spanish from the 16th century to the end of the 19th century. This dataset is made up of unpublished documents from three text types characterized by their communicative immediacy: inventories of goods, witnesses' testimonies in criminal trials, and surgeons' reports on the state of an injured or dead person. In this work, we detail the design process and methodology employed in creating this corpus: from digital edition in TEI -XML to the application of natural language processing tools on the TEITOK platform. We also explain, providing several examples, the possibilities for exploiting this digital resource via its query interface.

**KEYWORDS**

Corpus linguistics, Digital edition, TEITOK, History of Spanish, Early modern period.

[1] Esta publicación es parte del Proyecto PID2022-136256NB-I00 financiado por MICIU/ AEI /10.13039/501100011033 y por FEDER, UE. También es parte del proyecto de I+D+i C-HUM-038-UGR23, cofinanciado/a por la Consejería de Universidad, Investigación e Innovación y por la Unión Europea con cargo al Programa FEDER Andalucía 2021-2027. Agradezco los atinados comentarios de dos evaluadores anónimos a una versión previa del artículo. Cualquier error que persista en el texto es de mi entera responsabilidad.



Revista de Humanidades Digitales http://revistas.uned.es/index.php/RHD/index
Recibido 07/02/2024 – Aceptado 18/03/2024
Editado por Gimena del Rio Riande

## 1. INTRODUCCIÓN

En un trabajo de Merja Kytö sobre el estado actual de los corpus en el ámbito de la lingüística histórica, aclara esta autora que la expresión lingüística histórica de corpus constituye en realidad una tautología, dado que "apart from re-construction, all historical linguistics is in a wide sense corpus-based" (Kytö, 2011, p. 418). Si a esto añadimos las indiscutibles ventajas que trajo consigo el uso de los ordenadores para las aproximaciones empíricas al estudio del lenguaje, entenderemos la relevancia que tienen los corpus informatizados para la investigación de la historia de la lengua. En pocas palabras, el análisis lingüístico en diacronía no se concibe actualmente sin recurrir a los corpus en línea.

Un corpus puede definirse como una colección de textos orales o escritos producidos en condiciones naturales, representativos de una lengua o variedad de lengua, almacenados en soporte informático y destinados al análisis lingüístico (McEnery et al. 2006, p. 5; Gries, 2009, p. 7; Rojo, 2021, p. 1; García-Miguel, 2022, p. 11). En el ámbito hispánico y diacrónico contamos con varios corpus de acceso libre en la red, como son la versión histórica del Corpus del Español (CdE-hist), el Corpus Diacrónico del Español (CORDE) o el Corpus del Diccionario Histórico de la Lengua Española (CDH). Junto a estos corpus de referencia, diseñados para proporcionar una amplia variedad de textos del español desde sus orígenes hasta la época actual, en los últimos años se están construyendo numerosos corpus históricos de carácter más específico y tamaño más reducido que los anteriores. Estos corpus históricos suelen estar concebidos para analizar un área geográfica concreta (por ejemplo, CODEMA. Corpus diacrónico de documentación malagueña), para investigar materiales de un tipo determinado (por ejemplo, CORLEXIN: Corpus Léxico de Inventarios) o para profundizar en una época histórica delimitada (por ejemplo, CODEA+ 2015. Corpus de Documentos Españoles Anteriores a 1800)[2]. Frente a los corpus de propósito general o corpus de referencia, consideramos que este tipo de corpus, conformados por textos históricos que obedecen a algún denominador común, sea del tipo que sea, pueden ser catalogados como corpus (históricos) especializados[3].

En este artículo presentamos el corpus Oralia diacrónica del español (en adelante, ODE), un corpus histórico especializado de reciente creación, de acceso libre y constituido por documentación manuscrita del español clásico y moderno (Calderón Campos, 2019a; Calderón Campos y

[2] El Portal CORHIBER (Torruella y Kabatek, 2024) recopila una lista bastante extensa y actualizada de corpus históricos en lenguas iberorrománicas: http://www.corhiber.org/.

[3] La denominación *corpus especializado* puede resultar confusa, como nos ha hecho ver uno de los evaluadores anónimos de este artículo, ya que se ha venido utilizando en sentido restringido para referirse a corpus que centran su interés en los llamados lenguajes de especialidad, esto es, en textos producidos entre especialistas de una determinada materia o ámbito de conocimiento. En este trabajo lo usamos en sentido amplio para referirnos simplemente a "corpora which are designed with particular research projects in mind" (Kennedy, 1998, p. 20), en contraposición a los corpus generales o de referencia. Con este sentido, entendemos que un corpus puede ser más o menos especializado, dependiendo del propósito para el que fue diseñado (Hunston, 2002, p. 14; Koester, 2021, p. 50).

Vaamonde, 2020)[4]. En las páginas que siguen daremos respuesta a tres preguntas en torno a este corpus histórico: (i) cuál es su finalidad esencial, esto es, qué características presentan los textos que lo componen y cuál es el denominador común que lo convierten en un corpus histórico especializado, (ii) cómo se ha diseñado, esto es, qué metodología se ha seguido en su construcción y qué ventajas se derivan de esta metodología y (iii) cómo se puede explotar eficazmente, esto es, qué tipo de búsquedas admite y cómo se pueden implementar desde la interfaz de consulta. Anticipamos por el momento, y a modo de introducción, las dos características principales que singularizan a ODE frente a los corpus históricos generales. Por un lado, su concepción tiene como finalidad servir a un doble propósito: la reconstrucción diacrónica de la oralidad y el estudio de la variación histórico-dialectal. Las particularidades de los tres tipos textuales que conforman este corpus –declaraciones de testigos, inventarios de bienes y certificaciones médicas– se ajustan claramente a este doble objetivo, como veremos. Por otro lado, la metodología implementada está pensada para ofrecer un corpus anotado y, simultáneamente, una edición académica digital de los manuscritos que lo conforman, aunando para este doble propósito métodos desarrollados tanto por la Lingüística de Corpus como por las Humanidades Digitales.

El presente artículo está estructurado del modo siguiente: en el segundo apartado exponemos las características de los tres tipos textuales que componen el corpus y apuntamos información básica sobre su composición. El tercer apartado está dedicado a la transcripción de los manuscritos en lenguaje XML-TEI y al producto resultante de esta tarea, esto es, la edición académica digital. En el apartado 4 explicamos el procesamiento lingüístico del corpus, desde la tokenización de los textos hasta su lematización. El apartado 5 se centra en la interfaz de búsqueda y la recuperación de información, que ilustraremos con varios ejemplos. Cerramos el artículo ofreciendo algunas conclusiones en el apartado 6.

## 2. COMPOSICIÓN DEL CORPUS: TIPOLOGÍA TEXTUAL Y EJE ESPACIO-TEMPORAL

Los corpus generales, también conocidos como corpus de referencia, están diseñados para ser representativos de variedades diferentes de una misma lengua, por lo que deben incluir textos de muy diversa naturaleza, que idealmente constituyen subcorpus seleccionables por el usuario al momento de realizar una consulta: textos líricos, épicos, dramáticos, periodísticos, religiosos, históricos, etcétera. Sin embargo, los corpus generales de tipo histórico suelen presentar una limitación importante: la escasa representación de la lengua hablada (Rodríguez Puente, 2018, p. 90). Para hacer frente a esta carencia, en los últimos años se han ido elaborando corpus históricos formados exclusivamente por documentación escrita de impronta oral, esto es, por tipos textuales que se aproximan, en mayor o menor medida, al polo de lo que se ha dado en llamar inmediatez comuni-

[4] Todos los datos aportados en este trabajo han sido extraídos de la página electrónica de ODE durante el mes enero de 2024: http://corpora.ugr.es/ode/.

cativa (Koch y Oesterreicher, 1990 [2007]). Nos referimos a textos oralizantes como las cartas privadas, los diarios de contenido personal, los textos dialogados o las autobiografías, por citar algunos ejemplos paradigmáticos.

En el ámbito anglófono, cabe citar al menos tres corpus que han sido compilados con esta finalidad: el Corpus of Early English Correspondence (Raumolin-Brunberg y Nevalainen, 2007), una familia de corpus constituidos por cartas privadas escritas entre 1410 y 1800 y con un tamaño total que supera los cinco millones de palabras; el Corpus of English Dialogues 1560-1760 (Kytö y Walker, 2006), formado por 1.200.000 palabras extraídas de diferentes tipos de discurso dialogado; y el Old Bailey Corpus (Huber, 2007), un corpus que asciende a 24 millones de palabras y recoge declaraciones de testigos del período comprendido entre 1720 y 1913.

En el ámbito hispánico, por su parte, cabe mencionar el corpus P. S. Post Scriptum (Vaamonde, 2018), formado por cartas privadas escritas en Portugal y España entre 1500 y 1830 y con un tamaño que ronda los dos millones de palabras –un millón por cada lengua: portugués y español–, o el Corpus Léxico de Inventarios (Morala, 2014), constituido por relaciones de bienes de la época áurea y con un tamaño actual de unas 1.800.000 palabras. El corpus ODE contribuye a engrosar esta exigua lista de corpus históricos del español orientados a la observación de la oralidad en la escritura, y lo hace integrando en un único recurso tres tipos textuales que resultan adecuados para estudiar, desde diferentes perspectivas, el plano de lo oral y la esfera de lo cotidiano en las sociedades del pasado: las declaraciones de testigos, los inventarios de bienes y las certificaciones médicas. Presentamos a continuación las aportaciones fundamentales de estas fuentes documentales para la historia del español.

Las declaraciones de testigos son el resultado de los interrogatorios realizados en el transcurso de un proceso judicial. Testigos y acusados son llamados a declarar y el escribano deja constancia por escrito de los testimonios que oye. El interés de las declaraciones "procede del hecho de que se trata de textos en los que no interviene ningún tipo de planificación, así como en la procedencia generalmente humilde de sus protagonistas" (Blas Arroyo, 2012, p. 1743), características que aproximan estos textos a la dimensión de lo oral. Además, merecen especial atención las expresiones en estilo directo que aparecen insertadas ocasionalmente en las declaraciones; estas últimas conforman, sin duda, el subcorpus de ODE más próximo al discurso hablado. Por ejemplo:

(1) […] Y, enfadado de ello Mayorgas, siempre que salia dicha Ana Josefa Matheos la perseguia tocando un pito, diciendole varias injurias de puta y otras. Y estando un dia en la plaza la azio de la mantilla y a el volver la cara la escupio. Y una noche se entro en la casa de Palomas y la emvistio diciendola: «Mira picara, que te mato»; expresandole esto a la otra criada, que abrio la puerta pensando que era la dicha Ana Josefa, con otras injurias. Y, haviendo contrahido matrimonio, encontro a su marido en la calle llamada Sin Casas dicho don Juan Maria Mayorgas, y le digo: «¿Es cierto que vsted se a casado con Ana Matheos? Pues tientele vsted el vientre, y vera como esta defondocado de parir muchachos» (ODE, GR1784D9058).

