ANOTACIÓN Y ANÁLISIS PARA EL ESTUDIO DIACRÓNICO DE ARABISMOS EN ORALIA DIACRÓNICA DEL ESPAÑOL (ODE)*

Inmaculada González Sopeña
Gael Vaamonde
Universidad de Granada

Resumen: El estudio de léxico de origen árabe en lengua española ha suscitado el interés de numerosos investigadores en el campo filológico debido al contacto temporal acaecido entre el árabe hispánico y los romances peninsulares durante casi ocho siglos. Los trabajos publicados sobre esta cuestión han priorizado sobre todo el periodo medieval (Neuvonen 1941), mientras que los datos posteriores al siglo xvi no han recibido tanta atención. Para el español clásico y moderno, los estudios han versado, en especial, sobre el análisis de casos específicos (Calderón Campos 2010, Morala Rodríguez 2012a) y sobre la sustitución, pérdida u obsolescencia de arabismos en periodos muy concretos (Giménez-Eguíbar 2015). Sin embargo, faltan trabajos que, con una visión de conjunto que abarque cuatro siglos de historia, ofrezcan datos cuantitativos acerca de la mayor o menor presencia de arabismos en nuestra lengua.

Es evidente, en este sentido, la poca preocupación que ha existido por preparar materiales orientados al estudio de arabismos y de su evolución a lo largo tiempo. Así, los corpus históricos de referencia (CORDE, CDH), empleados normalmente para apoyar los mencionados estudios, no incluyen ningún tipo de anotación a nivel léxico que facilite la recuperación automática de información etimológica, lo que dificulta su tratamiento estadístico. Por todo ello, este trabajo tiene un doble objetivo. En primer lugar, se explica la metodología adoptada para recuperar automáticamente todos los arabismos contenidos en ODE (Oralia Diacrónica del

* Este trabajo ha sido realizado en el marco del Proyecto PID2022-136256NB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE. Asimismo, 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.

https://doi.org/10.31819/9783968699554_005

Español), un corpus histórico especializado e integrado por textos de inmediatez comunicativa producidos entre el siglo XVI y el XIX. En segundo lugar, se analizan las frecuencias de uso de arabismos obtenidas de este corpus, plasmadas en un dendrograma que refleja un punto de inflexión claro en el uso de este tipo de léxico a partir de la segunda mitad del siglo XVI. En definitiva, este trabajo pretende aportar una perspectiva empírica al estudio de la vitalidad y pérdida de los arabismos en el español clásico y moderno.

PALABRAS CLAVE: arabismos, ODE, TEITOK, Edad Moderna, periodización.

Annotation and Analysis for the Diachronic Study of Arabisms in Oralia Diacrónica del Español (ODE)

ABSTRACT: The study of Arabic-origin lexicon in the Spanish language has attracted the interest of numerous researchers in the philological field due to the historical contact between Hispanic Arabic and the peninsular Romance languages for almost eight centuries. Works published on this matter have primarily focused on the medieval period (Neuvonen 1941), while data from the sixteenth century onwards has not received as much attention. For classical and modern Spanish, studies have particularly centered on the analysis of specific cases (Calderón Campos 2010, Morala Rodríguez 2012a) and on the substitution, loss, or obsolescence of Arabisms in very specific periods (Giménez-Eguíbar 2015). However, there is a lack of comprehensive studies spanning four centuries of history that provide quantitative data regarding the greater or lesser presence of Arabisms in our language.

It is evident, in this regard, that there has been little concern for preparing materials aimed at the study of Arabisms and their evolution over time. Thus, the historical reference corpora (CORDE, CDH), typically used to support the aforementioned studies, do not include any lexical-level annotation that facilitates the automatic retrieval of etymological information, making its statistical treatment challenging. For all these reasons, this work has a dual objective. Firstly, the methodology adopted to automatically retrieve all Arabisms contained in ODE (Oralia Diacrónica del Español) is explained, which is a specialized historical corpus consisting of texts with communicative immediacy produced between the 16th and 19th centuries. Secondly, the frequencies of use of Arabisms obtained from this corpus are analyzed and represented in a dendrogram that reflects a clear turning point in the use of this type of lexicon from the second half of the 16th century onwards. In conclusion, this work aims to provide an empirical perspective on the study of the vitality and loss of Arabisms in classical and modern Spanish.

KEYWORDS: Arabisms, ODE, TEITOK, Modern Age, periodization.

1. Introducción

El contacto de las variedades romances peninsulares y del árabe hispánico desde el 711 hasta la total expulsión de los moriscos ha suscitado enorme interés entre la comunidad científica. No obstante, a pesar de la existencia de cientos de estudios al respecto, faltan investigaciones apoyadas en datos cuantitativos que vengan a respaldar las líneas que se han arrojado sobre la historia del arabismo en lengua española.