Los inventarios de bienes incluyen "cualquier texto hecho con la finalidad de enumerar, de la forma más minuciosa posible, los bienes de una persona o una institución" (Morala, 2012, p. 200). Se trata de documentación notarial caracterizada por registrar listados de objetos. Nos referimos, entre otros, a inventarios *post mortem*, tasaciones, repartos de herencias, subastas públicas, testamentos, donaciones, cartas de dote o embargos judiciales.

Desde el punto de vista lingüístico, los inventarios presentan una estructura repetitiva y de escaso interés discursivo, siendo de poca utilidad para la observación de aspectos morfosintácticos –a diferencia de lo que ocurre con las declaraciones–; sin embargo, son un tipo textual de primer nivel para el estudio histórico del léxico, así como de ciertos aspectos de la morfología nominal. Su valor, en este sentido, es doble: por un lado, constituyen una ventana a la realidad cotidiana del pasado y permiten documentar vocabulario que no siempre aparece en los corpus de referencia[5], por otro lado, se trata de textos ricos en léxico autóctono e inequívocamente localizados en el eje espacio-temporal, proporcionando así una fuente de datos excepcional para realizar investigaciones histórico-dialectales. Repárese, en el ejemplo (2), en la voz *rasera* (espumadera), que solo se documenta en el oriente andaluz (Nieto Jiménez y Alvar Ezquerra, 2007), o en el empleo del diminutivo *-ico*, característico también de Andalucía oriental (véase Figura 14), como en este caso:

(2) […] paso a las casas que quedan por el fallezimiento de Jun Gonzales Grano de oro, vezino que fue de esta uilla, para efecto de hazer el ynventario de sus vienes; y por dicho alguazil se ejecuto en la forma siguiente: […] Un cazo pequeño usado. Una rasera grande a medio seruir. Un asador grande. Una cuchara de hierro a medio seruir. Una espetera con zinco ganchos. Un mortero con su mano. Una pilica, varro de Ubeda. Un salero, varro de Ubeda. Un peso de esparto con valanza de pino. Un cuchillo de la cozina. Unas pintaderas para pan. Dos abujas de hierro para coser plata. Una zenzerrica. Dos cadenas para mulas (ODE, AL1714I0003).

Finalmente, las certificaciones médicas se refieren a las declaraciones realizadas por cirujanos, barberos o sangradores con el objeto de describir minuciosamente las heridas sufridas por las víctimas de una agresión. Este tipo de documentación, de reciente incorporación a ODE, está siendo extraída de pleitos y probanzas criminales del siglo XVIII y resulta especialmente interesante "para estudiar el léxico médico que se empleaba en el día a día de la práctica terapéutica, porque recoge tanto las voces populares como los tecnicismos médicos y anatómicos de la época" (Calderón Campos, 2018, p. 427). Por ejemplo:

(3) […] Y, preguntado sobre dicha cabeza de prozeso y herida a don Ambrosio Fernandez, maestro de zirujano y vezino de este lugar, dijo que a bisto y curado en la carzel de este lugar a Luis Berdejo, vezino de el, de una herida que tiene en la caueza sobre el güeso pariettal derecho, al parezer dada con ynstrumto contundente de palo o piedra v otro semejante, ronpiendo cuero, cordura y menbrana carnosa, con vna equimoses sobre toda la zircunferencia de la herida. Y, segun la principalidad del sitio que ocupa y aszidentes que le pueden sobrebenir, tiene peligro de muerte, y esto que a dicho es la berdad so cargo del juramento que lleua fecho. Y lo firmo, y dicho señor alcalde y maestro, doi fee (ODE, GR1713C4061).

[5] En el subcorpus de inventarios de ODE aparecen términos como *crehuela, agriaz, acanalador, chaconada, parella, espiocha, jarapa, zafero* o *jamuga*, inexistentes o infrarrepresentados en los corpus históricos generales.

Al momento de redactar estas líneas, ODE asciende a algo más de un millón de palabras, cuya distribución por tipo textual es la que se recoge en la Tabla 1. La menor representación de certificaciones médicas –en número de palabras– viene dada por su reciente incorporación a ODE, así como por la mayor brevedad de estos textos en comparación con las declaraciones o los inventarios[6]. Por lo que respecta al eje temporal (Tabla 2), ODE abarca un período de cuatro siglos, desde principios del siglo XVI hasta finales del siglo XIX, aunque con un claro predominio del siglo XVIII. En el futuro, se prevé incrementar el número de palabras de los siglos menos representados y, en especial, del español decimonónico.

| **Tipo textual** | **N° de palabras** | **%** |
|---|---:|---:|
| inventarios de bienes | 860.090 | 71.74 |
| declaraciones de testigos | 250.258 | 20.87 |
| certificaciones médicas | 72.790 | 6.07 |
| otros | 15.783 | 1.32 |
| TOTAL | 1.198.921 | 100.00 |

Tabla 1. Composición de ODE por tipo textual. Fuente: elaboración propia.

| **Siglo** | **N° de palabras** | **%** |
|---|---:|---:|
| XVI | 241.303 | 20.13 |
| XVII | 270.069 | 22.53 |
| XVIII | 576.984 | 48.12 |
| XIX | 110.565 | 9.22 |
| TOTAL | 1.198.921 | 100.00 |

Tabla 2. Composición de ODE por siglo. Fuente: elaboración propia.

| **Área** | **N° de palabras** | **%** |
|---|---:|---:|
| Andalucía | 776.718 | 64.78 |
| Extremadura[7] | 203.401 | 16.97 |
| Madrid | 93.655 | 7.81 |
| Castilla y León | 90.991 | 7.59 |
| otros | 34.156 | 2.85 |
| TOTAL | 1.198.921 | 100.00 |

Tabla 3. Composición de ODE por área. Fuente: elaboración propia.

[6] Actualmente, se han incorporado 745 inventarios, 116 declaraciones y 111 certificaciones médicas, lo que arroja una media de palabras por tipo documental de 1.150, 2.160 y 650, respectivamente.
[7] Para más información sobre el corpus extremeño de ODE, véase González Sopeña (2022).

| Área | N° de palabras | % |
|---|---|---|
| Andalucía oriental | 468.182 | 60.28 |
| Andalucía occidental | 204.682 | 26.35 |
| Andalucía central | 103.854 | 13.37 |
| TOTAL | 776.718 | 100.00 |

Tabla 4. Composición de ODE por área (Andalucía). Fuente: elaboración propia.

En cuanto al ámbito geográfico (Tablas 3 y 4), ODE es continuación de un proyecto anterior, el Corpus diacrónico del español del reino de Granada (CORDEREGRA), lo que explica el predominio del área andaluza y, en especial, de las actuales provincias de Granada, Málaga y Almería. No obstante, se están incorporando paulatinamente datos del centro y norte peninsular, así como del resto del territorio andaluz, que funcionan como subcorpus de control con respecto a los datos más meridionales a la vez que permiten extender la investigación a diferentes puntos de la geografía española.

## 3. TRANSCRIPCIÓN EN XML-TEI: LA EDICIÓN ACADÉMICA DIGITAL

Construir un corpus implica, como es obvio, almacenar en formato digital los textos que se hayan seleccionado previamente. Si el corpus está basado en documentación manuscrita, lo que suele ser habitual en el caso de los corpus históricos, esta fase del proceso coloca al compilador ante una disyuntiva importante, debido a las implicaciones que se derivan de la decisión que adopte. Una opción es aprovechar las ediciones modernas ya existentes de los correspondientes textos manuscritos y digitalizarlas o, si están accesibles en formato electrónico, importarlas directamente al corpus; podemos referirnos a esta opción como *philological outsourcing* (Dollinger, 2004, p. 6). La otra opción es partir de las fuentes manuscritas para crear ediciones propias que sean fieles al original y asumir, por tanto, el trabajo filológico que conlleva esta tarea; podemos referirnos a esta opción como *philological computing* (Meurman-Solin, 2001, p. 18). La primera alternativa permite ahorrar tiempo y esfuerzo y es adoptada normalmente por los corpus generales debido a su gran tamaño, aunque suele implicar la introducción de discrepancias textuales o la pérdida de grafías originales, entre otros problemas conocidos que limitan su explotación lingüística (Claridge, 2008, pp. 250-251; Honkapohja et al., 2009, pp. 456-460; Marttila, 2014, pp. 132-133). La segunda permite ofrecer ediciones más cuidadas y garantiza una composición homogénea del corpus, pero solo resulta factible en corpus de tamaño reducido debido al coste, en tiempo y esfuerzo, que supone su puesta en práctica. En ODE apostamos claramente por esta segunda línea de trabajo[8].

[8] Esta metodología no es, desde luego, nueva en el ámbito hispánico. Ya a comienzos de los años setenta, la creación del *Dictionary of Old Spanish Language* en el seno del Hispanic Seminary of Medieval Studies exigió la encomiable preparación de transcripciones paleográficas electrónicas de textos antiguos, precisamente para evitar posibles perjuicios editoriales.

En consonancia con las prácticas actuales en el campo de las humanidades digitales, la transcripción de documentos se está realizando con el lenguaje de marcado XML (eXtensive Mark-up Language) y aplicando los estándares de codificación propuestos por el consorcio internacional de la Text Encoding Initiative (TEI) para la representación de textos en formato electrónico. Publicadas sus guías directrices inicialmente en el año 1987, la codificación TEI representa en la actualidad un estándar plenamente consolidado en la comunidad científica de las Humanidades Digitales. El modelo XML-TEI no solo posibilita la recuperación de información a partir de datos estructurados, sino que garantiza su preservación, su reutilización o su posible integración en repositorios digitales, entre otras ventajas conocidas (Burnard, 2014; Fradejas Rueda, 2009-2010, p. 226-227; Allés-Torrent, 2015, p. 19).

En ODE, cada documento seleccionado constituye un archivo XML y la información almacenada en cada archivo está dividida en dos bloques, siguiendo el estándar TEI: una cabecera (<teiHeader>), que incluye los metadatos del documento, y el texto propiamente dicho (<text>), que incluye una transcripción conservadora del contenido del manuscrito. Los metadatos principales recogidos en ODE son el título (<title>), la ubicación del manuscrito (<msidentifier>), el tipo textual (<catRef/>), el lugar y la fecha de creación (<setting>). La estructura TEI que se ha adoptado para marcar esta información se ilustra en la Figura 1. La estructura aquí recogida está simplificada por razones de claridad y corresponde a la declaración de testigos que se ha mostrado previamente en el ejemplo (1). A continuación, explicamos brevemente cada uno de estos segmentos de metadatos.

```
<teiHeader>
  <fileDesc>
     <titleStmt>
        <title>Probanza. Manuel Espadas contra Juan María Mayorgas, vecinos ambos de Loja,
        sobre palabras denigrativas contra la hija de Manuel Espada</title>
[...]
     </titleStmt>
[...]
     <sourceDesc>
        <msDesc>
           <msIdentifier>
              <country>España</country>
              <settlement>Granada</settlement>
              <institution>Archivo de la Real Chancillería de Granada</institution>
              <repository>Probanzas</repository>
              <idno>ARCHGR 10723/10</idno>
           </msIdentifier>
[...]
     </sourceDesc>
  </fileDesc>
[...]
  <profileDesc>
     <textClass>
        <catRef target="dec"/>
     </textClass>
        <settingDesc>
            <setting>
                <name type="place" geo="37.1664839 -4.1496374">España, Granada, Loja</name>
                <date when="1784" when-custom="XVIII">1784</date>
            </setting>
        </settingDesc>
[...]
</teiHeader>
```

Figura 1. Ejemplo simplificado de una cabecera de ODE en XML-TEI. Fuente: elaboración propia.

Con respecto al título (<title>), este se utiliza en ODE a modo de breve descripción del documento. Se trata de un campo de contenido libre y redactado por el propio responsable de la transcripción del manuscrito, por lo que la casuística es muy variada, como se refleja en los ejemplos recogidos en el ejemplo (4). No está pensado, por tanto, para realizar búsquedas sistemáticas, aunque sí puede ser de utilidad para recuperar, por ejemplo, subtipos documentales dentro de la categoría de inventarios (piénsese en palabras como *dote* o *testamento*), como en el siguiente caso:

(4) a. <title>Inventario de los bienes de doña Luisa Francisca de Olerson para su venta a José del Villar</title> (ODE, CA1708I7068).
b. <title>Carta de pago de dote de Juan Serrano para María Núñez</title> (ODE, BA1703I7139).
c. <title>Testamento de María Delgado, vecina de Malpartida, jurisdicción de Cáceres</title> (ODE, CC1615I2566)
d. <title>Declaración de varios testigos sobre las cuchilladas que le dieron a Alonso Gallego, vecino del lugar de Dos Hermanas</title> (ODE, SE1537D2778)
e. <title>Aprecio de los bienes de don Bartolomé Muñoz</title> (ODE, CA1738I7077)
f. <title>Certificado del cirujano Salvador de Olivares sobre las heridas de Juan José de Almodóvar, vecino de Atarfe</title> (ODE, GR1737C9031)

Con respecto a la ubicación del manuscrito (<msIdentifier>), esta información se organiza en ODE en cinco ítems, que siguen una jerarquía descendente, comenzando desde lo más general hasta lo más específico: país (<country>), ciudad (<settlement), nombre del archivo (<institution>), fondo documental (<repository>) y signatura (<idno>). El nombre del archivo resulta de especial interés, por ser el nivel intermedio, y por esta razón constituye un campo de búsqueda propio desde la interfaz de consulta (*Archivo* en la Figura 11).

Al momento de redactar estas líneas, se han visitado casi una veintena de archivos históricos, entre los que cabe destacar las Reales Chancillerías de Valladolid y Granada. Desde finales del siglo XV, como es sabido, el reino de Castilla dividió judicialmente su territorio en dos partes, utilizando el río Tajo como línea natural de separación. Valladolid se convirtió en la sede de la Chancillería del norte, mientras que Granada albergó la del sur. En consecuencia, los documentos producidos ambas Chancillerías son una valiosa fuente para contrastar variedades dialectales en diferentes enclaves del norte y sur peninsulares. Además, también se ha seleccionado documentación en un nutrido número de archivos históricos provinciales de la zona meridional (Loja, Lorca, Almería, Badajoz, Cáceres, Cádiz, Huelva, Jaén, Málaga, Sevilla), a los que cabe sumar, en áreas más norteñas, el Archivo Histórico de Protocolos de Madrid o Archivo Histórico Provincial de Burgos.

Por lo que respecta al tipo textual, en ODE se ha establecido una taxonomía sencilla formada por cuatro categorías (Figura 2) y susceptible de ampliación en el futuro. La categoría correspondiente a cada documento se marca mediante el elemento <catRef/> y, concretamente, a través del atributo @target, que adopta uno de los cuatro valores previamente definidos en la taxonomía: *inv* (inventarios de bienes), *cer* (certificados médicos), *dec* (declaraciones de testigos) y *oth* (otros). Este último valor se reserva actualmente para un reducido número de documentos de

especial interés en ODE, pero que no encajan en el resto de categorías (por ejemplo, cartas particulares).

```
<taxonomy>
    <category id="inv">
        <catDesc>inventarios de bienes</catDesc>
    </category>
    <category id="cer">
        <catDesc>certificados médicos</catDesc>
    </category>
    <category id="dec">
        <catDesc>declaraciones de testigos</catDesc>
    </category>
    <category id="oth">
        <catDesc>otros</catDesc>
    </category>
</taxonomy>
```

Figura 2. Taxonomía de ODE en XML-TEI para la clasificación del tipo textual. Elaboración propia.

La información relativa al lugar y la fecha de creación del documento se marca conjuntamente dentro del elemento <setting>, que se organiza en ODE en dos elementos de nivel inferior: <name> y <date>. El contenido textual del primero captura la información geográfica en una estructura jerárquica del tipo *país, provincia, ciudad*, esto es, con cada nivel de información separado por comas (por ejemplo, *España, Granada, Loja*). Esta informacion es utilizada posteriormente para obtener las coordenadas geográficas correspondientes a cada localización, que son almacenadas en un atributo @geo dentro del elemento <name> (Figura 1). La inclusión de las coordenadas dentro del archivo XML se realiza automáticamente a partir del contenido textual de <name> y mediante un *script* en lenguaje Perl. Este *script* añade el atributo @geo, y su correspondiente valor, a partir del motor de búsqueda *Nominatim*, desarrollado y mantenido por la comunidad *OpenStreetMap* (OSM). El pseudocódigo del *script* se puede resumir en los pasos siguientes (*5*):

(*5*) - Descargar librerías necesarias para manejar archivos XML y realizar solicitudes web.
- Abrir y leer el contenido del archivo XML.
- Comprobar si el XML ya tiene información de geolocalización (o sea, comprobar si ya existe una ruta "//setting/name[@geo]").
- Si ya tiene información de geolocalización, mostrarla y terminar el programa.
- Extraer el contenido de la ruta "//setting/name" y almacenar el nombre del lugar en una variable.
- Construir una URL utilizando el nombre del lugar como parámetro de búsqueda.
- Utilizar el motor de búsqueda *Nominatim* para obtener los datos de geolocalización en formato XML.
- Extraer la información de geolocalización y agregarla a la ruta "//setting/name[@geo]" dentro del XML.
- Guardar los cambios en el XML y terminar el programa.

La ventaja evidente de este proceso de geolocalización es que aumenta las posibilidades de visualización de los resultados, ya que se pueden proyectar en un mapa las consultas realizadas sobre el corpus, como mostraremos en el apartado 5. Por lo que se refiere a la fecha, el elemento <date> contiene en ODE dos atributos: @when y @when-custom. El primero permite marcar la fecha concreta en que se generó el manuscrito, mientras que el segundo sirve para dejar constancia del siglo correspondiente. El contenido de ambos constituye también sendos campos de búsqueda en la interfaz de consulta (*Año* y *Siglo* en la Figura 11, respectivamente), lo que que facilita la recuperación de información dentro de rangos temporales de diferente extensión, que son definidos en función de los intereses del usuario.

Por lo que atañe a la marcación del texto (<text>), en ODE se parte de una edición paleográfica del manuscrito, esto es, una transcripción fiel a la ortografía original y en la que se respetan las particularidades sobre disposición, edición y apariencia del texto, como son adiciones (<add>), cancelaciones (<del>), abreviaturas (<abbr>), conjeturas (<supplied>) o lagunas textuales (<gap/>). La fidelidad a la ortografía original resulta crucial para realizar investigaciones sobre fonética histórica, mientras que la marcación de aspectos peritextuales como los mencionados anteriormente confiere rigor filológico al corpus, aumenta las opciones de búsqueda y evita errores de interpretación en el análisis lingüístico (Meurman-Solin y Tyrkkö, 2013).

La Tabla 5 muestra los elementos principales utilizados en ODE para la transcripción del contenido textual de los manuscritos. A continuación de esa tabla, y a modo de ejemplo, recogemos en la Figura 3 la transcripción en TEI-XML correspondiente al fragmento del ejemplo (1):

| Elemento | Descripción | Elemento | Descripción |
|---|---|---|---|
| <pb/> | inicio de folio | <supplied> | texto conjeturado |
| <lb/> | inicio de línea | <gap/> | texto omitido |
| <p> | párrafo | <hi> | texto resaltado |
| <add> | adición textual | <surplus> | texto redundante |
| <del> | cancelación | <sic> | texto incorrecto |
| <foreign> | texto en otra lengua | <quote> | estilo directo |
| <handShift/> | cambio de mano | <q> | cita textual |
| <abbr> | abreviatura | <ex> | desarrollo de abreviatura |

Tabla 5. Elementos TEI usados en ODE para la transcripción de manuscritos. Fuente: elaboración propia.

```
<text>
    <body>
[...]
Y, enfadado de ello Mayorgas, siempre <pb n="3r" facs="IMG_4986.JPG"/>
<lb n="1"/> que salia dicha Ana Josefa Matheos la perseguia tocando
<lb n="2"/> un pito, diciendole varias injurias de puta y otras. Y
estan<lb n="3"/>do un dia en la plaza la azio de la mantilla y a el
<lb n="4"/> volver la cara la escupio. Y una noche se entro en la
<lb n="5"/> casa de Palomas y la emvistio diciendola: <quote>Mira picara,
<lb n="6"/> que te mato</quote>; expresandole esto a la otra criada, que
<lb n="7"/> abrio la puerta pensando que era la dicha Ana Josefa,
<lb n="8"/> con otras injurias. Y, haviendo contrahido
matri<lb n="9"/>monio, encontró a su marido en la calle llamada
<lb n="10"/> Sin Casas dicho dn Juan Maria Mayorgas, y le digo: <quote>¿Es
<lb n="11"/> cierto que V se a casado con Ana Matheos? Pues
tien<lb n="12"/>tele V el vientre, y vera como esta defondocado de
<lb n="13"/> parir muchachos</quote>.
[...]
    </body>
</text>
```

Figura 3. Ejemplo simplificado de transcripción en XML-TEI. Fuente: elaboración propia.

El fragmento de la Figura 3 permite ilustrar algunas de las etiquetas TEI-XML más utilizadas en ODE, como son los inicios de folio (<pb/>), los inicios de línea (<lb/>), las abreviaturas (<abbr>) o, en el caso de las declaraciones de testigos, los segmentos en estilo directo (<quote>). Nótese que el elemento <pb/> incluye un atributo @facs que sirve para asociar el folio con la imagen correspondiente. De hecho, ODE incluye para su consulta y descarga las imágenes facsimilares de todos los documentos transcritos, de forma que siempre es posible cotejar con el manuscrito original cualquier aspecto relativo a la edición digital. Nótese, igualmente, que cuando un inicio de línea con el elemento <lb/> interrumpe una palabra, como en el caso de *estan<lb n="3"/>do*, se omite en la transcripción cualquier espacio antes y después del salto de línea. Esta convención facilita la posterior *tokenización* automática del texto, esto es, su división en palabras ortográficas, ya que el espacio se convierte en un indicador distintivo entre ellas.