En este estudio intentamos ir un paso más allá a través de un objetivo doble. Por un lado, explicamos detalladamente la metodología adoptada para la recuperación automática de los arabismos contenidos en el corpus ODE, un corpus histórico especializado e integrado por textos de inmediatez comunicativa producidos entre el siglo XVI y el XIX. En segundo lugar, analizamos las frecuencias de uso de arabismos obtenidas y aportamos algunas estadísticas descriptivas, incluyendo la generación de un dendrograma. En definitiva, este trabajo pretende aportar una perspectiva empírica al estudio de la vitalidad y pérdida de los arabismos en el español clásico y moderno.

Hasta ahora, muchas investigaciones han descrito cómo el estudio del componente léxico en perspectiva general resulta de por sí complejo, dado que se trata del plano de la lengua más superficial, variable y menos sistematizable. Estos trabajos han de apoyarse necesariamente en fuentes documentales de todo tipo que permitan rastrear todos los fenómenos lingüísticos asociados a un término para presentarlos en grandes diccionarios históricos: acepciones, ampliaciones y reducciones semánticas, sustituciones léxicas, periodos de incorporación a la lengua, cuestiones etimológicas, ejemplos reales de uso, etcétera. Por ello, el estudio histórico del léxico a través de corpus lingüísticos elaborados a partir de documentos alejados de lo culto es una de las metodologías más fructíferas actualmente. Esos fondos contribuyen a ampliar y completar el caudal léxico de la lengua, registrando formas que van más allá de lo que podemos considerar que constituye el léxico más normalizado y académico (Morala Rodríguez 2012b: 200).

En el caso concreto del léxico de origen árabe, se pueden observar problemas específicos tales como la variación ortográfica tan acusada que muestran los documentos donde aparecen, dificultando la labor de presentar todas las variantes que se asocian a un mismo arabismo. De forma adicional, dado que en muchos corpus han primado documentos vinculados a registros formales, no había sido posible analizar qué sucede con este tipo de voces en otros documentos hasta hace poco.

Teniendo en cuenta todo ello, en este estudio se exponen algunas de las ventajas que ofrecen las nuevas metodologías en las humanidades digitales y que han venido a mejorar la investigación de los arabismos en lengua española a través de la tecnología que se ha empleado en la elaboración del corpus Oralia Diacrónica del Español (ODE). La metodología y la tecnología que operan sobre este corpus se cimentan en la aplicación de una de las iniciativas que está teniendo gran acogida entre los investigadores dentro del campo de la lingüística de corpus: el uso del estándar internacional TEI (Text Encoding Initiative) para marcar y estructurar documentos con el lenguaje XML (eXtensible Markup Language). Adicionalmente, la plataforma que se está convirtiendo en referencia de uso para alojar corpus es TEITOK, por estar específicamente diseñada para tal fin y cuyo funcionamiento se basa en el empleo de transcripciones documentales en XML-TEI.

Con ello, conseguiremos aportar un estudio pionero sobre arabismos en un corpus determinado que permita contrastar los datos obtenidos con la teoría arrojada sobre sus periodos de uso y de pérdida en lengua española.

2. El estudio de los arabismos en lengua española. Antecedentes

2.1. *Los préstamos léxicos de origen árabe*

El estudio de léxico de origen árabe en lengua española ha suscitado el interés de numerosos investigadores en el campo filológico. En parte, ello se debe a los casi ocho siglos de contacto estrecho que se produce entre las diferentes variedades del árabe y de romances en la península ibérica a partir de la invasión musulmana en el año 711 (Corriente 1977). Este contacto lingüístico propició que, poco a poco, se fuera configurando la base del *hispanoárabe* o *andalusí* (Corriente 2004), al tiempo que la continuación directa del latín traído a la Bética recibió influencia del árabe, variedad que se conoce como *mozárabe* o *romanandalusí*. Así, un contacto temporal tan prolongado explica la enorme cantidad de estudios que existen dedicados al arabismo en español. Por ello, contamos con una tipología de investigaciones sobre este componente muy dilatada y compleja de sintetizar en pocas líneas.

Antes de profundizar en los diferentes tipos de estudios dedicados al arabismo en español, resulta necesario señalar que el análisis del componente léxico de una lengua presenta numerosas dificultades, entre otras cosas, por

ser el plano lingüístico más voluble y proclive a reflejar todo tipo de cambios sociales, culturales, políticos, etcétera, motivados por condicionamientos externos a una lengua concreta en su diacronía. Dejando a un lado los diferentes procedimientos morfológicos con los que cuenta una lengua para la formación de palabras nuevas (la derivación y la composición), el plano léxico se ve sustancialmente multiplicado a través de la adopción de *préstamos*, entendidos como aquellos elementos que una lengua toma de otra (Gómez Capuz 2004).