Somos conscientes de que la tarea de transcripción utilizando elementos TEI-XML puede ser un proceso lento y costoso debido a la atención minuciosa que requiere para marcar correctamente cada elemento del texto. Sin embargo, los beneficios que ofrece en términos de visualización y recuperación de información son significativos. Piénsese, por ejemplo, en los segmentos en estilo directo, que son de especial interés en ODE. Una vez marcados, pueden ser resaltados en la edición digital con una simple orden en lenguaje CSS (*Cascading Style Sheet*) —y lo mismo con las abreviaturas—, de forma que son fácilmente identificables por el usuario (Figura 4). Y, como veremos en el apartado 5, la información textual que contienen dichos segmentos puede ser filtrada y recuperada desde la interfaz de consulta, convirtiéndose así en un subcorpus de gran valor lingüístico.

que salia dicha Ana Josefa Matheos la perseguia tocando un pito, diciendole varias injurias de puta y otras. Y estando un dia en la plaza la azio de la mantilla y a el volver la cara la escupio. Y una noche se entro en la casa de Palomas y la emvistio diciendola: *Mira picara, que te mato*; expresandole esto a la otra criada, que abrio la puerta pensando que era la dicha Ana Josefa, con otras injurias. Y, haviendo contrahido matrimonio, encontró a su marido en la calle llamada Sin Casas dicho don Juan Maria Mayorgas, y le digo: *¿Es cierto que Vsted se a casado con Ana Matheos? Pues tientele Vsted el vientre, y vera como esta defondocado de*

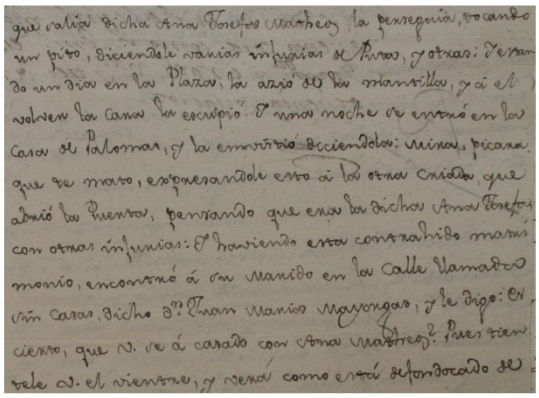

Figura 4. Ejemplo de visualización en paralelo: transcripción y facsímil. Fuente: elaboración propia.

## 4. PROCESAMIENTO LINGÜÍSTICO: EL CORPUS ANOTADO

La distinción entre un corpus y un archivo de textos es, en teoría, fácil de establecer. Mientras que "a corpus designed for linguistic analysis is normally a systematic, planned and structured compilation of text, an archive is a text repository, often huge and opportunistically collected, and normally not structured" (Kennedy, 1998, p. 4). De acuerdo con esta distinción y aplicándola a ODE, se podría decir que el resultado del proceso anterior ya constituye de hecho un corpus, puesto que las ediciones digitales elaboradas en el seno de este proyecto responden a un interés lingüístico, resultan de un trabajo de selección y presentan una marcación sistemática y estructurada.

No obstante, hay al menos dos aspectos que permiten aumentar las posibilidades de explotación de un corpus: la anotación lingüística y la interfaz de consulta. Es obvio que un corpus lematizado y etiquetado morfosintácticamente podrá ser interrogado sobre más —y más abstractas— cuestiones que un corpus al que no se le haya añadido información gramatical alguna, lo que convierte la anotación del corpus en un factor determinante de su explotación posterior y, en definitiva, de su utilidad como herramienta de investigación (Rojo, 2021, p. 113). Además, resulta pertinente contar con un sistema de consulta que permita recuperar y presentar la información de forma eficaz, para rentabilizar así todo el trabajo previo de marcación textual y de anotación lingüística. En el caso de ODE, esta lista de necesidades es cubierta por una única herramienta llamada TEITOK.

TEITOK (Janssen, 2014; Janssen y Vaamonde, 2020) es una plataforma web diseñada para crear y gestionar corpus basados en lenguaje XML-TEI. Su funcionalidad es doble: por un lado, permite al usuario navegar a través de una edición digital y, al mismo tiempo, consultar un corpus anotado; por otro lado, permite al compilador editar los datos en XML-TEI y aplicar herramientas de procesamiento automático de textos. En palabras de Maarten Janssen, 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).

Originalmente, esta plataforma web se creó para responder a las necesidades de visualización y gestión que demandaba el corpus P. S. Post Scriptum (Vaamonde 2015; 2018), también de carácter histórico, pero actualmente da soporte a corpus de diferente naturaleza y tamaño. En ODE, se ha adoptado la misma metodología que se siguió en P. S. Post Scriptum —basada en el modelo XML-TEI para la transcripción de manuscritos y en el modelo TEITOK para la creación y edición del corpus anotado— y se han incorporado además algunas de las nuevas funcionalidades que ha ido desarrollando esta herramienta desde su lanzamiento en 2014.

Al hablar de *modelo TEITOK* nos referimos, en particular, a la anotación lingüística del corpus a través de esta plataforma, un proceso que en ODE comprende al menos cuatro fases, descritas brevemente a continuación: (i) delimitación del texto en unidades mínimas o *tokens*, (ii) normalización ortográfica, (iii) etiquetado morfosintáctico y (iv) lematización. El resultado de este proceso, que se realiza de forma secuencial, se codifica en lenguaje XML y se guarda en los propios archivos que contienen la edición digital, como veremos a continuación.

### 4.1. *Tokenización*

La *tokenización* se refiere al proceso de segmentar un texto para identificar y delimitar las unidades del corpus que van a ser anotadas —los *tokens*—, habitualmente palabras, números y signos de puntuación (Gries y Berez, 2017, p. 383). En TEITOK, esta tarea se ejecuta de manera automática y consiste en la adición de un elemento <tok>, acompañado de un identificador único, para marcar cada uno de los tokens que contiene el texto. Así, una expresión como *una sauana de estopa y canamo viexa;*, perteneciente al documento con referencia GR1723I9022 en ODE, resulta en la marcación que se recoge en la Figura 5, una vez transcrita y tokenizada:

```
1 <tok id="w-61">una</tok>
2 <tok id="w-62">sauana</tok>
3 <tok id="w-63">de</tok>
4 <tok id="w-64">estopa</tok>
5 <tok id="w-65">y</tok>
6 <tok id="w-66">canamo</tok>
7 <tok id="w-67">viexa</tok>
8 <tok id="w-68">;</tok>
```

Figura 5. Resultado de la tokenización en TEITOK. Fuente: elaboración propia.

Cierto es que el propio estándar TEI ya contempla un elemento específico para marcar palabras (<w>). En este punto, sin embargo, TEITOK se distancia del estándar TEI y opta por un elemento propio (<tok>), debido a que los signos de puntuación también se consideran tokens, pero no palabras. Así, la transcripción original se marca siempre como contenido del elemento <tok> y las sucesivas campañas de anotación que se van aplicando sobre el corpus se van marcando mediante diferentes atributos dentro de este mismo elemento, como veremos. De esta forma, el elemento <tok> facilita las tareas de procesamiento del lenguaje natural que se contemplan en TEITOK (Janssen, 2016, p. 4038).

### 4.2. Normalización ortográfica

Por normalización nos referimos al proceso de asignar a cada palabra transcrita su correspondencia ortográfica según el estándar actual. La normalización ortográfica presenta al menos tres ventajas importantes para un corpus histórico. En términos de visualización, permite acompañar la transcripción del manuscrito original de una versión normalizada, facilitando así la lectura del texto a quienes no estén familiarizados con el español no contemporáneo (Figuras 6 y 7).

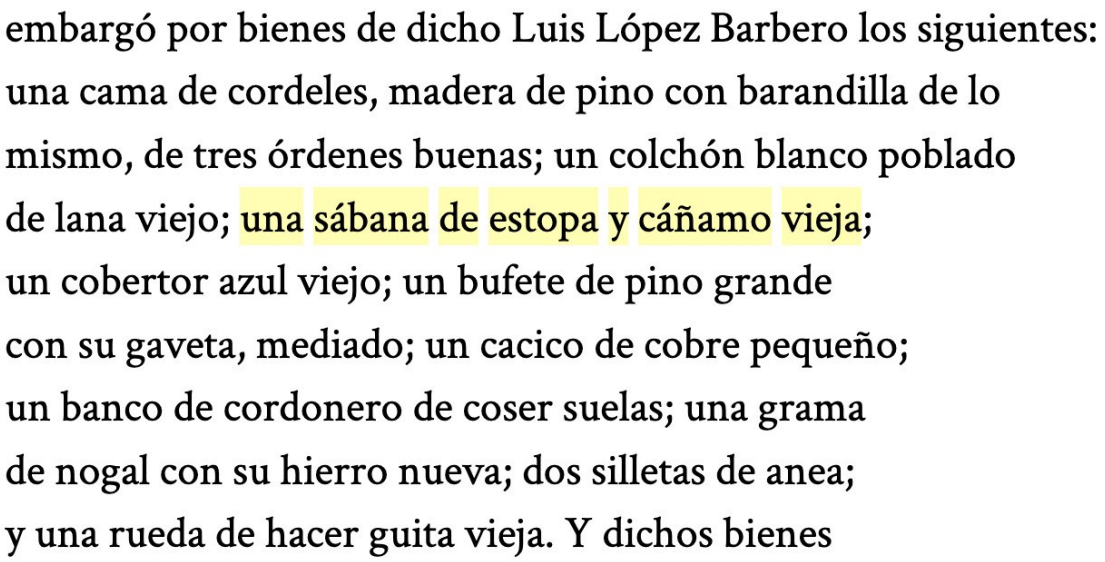
embargó por bienes de dicho Luis López Barbero los siguientes:
una cama de cordeles, madera de pino con barandilla de lo
mismo, de tres órdenes buenas; un colchón blanco poblado
de lana viejo; una sábana de estopa y cáñamo vieja;
un cobertor azul viejo; un bufete de pino grande
con su gaveta, mediado; un cacico de cobre pequeño;
un banco de cordonero de coser suelas; una grama
de nogal con su hierro nueva; dos silletas de anea;
y una rueda de hacer guita vieja. Y dichos bienes

Figura 6. Ejemplo de visualización de la transcripción original. Fuente: ODE.

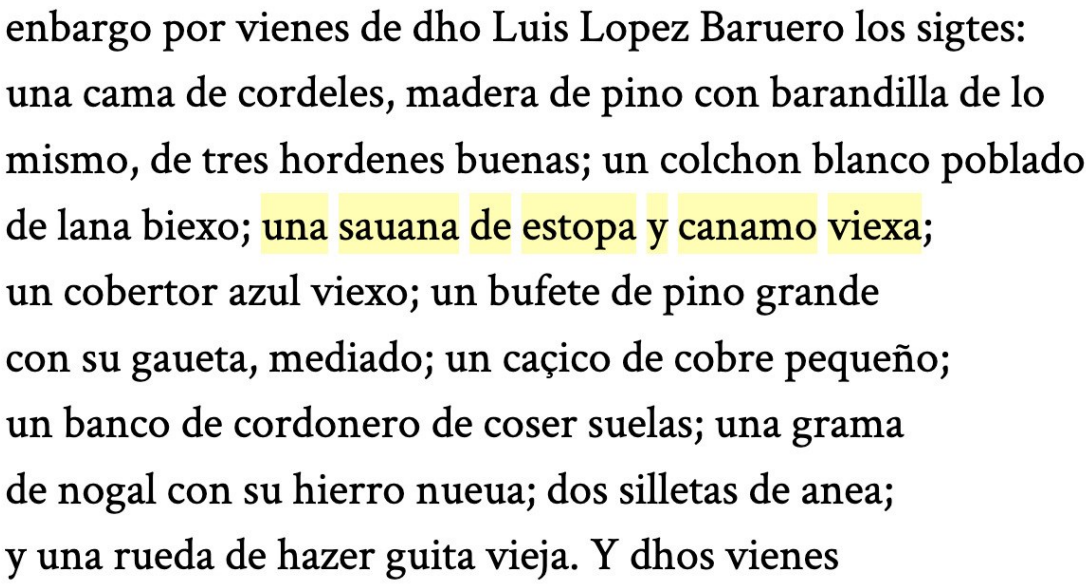
enbargo por vienes de dho Luis Lopez Baruero los sigtes:
una cama de cordeles, madera de pino con barandilla de lo
mismo, de tres hordenes buenas; un colchon blanco poblado
de lana biexo; una sauana de estopa y canamo viexa;
un cobertor azul viexo; un bufete de pino grande
con su gaueta, mediado; un caçico de cobre pequeño;
un banco de cordonero de coser suelas; una grama
de nogal con su hierro nueua; dos silletas de anea;
y una rueda de hazer guita vieja. Y dhos vienes

Figura 7. Ejemplo de visualización de la transcripción normalizada. Fuente: ODE.

En términos de explotación del corpus, permite obtener las variantes ortográficas de una misma palabra, pues todas ellas están vinculadas a una única forma normalizada; ofrece, por tanto, una vía de análisis interesante para realizar estudios sobre ortografía y fonética históricas (ver la Tabla 6 más abajo). En términos de procesamiento, finalmente, simplifica el etiquetador morfosintáctico, que se puede aplicar sobre un conjunto mucho más restringido de posibilidades léxicas –las formas normalizadas– para, de esta forma, reducir los errores de etiquetado automático (Sánchez Marco *et al.*, 2012). En ODE, la normalización se ejecuta automáticamente sobre el resultado de la tokenización y consiste en la adición de un atributo @nform (i.e. *normalized form*) con el valor correspondiente dentro del elemento <tok> (véase, por ejemplo, la forma *sábana* en la Figura 8). Posteriormente, el equipo de lingüistas de ODE edita el resultado, si requiere corrección manual, desde la propia interfaz de TEITOK.

```
<tok id="w-61">una</tok>
<tok id="w-62" nform="sábana">sauana</tok>
<tok id="w-63">de</tok>
<tok id="w-64">estopa</tok>
<tok id="w-65">y</tok>
<tok id="w-66" nform="cáñamo">canamo</tok>
<tok id="w-67" nform="vieja">viexa</tok>
<tok id="w-68">;</tok>
```

Figura 8. Resultado de la normalización en TEITOK. Fuente: elaboración propia.

### 4.3. Etiquetado morfosintáctico

El etiquetado morfosintáctico consiste en la asignación de una etiqueta con información gramatical a cada palabra del corpus. En TEITOK, esta tarea se ejecuta automáticamente sobre el resultado de la normalización y consiste en la adición de un atributo @pos (siglas de *Part Of Speech*) con el valor correspondiente dentro del elemento <tok> (Figura 9). El resultado requiere revisión manual por parte del equipo de lingüistas de ODE.

```
<tok id="w-61" pos="DI0FS0">una</tok>
<tok id="w-62" nform="sábana" pos="NCFS000">sauana</tok>
<tok id="w-63" pos="SPS00">de</tok>
<tok id="w-64" pos="NCFS000">estopa</tok>
<tok id="w-65" pos="CC">y</tok>
<tok id="w-66" nform="cáñamo" pos="NCMS000">canamo</tok>
<tok id="w-67" nform="vieja" pos="AQ0FS0">viexa</tok>
<tok id="w-68" pos="Fx">;</tok>
```

Figura 9. Resultado del etiquetado morfosintáctico en TEITOK. Fuente: elaboración propia.

Actualmente, TEITOK utiliza el etiquetador morfosintáctico NeoTag (Janssen, 2012) para llevar a cabo esta tarea. El sistema de etiquetas utilizado por NeoTag para la anotación del corpus ODE está basado en la propuesta del grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas (Leech y Wilson, 1996). Este sistema se rige por una lógica de posiciones: cada etiqueta consta de una secuencia de letras y números, que representan rasgos morfosintácticos determinados dependiendo de su posición dentro de la secuencia completa. Por ejemplo, a la forma *sábana* se le ha asignado la etiqueta NCFS000, que indica los rasgos *nombre* (N), *común* (C), *femenino* (F), *singular* (S); el cero (0) representa información gramatical no aplicable para esa categoría y posición en la lengua que es objeto de análisis. El conjunto de etiquetas morfosintácticas que se ha utilizado para la anotación lingüística de ODE se puede consultar en http://corpora.ugr.es/ode/index.php?action=tagse.

### 4.4. Lematización

La última fase del procesamiento lingüístico consiste en la asignación del lema correspondiente a cada token, esto es, la forma que representa al conjunto de variantes morfológicas de una palabra y encabeza la entrada de un diccionario. En TEITOK, esta información se codifica dentro del atributo @lemma y, al igual que las dos tareas precedentes, se realiza automáticamente y el resultado se revisa de forma manual (Figura 10):

```
<tok id="w-61" pos="DI0FS0" lemma="un">una</tok>
<tok id="w-62" nform="sábana" pos="NCFS000" lemma="sábana">sauana</tok>
<tok id="w-63" pos="SPS00" lemma="de">de</tok>
<tok id="w-64" pos="NCFS000" lemma="estopa">estopa</tok>
<tok id="w-65" pos="CC" lemma="y">y</tok>
<tok id="w-66" nform="cáñamo" pos="NCMS000" lemma="cáñamo">canamo</tok>
<tok id="w-67" nform="vieja" pos="AQ0FS0" lemma="viejo">viexa</tok>
<tok id="w-68" pos="Fx" lemma=";">;</tok>
```

Figura 10. Resultado de la lematización TEITOK. Fuente: elaboración propia.

El resultado de este proceso, que se realiza de forma secuencial, se codifica en lenguaje XML y se almacena dentro de la propia edición digital de los textos, para ser después recuperado mediante consultas desde la interfaz de búsqueda. En esencia, por tanto, ODE está constituido por un número creciente de archivos XML que contienen (i) metadatos descriptivos (Figura 1), (ii) transcripciones paleográficas basadas en el estándar TEI (Figura 3) y (iii) anotación lingüística de cada token codificada de acuerdo con el procedimiento descrito a lo largo de este apartado, esto es, tokenización (Figura 5), normalización (Figura 8), etiquetado morfosintáctico (Figura 9) y lematización (Figura 10). Al momento de redactar estas líneas, el corpus que conforma ODE se nutre del contenido filológico-lingüístico almacenado en 999 archivos XML.

La metodología explicada en los párrafos precedentes demanda, en último término, contar con una interfaz de consulta que permita extraer toda esta información de forma accesible y eficaz, y que permita igualmente presentarla y analizarla según los formatos habituales empleados en lingüística de corpus: concordancias, listas de palabras, frecuencias de uso, etcétera. En respuesta a esta necesidad, el sistema TEITOK convierte automáticamente el contenido de los archivos XML en un corpus basado en el formato CWB (*Corpus WorkBench*) (Christ *et al.*, 1999). La ventaja primordial de este formato estriba en la posibilidad de realizar búsquedas mediante sintaxis CQL (*Corpus Query Language*), un lenguaje de consulta diseñado para recuperar eficazmente información de un corpus lingüístico, incluyendo patrones gramaticales complejos mediante el uso de expresiones regulares[9].

Dedicaremos el resto del presente trabajo a presentar algunos ejemplos de búsquedas para ilustrar las posibilidades de explotación del corpus ODE, acompañados de la sintaxis CQL correspondiente siempre que sea posible. Conviene señalar, no obstante, que ODE incorpora en su interfaz de búsqueda un constructor automático de consultas en sintaxis CQL, pensado para los usuarios que no estén familiarizados con este tipo de lenguaje.

[9] En la bibliografía especializada, este lenguaje también aparece referido en ocasiones con las siglas CQP (*Corpus Query Processor*) (Evert y Hardie, 2011, p. 8). Mantenemos aquí las siglas CQL por ser las más extendidas y las que se utilizan en TEITOK y, por extensión, en ODE.

## 5. LA EXPLOTACIÓN DEL CORPUS

Ya hemos comentado que la información relativa a cada documento se distribuye en dos bloques principales: la cabecera (<teiHeader>), donde se incluyen los metadatos que describen el documento, y el contenido textual (<text>), basado en una transcripción paleográfica del manuscrito y sobre la que se van añadiendo sucesivas campañas de edición que enriquecen lingüísticamente el corpus. Paralelamente, el corpus ODE puede ser interrogado desde al menos dos perspectivas: la búsqueda de documentos y la búsqueda de patrones textuales. La interfaz de búsqueda de ODE, una vez desplegado el constructor de consultas, refleja esta doble posibilidad (Figura 11).

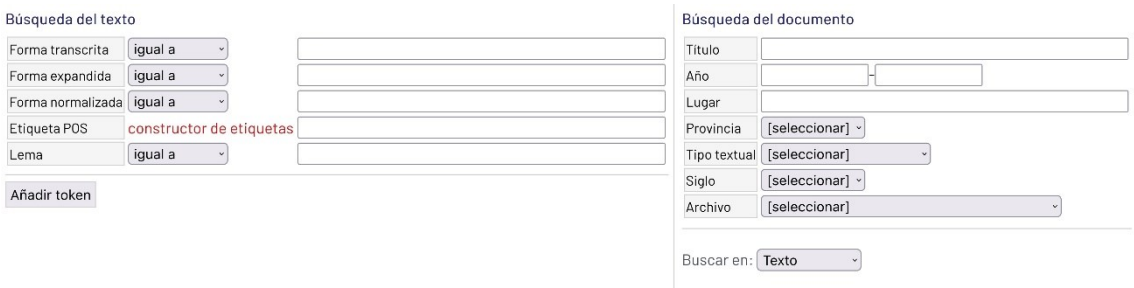


Figura 11. Interfaz de búsqueda. Fuente: ODE.

La primera opción permite obtener la lista de documentos que cumplen las características metatextuales seleccionadas al momento de realizar la búsqueda. Se trata de una opción especialmente adecuada para historiadores y filólogos que deseen obtener documentación histórica manuscrita de una determinada zona, época o tipología. Por ejemplo, podemos recuperar la lista de inventarios de bienes producidos durante el siglo XVIII en la provincia de Huelva (Figura 12). Nótese que el valor de la columna *ID* en este tipo de resultados es en realidad un enlace que reencamina al usuario a la edición digital del manuscrito correspondiente.

| ID | Título | Año |
|---|---|---|
| HU1701I0316.xml | Carta de dote. Bienes que Tomás Martín recibe de su esposa Ana de la Cruz | 1701 |
| HU1701I0317.xml | Testamento del Licenciado Fernando Díaz Arroyo | 1701 |
| HU1701I0318.xml | Bienes que Juan Suarez recibe como dote por su matrimonio con María Francisca | 1701 |
| HU1702I0315.xml | Bienes que Diego Díaz recibe como dote | 1702 |
| HU1704I0319.xml | Recibo de dote. Bienes que Joaquín Alonso recibe como dote por su matrimonio con Catalina de Rojas | 1704 |
| HU1704I0320.xml | Recibo de dote. Bienes que Francisco de Paula recibe como dote de María de Santiago | 1704 |
| HU1704I0326.xml | Testamento de Polonia de Sanjuán | 1704 |
| HU1709I0324.xml | Testamento de Bernarda Antonia | 1709 |
| HU1709I0325.xml | Testamento de María de la Concepción | 1709 |

Figura 12. Resultado de consulta sobre metadatos (lugar, tipo textual y siglo). Fuente: ODE.
CQL: <text> [] :: match.text_province = "Huelva" & match.text_cat = "inv" & match.text_century = "XVIII".

La segunda opción, orientada a la investigación lingüística, permite obtener listas de ejemplos que cumplen determinados criterios, ya sean relativos a una palabra o a una construcción gramatical

más compleja. Podemos recuperar, por ejemplo, las ocurrencias de la voz *sencillo* documentadas en ODE (Figura 13). Al tratarse de una consulta sobre un único token, basta con incluir la propia palabra en el cajón de búsqueda y el sistema, por defecto, devuelve todas las ocurrencias de esta palabra. Salvo que se indique lo contrario, las búsquedas se aplican sobre la versión normalizada del corpus, lo que significa que, para el caso que nos ocupa, se obtendrán todas las formas que hayan sido normalizadas a *sencillo*, incluyendo las que ya han sido transcritas así originalmente.

En ODE, la tabla de resultados que se obtiene con este tipo de consultas textuales consta siempre de cuatro columnas, que de izquierda a derecha ofrecen la información siguiente: (i) un enlace al documento completo del que se ha extraído el ejemplo, (ii) el ejemplo en formato de concordancia o *KWIC* (*Key Word In Context*), esto es, con la(s) palabra(s) buscada(s) en posición central y acompañada(s) a izquierda y derecha de su contexto inmediato, cuya extensión es personalizable, (iii) el año en que fue escrito el manuscrito y (iv) el lugar donde fue escrito. Además, las concordancias pueden mostrarse, bien en la versión con grafía original, o bien de acuerdo con la versión normalizada del corpus, también a elección del usuario.

| | | | |
|---|---|---|---|
| contexto | rs. \| Otro de tafetan **çençillo** en treinta y tres rs | 1703 | Granada |
| contexto | de la China forrado \| en **çençillo** , en dosçientos rs. | 1703 | Granada |
| contexto | , una colcha de tafetan **çençillo** , biexa, colorada \| y | 1663 | Cáceres |
| contexto | piezas de tafetan \| doble y **cencillo** , y pañuelos menssionados \| en | 1705 | Granada |
| contexto | , otro pedaço de tafetan **sençillo** pardo; vn pedaço de | 1661 | Badajoz |
| contexto | : otras ocho de \| tafetan **sencillo** y quatro de gasa encarnada | 1711 | Madrid |
| contexto | . \| Yd otro grande mas **sencillo** y usado en 60. | 1855 | Madrid |
| contexto | berdoso \| Quatro piezas de bonbasi **sensillo** \| Una pieza de fustan \| Una | 1645 | Badajoz |
| contexto | Veynte baras de tafetan encarnado **sensillo** \| Çiento y çinquenta baras de | 1645 | Badajoz |
| contexto | . \| Un delanttar de tafettan **sensillo** . \| Un jubon de felpa | 1700 | Málaga |

Figura 13. Concordancias de la forma *sencillo* (visualización con grafía original). Fuente: ODE.
*CQL: [nform = "secillo"].*

También podemos obtener resultados sobre expresiones más complejas como, por ejemplo, el sintagma *arroba de NC*, donde *NC* equivale a cualquier nombre común (Figura 14). En este caso, es necesario especificar los criterios de cada uno de los tokens que conforman la construcción pretendida: el lema *arroba* seguido de la forma normalizada *de* seguida de un nombre común, esto es, de una palabra cuya etiqueta morfosintáctica comience por las siglas *NC*.

| | | | | | |
|---|---|---|---|---|---|
| contexto | tres reales. \| Una media | **arrobas de hierro** | , diez reales. \| Tres | 1748 | Jaén |
| contexto | de a ocho. \| Diez | **arrobas de higos** | al precio corriente de | 1715 | Málaga |
| contexto | peso de garabatos con media | **arrobas de hierro** | en \| pesas. \| Una alcuza | 1754 | Jaén |
| contexto | sesenta reales. \| Ídem Treinta | **arrobas de indias** | a doce reales. \| Ídem | 1845 | Madrid |
| contexto | en dos reales. \| Media | **arroba de lana** | en quince reales. \| Una | 1774 | Almería |
| contexto | , \| por considerarle más de | **arrobas de lana** | \| a cada cabecera. \| Otra | 1713 | Jaén |
| contexto | treinta reales. \| De seis | **arrobas de lana** | , a veinte y cinco | 1748 | Jaén |
| contexto | cáñamo en 120. \| Siete | **arrobas de lana** | para henchimiento en d280. | 1795 | Jaén |
| contexto | cinco reales. \| Las quince | **arrobas de maíz** | tardío y las quince arrobas | 1715 | Málaga |
| contexto | reales. \| Doscientas y sesenta | **arrobas de mosto** | , a dos reales y | 1748 | Jaén |

Figura 14. Concordancias de la expresión *arroba de NC* (visualización con grafía normalizada). Fuente: ODE.
*CQL: [lemma = "arroba"] [nform="de"] [pos="NC.+"].*

Finalmente, y aunque no se recoge en las Figuras 13 y 14, cabe destacar que ODE facilita recuentos sobre la consulta que se ha ejecutado, como son el número de resultados obtenidos –65 ocurrencias de *sencillo* y 247 ocurrencias de *arroba de NC*– o su frecuencia normalizada expresada en casos por millón de palabras –54.22 y 206.02, respectivamente–.

El corpus permite, además, agrupar cualquier resultado según la frecuencia de aparición de uno o más criterios de búsqueda. Así, podemos buscar una palabra y posteriormente ordenar el resultado en función de su transcripción original. Por ejemplo, las 65 apariciones de la palabra *sencillo* se pueden agrupar en nueve variantes gráficas diferentes, tal como se recoge en la Tabla 6. Como ya se ha señalado, esta posibilidad de consulta es especialmente interesante para abordar estudios sobre fonética histórica: formas como *zenzillo*, *zencillo*, *sensillo* o *zensillo*, reflejan en la escritura lo que debían ser pronunciaciones ceceantes y seseante del escribano –o del declarante– en el plano oral[10].

| **Forma transcrita** | **N°** |
|---|---|
| senzillo | 33 |
| zenzillo | 7 |
| zencillo | 7 |
| sensillo | 7 |
| zensillo | 4 |
| çençillo | 3 |
| sencillo | 2 |
| cencillo | 1 |
| sençillo | 1 |
| TOTAL | 65 |

Tabla 6. Distribución de la voz sencillo según forma transcrita. Fuente: elaboración propia.
CQL: *Matches = [nform="sencillo"]; group Matches match form.*

[10] Este tipo de voces no suelen ser recogidas en las ediciones que utilizan los corpus históricos de referencia, lo que otorga a ODE un valor añadido que debe ser aquí destacado.

Utilizando esta estrategia de agrupación, también podemos obtener automáticamente los lemas que aparecen con mayor frecuencia en la tercera posición de la construcción *arroba de NC*. La lista resultante nos da una idea de los sustantivos que se documentaban habitualmente junto a esta unidad de medida, a la vez que nos acerca léxico representativo de la vida cotidiana del pasado (Tabla 7).

| Lema | N° | Lema | N° | Lema | N° |
|---|---|---|---|---|---|
| aceite | 44 | vaso | 5 | higo | 3 |
| lana | 42 | paja | 5 | tocino | 3 |
| vino | 23 | queso | 4 | barro | 3 |
| lino | 14 | manteca | 4 | pasa | 3 |
| cáñamo | 9 | tela | 4 | almendra | 2 |
| vinagre | 7 | agua | 4 | jabón | 2 |
| hierro | 7 | carbón | 4 | azúcar | 2 |
| cabida | 6 | miel | 3 | papa | 2 |

Tabla 7. Lemas más frecuentes en la construcción *arroba de NC*. Fuente: elaboración propia. CQL: *Matches = [lemma="arroba"] [nform="de"] @[pos="NC.+"]; group Matches target lemma*.

Por defecto, las búsquedas textuales realizadas en ODE se ejecutan sobre el texto completo, esto es, sobre el conjunto de tokens contenido en el elemento <text> dentro de cada archivo XML. Sin embargo, cabe la posibilidad de restringir nuestra búsqueda a segmentos textuales más específicos. Así, podemos examinar únicamente la información incluida en las expresiones de estilo directo que se recogen en las declaraciones de testigos. Estos segmentos textuales, marcados mediante el elemento <quote>, conforman el subcorpus de ODE más próximo al discurso hablado y, por esta razón, resultan de especial interés para el lingüista histórico. En la Figura 15 se recogen algunas concordancias de los lemas *pícaro* y *cornudo*, dos de los adjetivos documentados con mayor frecuencia dentro de las expresiones de estilo directo que integran el corpus.

| | | | | |
|---|---|---|---|---|
| contexto | salgan aquí todos los cabrones | **cornudos** | \| de tu linaje y cuantos | 1626 Almería |
| contexto | de tu linaje y cuantos | **cornudos** | están ahí a demandármelo | 1626 Almería |
| contexto | esta manera, que sois | **cornudos** | , \| y dijo Jerónimo: | 1626 Almería |
| contexto | embistió diciéndola: *Mira* | **pícara** | , \| que te mato; | 1784 Granada |
| contexto | \| dicho Salvador Moreno: | **Pícaro** | desvergonzado, \| ¿cómo te | 1656 Málaga |
| contexto | se \| ha de ir este | **pícaro** | en sus pies, y | 1626 Almería |
| contexto | dicho Juan Rodríguez dijo: | **pícaro** | , ¿tú no eres | 1626 Almería |
| contexto | dijo, *tú eres un* | **pícaro** | infame, y voto a | 1626 Almería |
| contexto | un golpe diciéndole: «» | **pícaro** | , cuando paséis por \| delante | 1627 Almería |
| contexto | : *¿Osa hablar aquel* | **pícaro** | desvergonzado desde allí por estar | 1615 Granada |

Figura 15. Concordancias de los lemas *pícaro* y *cornudo* en fragmentos de estilo directo. Fuente: ODE. *CQL: [lemma = "(pícaro|cornudo)"] within quote*.

La presentación de resultados en forma de concordancias y listas de frecuencia es la que suele utilizarse en interfaces de búsqueda y programas informáticos diseñados para recuperar y analizar datos de corpus. Las concordancias proporcionan una rápida visión general acerca del empleo de una determinada palabra o construcción sintáctica, mientras que las listas de frecuencia ofrecen información útil sobre su distribución en el (sub)corpus seleccionado (Stefanowitsch, 2020, p. 50). Ninguna de estas opciones, sin embargo, es la más adecuada para examinar la variación dialectal de un fenómeno lingüístico.

Lo idóneo para este último propósito es, obviamente, recurrir a la generación de mapas, una posibilidad con la que también cuenta ODE. Como ya hemos mencionado, las coordenadas geográficas correspondientes al lugar de creación de cada manuscrito se almacenan en los metadatos de la edición digital, dentro de un atributo @geo, y esta información se exporta al corpus para ser recuperada a petición del usuario. En otras palabras, puesto que los documentos de ODE están vinculados a un lugar concreto e inequívoco, los resultados de una consulta sobre el texto contenido en estos documentos pueden georreferenciarse y proyectarse en un mapa, facilitando así su interpretación espacial. Aportamos algunos ejemplos que muestran la utilidad de este modo de visualización –y, por extensión, del corpus ODE– para la investigación histórico-dialectal del español[11].

[11] En todos los casos, las cifras que recogen los mapas se refieren a número de ocurrencias, y no a número de documentos.

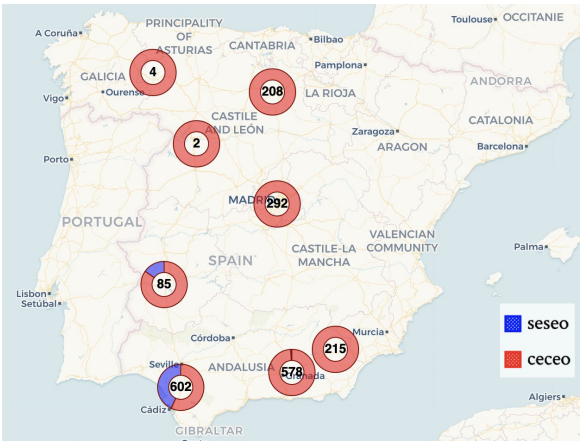


Figura 16. Distribución geográfica de las formas *sinco* y *cinco*. Fuente: ODE.


Figura 17. Distribución geográfica de las formas *lienso* y *lienzo*. Fuente: ODE.

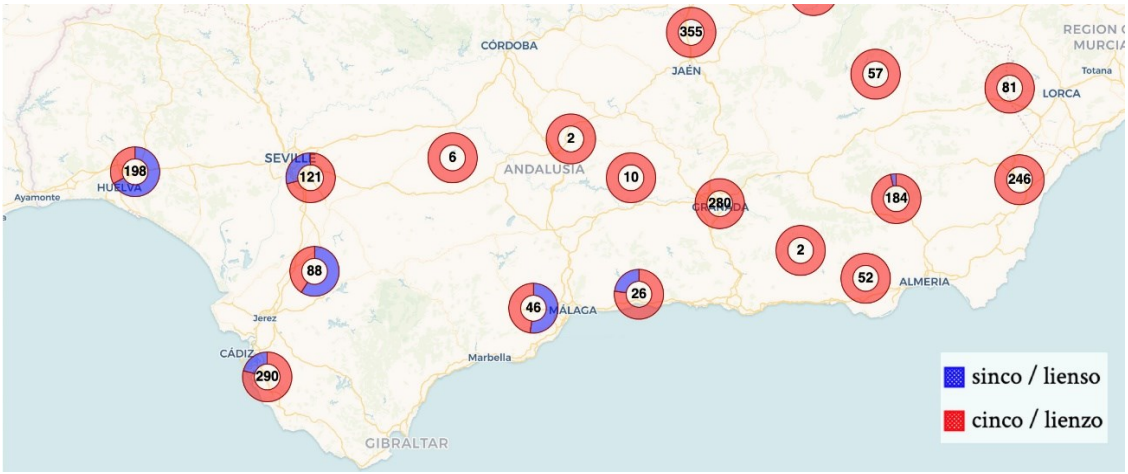


Figura 18. Distribución geográfica de las formas *sinco-lienso frente a cinco-lienzo* (Andalucía). Fuente: ODE.

En el primer ejemplo se comparan las variantes ortográficas *sinco* y *cinco* (Figura 16) y en el segundo ejemplo se hace lo propio con las variantes *lienso* y *lienzo* (Figura 17). Las primeras, claramente indicadoras de seseo, se documentan solo en la parte meridional de la península ibérica, que es la que tradicionalmente se ha asociado a la pronunciación seseante en territorio peninsular (Penny, 2001, p. 118-120; NGLE, 2011, §5.5k); en concreto, estas formas se atestiguan en ODE, principalmente, en áreas de la Andalucía occidental (Figura 18) y en tierras extremeñas meridionales. Las segundas, por el contrario, se registran en textos de todas las áreas geográficas representadas en el corpus[12]. Esta distribución sugiere que las formas *sinco* y *lienso* –y otras similares como *calsones*, *dose* o *asul*, bien documentadas en ODE y que arrojan distribuciones geográficas análogas– no pueden interpretarse como errores ortográficos casuales, sino como auténticas representaciones del habla en la escritura; y sugiere, además, que la procedencia geográfica de los

[12] La forma *cinco*, a su vez, podría estar reflejando casos de ceceo en algunas zonas de Andalucía. A la identificación de estos casos, sin embargo, solo podemos llegar mediante la observación de otras formas inequívocamente ceceantes en el mismo documento.

escribanos que producían estos textos debía ser muy próxima, cuando no idéntica, a la de los lugares donde eran escritos (Calderón, 2019b, p. 119).

Esta hipótesis cobra fuerza con los datos aportados por un segundo ejemplo de carácter dialectal: el empleo de los sufijos diminutivos *-ico* e *-ito* en el sur de la península (Figura 19). En este caso, se observa que el sufijo *-ico* tiene una presencia destacada en la zona de la Andalucía oriental, particularmente en las provincias de Almería y Granada –62 ocurrencias de *-ico* frente a 10 ocurrencias de *-ito*–, pero presenta una baja intensidad en el resto del territorio andaluz –7 ocurrencias de *-ico* frente a 59 de *-ito*–. La distribución resultante es congruente con las tendencias señaladas en la bibliografía especializada (Náñez, 2006) y coincide, de hecho, con recuentos previos sobre el uso de estos diminutivos en épocas pasadas (Calderón, 2019b, p. 122). Cabe destacar, además, que algunas de las terminaciones en *-ico* registradas en ODE se asocian a bases nominales no documentadas en los corpus históricos académicos, como *sabanica*, *caperucica*, *ovillico*, *tonelico* o *piletica*. La riqueza dialectal del corpus se demuestra también en estas construcciones morfológicas inéditas.


Figura 19. Distribución geográfica de los diminutivos *-ito* e *-ico* en Andalucía. Fuente: ODE.

Somos conscientes de que tanto el seseo como el empleo de los diminutivos en la historia del español peninsular arrojan panoramas dialectales históricamente complejos, que no son objeto de estudio en el presente trabajo. Constituyen, no obstante, usos lingüísticos conocidos y documentados que sirven para demostrar la validez de ODE como fuente de datos histórico-dialectales, a tenor de los mapas aquí recogidos.

## 6. CONCLUSIONES

La publicación en línea, hace ya más de dos décadas, de los primeros corpus diacrónicos generales para la lengua española, como son el CORDE o el CdE-hist, ha facilitado el acceso rápido y sencillo a una gran cantidad de ejemplos y ha supuesto un avance indiscutible para la lingüística histórica en el ámbito hispánico. Junto a ellos, en los últimos años se están compilando cada vez

más corpus diacrónicos especializados que, a expensas de reducir el número de textos recopilados, permiten dar un salto cualitativo con respecto a los grandes corpus de referencia.

Adoptando esta última perspectiva, se presenta el corpus ODE: un recurso de acceso libre en la red constituido por documentación manuscrita producida entre 1500 y 1900. Los textos que lo componen obedecen a tres tipos textuales que destacan por su inmediatez comunicativa: los inventarios de bienes, que permiten documentar el léxico de la vida cotidiana utilizado en diferentes regiones peninsulares, las declaraciones de testigos, que ofrecen la oportunidad de extraer expresiones en estilo directo de épocas pasadas, y las certificaciones médicas, que sirven para contrastar las voces populares con los tecnicismos médicos y anatómicos. ODE no solo se nutre de estas fuentes documentales –escasamente representados en corpus históricos–, sino que es resultado de un minucioso trabajo de selección realizado en los fondos documentales de diferentes archivos históricos peninsulares. Se busca, con ello, crear un corpus histórico especializado que sirva como herramienta para reconstruir diacrónicamente la oralidad e investigar la variación dialectal en el español clásico y moderno.

Con respecto a su diseño, ODE aúna métodos propios de las humanidades digitales y de la lingüística de corpus: hace uso del estándar XML-TEI para la edición digital de fuentes primarias y de la plataforma web TEITOK para el procesamiento lingüístico del corpus, así como para la edición y visualización de los resultados. En esencia, ODE está compuesto por un número creciente de archivos XML que contienen, por cada documentado seccionado, una cabecera con metadatos (<teiHeader>) y una transcripción paleográfica del manuscrito (<text>). La primera incluye información detallada sobre el texto y es usada en ODE para marcar aspectos como el título, la ubicación del manuscrito, el tipo textual al que pertenece, la fecha del documento o su lugar de creación, incluyendo las coordenadas geográficas correspondientes. La segunda, de carácter paleográfico en ODE, incluye la marcación de diferentes aspectos estructurales, visuales y editoriales tales como inicios de línea, adiciones, cancelaciones, abreviaturas, conjeturas o segmentos en estilo directo, entre otros. Además, el contenido textual de la edición digital es enriquecido mediante herramientas de procesamiento del lenguaje natural integradas en TEITOK: un tokenizador, un normalizador ortográfico, un etiquetador morfosintáctico basado en el estándar EAGLES y un lematizador.

Todo este proceso de marcación y anotación multiplica, como es lógico, las opciones de búsqueda y recuperación de la información. Para rentabilizar al máximo esta ventaja, ODE hace uso de la sintaxis CQL, un lenguaje utilizado para realizar búsquedas en corpus textuales y que permite al usuario formular consultas complejas de manera precisa y eficientes. Además, CQL también admite operadores lógicos y expresiones regulares, lo que permite realizar consultas aún más sofisticadas y precisas. Finalmente, la presentación de resultados tras realizar una búsqueda admite

en ODE diferentes tipos de visualización, entre las que cabe citar listas de documentos, listas de concordancias, agrupaciones de frecuencias o mapeado de datos.

En definitiva, creemos que tanto la naturaleza de los textos que contiene como la metodología aplicada en su elaboración convierten a ODE en un recurso útil y complementario de los corpus históricos generales en el ámbito hispánico. Confiamos en que los datos que ofrece este corpus permitan hallar algunas respuestas y plantear nuevos interrogantes, y en que su construcción contribuya a seguir avanzando en el estudio de la historia de la lengua española.

## REFERENCIAS BIBLIOGRÁFICAS

Allés-Torrent, S. (2015). Edición digital y algunas tecnologías aliadas. *Ínsula: revista de letras y ciencias humanas*, 822, 18-21.

Blas Arroyo, J. L (2012). Tras las huellas de la variación y el cambio lingüístico a través de textos de inmediatez comunicativa. Fundamentos de un proyecto de sociolingüística histórica. En F. J. de Cos Ruiz y M. Franco Figueroa (Coords.), *Actas del IX Congreso Internacional de Historia de la Lengua Española. Volumen 2* (pp. 1743-1762). Iberoamericana/Vervuert.

Burnard, L. (2014). *What is the Text Encoding Initiative? How to add intelligent markup to digital resources*. OpenEdition Press. https://doi.org/10.4000/books.oep.42.

Calderón Campos, M. (2018). Las declaraciones de esencia del siglo XVIII: un tipo textual para el estudio de la terminología anatómica. *Dynamis*, *38*(2), 427-452. https://dx.doi.org/10.4321/s0211-95362018000200000.

Calderón Campos, M. (2019a). La edición de corpus lingüísticos en la plataforma TEITOK. El caso de Oralia diacrónica del español (ODE). *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, 6, 21-36.

Calderón Campos, M. (2019b). La configuración de la variedad meridional en el reino de Granada. En E. Bustos Gisbert y J. P. Sánchez Méndez (Eds.), *La configuración histórica de las normas del castellano* (pp. 109-134). Tirant Humanidades.

Calderón Campos, M. y Vaamonde, G. (2020). Oralia Diacrónica del Español: un nuevo corpus de la Edad Moderna. *Scriptum Digital*, 9, 167-189. https://doi.org/10.5565/rev/scriptum.10.

Claridge, C. (2008). Historical corpora. In A. Lüdeling y M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 242-259). Walter de Gruyter.

Christ, O., Schulze, B. M., Hofmann, Anja y König, Esther (1999). *The IMS corpus workbench: Corpus query processor (CQP): User's manual*. Technical report, IMS, University of Stuttgart. https://corpora.ficlit.unibo.it/TCORIS/cqpman.pd.

Dollinger, S. (2004). 'Philological computing vs. 'philological outsourcing' and the compilation of historical corpora: a Late Modern English test case. *Vienna English Working Papers (VIEWS)*, *13*(2), 3-23

Evert, S. y Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 Conference*. University of Birmingham. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/paper-153.pd

Fradejas Rueda, J. M. (2009-2010). La codificación XML/TEI de textos medievales. *Memorabilia*, 12, 219-247.

García-Miguel, J. M. (2022). Lingüística de corpus: de los datos textuales a la teoría lingüística. *Estudios de Lingüística del Español*, 45, 11-42. https://raco.cat/index.php/Elies/article/view/40373.

González Sopeña, I. (2022). Documentación notarial extremeña del siglo XVII en Oralia Diacrónica del Español (ODE): el léxico de la vida cotidiana a través de inventarios de bienes pacenses. *Romanica Olomucensia*, 34(1), 13-30. https://doi.org/10.5507/ro.2022.00.

Gries, S. (2009). *Quantitative Corpus Linguistics with R: A practical introduction*. Routledge.

Gries, S. y Berez, A. L. (2017). Linguistic Annotation in/for Corpus Linguistics. En N. Ide y J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (pp. 379-410). Springer. https://doi.org/10.1007/978-94-024-0881-2_1.

Honkapohja, A., Kaislaniemi, S. y Marttila, V. (2009). Digital Editions for Corpus Linguistics: Representing manuscript reality in electronic corpora. En A. H. Jucker, D. Schreier y M. Hundt (Eds.): *Corpora: Pragmatics and Discourse* (pp. 451-475). Rodopi.

Huber, M. (2007). The Old Bailey Proceedings, 1674-1834: evaluating and annotating a corpus of 18th and 19th century spoken English. En A. Meurman-Solin y A. Nurmi (Eds.). *Annotating variation and change*. Helsinki: VARIENG. https://varieng.helsinki.fi/series/volumes/01/huber.

Hunston, S. (2002). *Corpora in Applied Linguistics, Cambridge*. Cambridge University Press.

Janssen, M. (2012). NeoTag: A POS Tagger for Grammatical Neologism Detection. En N. Calzolari *et al.* (Eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation* (pp. 2118-2124). ELRA. http://www.lrec-conf.org/proceedings/lrec2012/pdf/1098_Paper.pd.

Janssen, M. (2014). TEITOK - a Tokenized TEI enviroment. http://www.teitok.org/index.ph.

Janssen, M. (2016). TEITOK: Text-Faithful Annotated Corpora. En N. Calzolari *et al.* (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation* (pp. 4037-4043). ELRA. https://aclanthology.org/L16-1637.pd.

Janssen, M. y Vaamonde, G. (2020). Da edición dixital á análise lingüística. A creación de corpus históricos na plataforma TEITOK. En R. Álvarez y E. González Seoane (Eds.), *Calen barbas, falen cartas. A escrita en galego na Idade Moderna* (pp. 271-292). Consello da Cultura Galega (Ensaio & Investigación). https://consellodacultura.gal/publicacion.php?id=437.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. Longman.

Koch, P. y Oesterreicher, W. (1990[2007]). *Lengua hablada en la Romania: español, francés, italiano*. Gredos. Versión española de Araceli López Serena a patir del original alemán de 1990: *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch*. Tübingen: Niemeyer.

Koester, A. (2022). Building small specialised corpora. En A. O'Keeffe y M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 48-61). Routledge. https://doi.org/10.4324/9780367076399-.

Kytö, M. (2011). Corpora and historical linguistics. *Revista Brasileira de Linguística Aplicada*, *11*(2), 417-457. https://doi.org/10.1590/S1984-63982011000200000.

Kytö, M. y Walker, T. (2006). *Guide to A Corpus of English Dialogues 1560-1760*. Acta Universitatis Upsaliensis.

Leech, G. y Wilson, A. (1996). *EAGLES. Recommendations for the Morphosyntactic Annotation of Corpora*. http://www.ilc.cnr.it/EAGLES96/annotate/annotate.htm.

Marttila, V. (2014). *Creating Digital Editions for Corpus Linguistics. The Case of Potage Dyvers, a Family of six Middle English recipe collections* [Tesis doctoral]. Universidad de Helsinki. http://hdl.handle.net/10138/13558.

McEnery, T, Xiao, R. y Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.

Meurman-Solin, A. (2001). Structured text corpora in the study of language variation and change. *Literary and Linguistic Computing*, *16*(1), 5-27.

Meurman-Solin, A. y Tyrkkö, J. (2013). Introduction. En A. Meurman-Solin y J. Tyrkkö (Eds.), *Principles and Practices for the Digital Editing and Annotation of Diachronic Data*. VARIENG. https://varieng.helsinki.fi/series/volumes/14/introduction.htm.

Morala, J. R. (2012). Léxico e inventarios de bienes en los Siglos de Oro. En G. Clavería Nadal, M. Freixas, M. Prat Sabater y J. Torruella (Eds.), *Historia del léxico: perspectivas de investigación* (pp. 199-218). Iberoamericana/Vervuert.

Morala, J. R. (2014). El CorLexIn, un corpus para el estudio del léxico histórico y dialectal del Siglo de Oro. *Scriptum Digital*, 3, 5-28. https://doi.org/10.5565/rev/scriptum.47.

Náñez Fernández, E. (2006). *El diminutivo. Historia y funciones en el español clásico moderno*. UAM Ediciones.

Nieto Jiménez, L. y Alvar Ezquerra, M. (2007). *Nuevo tesoro lexicográfico del español (s. XVI – 1726)*. Arco Libros.

NGLE = RAE / ASALE (2011): *Nueva gramática de la lengua española. Fonética y fonología*. Espasa.

ODE = Calderón Campos, M. y María Teresa García-Godoy (2010-): *Oralia Diacrónica del Español (ODE)*. [enero de 2024]. http://corpora.ugr.es/od.

Penny, R. (2001). *Variation and change in Spanish*. Cambridge University Press.

Raumolin-Brunberg, H. y Nevalainen, T. (2007). Historical Sociolinguistics: The Corpus of Early English Correspondence. En J. C. Beal, K. P. Corrigan y H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora. Volume 2: Diachronic Databases* (pp. 148-171). Palgrave Macmillan.

Rodríguez Puente, P. (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*, 89-127. https://doi.org/10.18778/2392-0718.05.0.

Rojo, G. (2021). *Introducción a la lingüística de corpus en español*. Routledge. https://doi.org/10.4324/9781003119760.

Sánchez Marco, C., Fontana Méndez, J. M. y Domingo, J. (2012). Anotación automática de textos diacrónicos en español. En E. Montero Cartelle y C. Manzano Rovira (Coords.), *Actas del VIII Congreso Internacional de Historia de la Lengua Española. Vol. 2* (pp. 1709-1720). Meubook.

Sinclair, J. (2004). *Trust the text. Language, corpus and discourse*. Routledge.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press. https://doi.org/10.5281/zenodo.373582.

Torruella, J. y Kabatek, J. (2024). *Portal de Corpus Históricos Iberorrománicos* (versión 6.0). http://corhiber.org.

Vaamonde, G. (2015). P. S. Post Scriptum. Dos corpus diacrónicos de escritura cotidiana. *Procesamiento del Lenguaje Natural*, *55*, 57-64.

Vaamonde, G. (2018). La multidisciplinariedad en la creación de corpus históricos: El caso de Post Scriptum. *Artnodes*, 22, 118-127. https://doi.org/10.7238/a.v0i22.323.