El concepto de *préstamo* ha sido profundamente tratado y discutido desde múltiples enfoques, tanto formalistas y estructuralistas como sociolingüísticos. Cada uno de estos enfoques ha dado lugar a diferentes clasificaciones en torno al préstamo lingüístico. Por ejemplo, Bloomfield (1933) estableció una dicotomía entre *préstamos dialectales* e *íntimos*. Más tarde, el formalista Bezt (1949) propuso varios conceptos para esta cuestión, entre los que destacan tres tipos de préstamos léxicos: el *integral*, el *calco léxico* y el *calco semántico*. Asimismo, esta clasificación fue retomada por Haugen (1950) y Weinreich (1953), quienes establecieron los conceptos de *importación* y *sustitución*, los cuales son equivalentes a la diferencia actual de corte europeo entre *préstamo* y *calco*. También se ha propuesto clasificar los préstamos léxicos en función de si han sido adoptados de una lengua fuente o si lo han hecho a través de varias lenguas (*inmediatos* o *remotos*) (Dworkin 2012). Existen, además, otros conceptos asociados a los préstamos, como pueden ser los de *extranjerismo*, *préstamo adaptado* o *préstamo crudo* (Álvarez de Miranda 2009).

Como se ha señalado más arriba, los préstamos suponen, además, un contacto entre lenguas o variedades de lenguas, factor que ha terminado por ser considerado como determinante en la historia del léxico español (Dworkin 2012). De todas las formas de contacto lingüístico existentes, la más frecuente suele ser el contacto social directo, ya sea por razones políticas, económicas, culturales o religiosas. El contacto entre lenguas se mide por la *influencia lingüística* y por la *intensidad de contacto* (García González 2008). Estos dos parámetros de medida han permitido describir situaciones de contacto de diferentes grados de intensidad (alto, medio y bajo).

Teniendo en cuenta lo anterior, en el caso de la incorporación de arabismos al español tales parámetros se manifiestan en la conjunción de aspectos lingüísticos y extralingüísticos. Entre los factores lingüísticos no se pueden dejar de lado procesos fonéticos, como la existencia de un molde fonético apropiado; morfológicos, como los mecanismos de derivación con sufijos y prefijos obsoletos; y semánticos, como sucede en el caso de fenómenos como la sino-

nimia o los conflictos homonímicos (Giménez Eguíbar 2011: 24). Desde el punto de vista extralingüístico, es fundamental tener en cuenta, por ejemplo, las diferencias y roles socioculturales entre musulmanes y cristianos en los diferentes momentos temporales de su contacto; o la introducción de nuevas realidades y conceptos que establecieron los musulmanes en el comercio, la organización social o militar (Lörinczi 1969: 65, Oliver Pérez 2004: 1076). De hecho, son estos factores extralingüísticos los que pusieron sobre la mesa la importancia que juega el *prestigio* en la adopción de voces de otras lenguas (Gimémez-Eguíbar 2016).

Con todo, nos interesan aquí los préstamos léxicos integrales o totales, adoptados del haz dialectal andalusí y cuya integración a los romances peninsulares ha conllevado ciertas alteraciones fonéticas y morfológicas debidas a la enorme distancia estructural que existe entre el sistema lingüístico semítico y el latino (Giménez-Eguíbar 2024). Esa distancia entre lenguas ha ocasionado un inmenso polimorfismo ortográfico en muchos arabismos, que, asimismo, se explica por su vía de penetración eminentemente oral al español (Giménez Eguíbar 2024: 363). Por una parte, dicha inserción se produjo a partir del siglo VIII de la mano de los cristianos que permanecieron en territorio musulmán en sus contactos con árabes y bereberes. Esas conexiones ocasionaron un lento pero progresivo proceso de arabización e, incluso, de bilingüismo entre los mozárabes (García González 2008, Steiger 1932). Este estado de cosas se vio interrumpido posteriormente por la caída del Califato de Córdoba y por la llegada de musulmanes fundamentalistas desde finales del siglo XI, acontecimientos que ocasionaron la huida masiva de mozárabes hacia el norte peninsular, donde buscaron refugio en los incipientes reinos cristianos que, poco a poco, avanzaban en su reconquista. De este modo, durante los siglos XII, XIII y, en parte, el XIV, se constata otra de las vías de contacto entre el árabe y el romance (Arié 1993). Numerosos investigadores señalan este periodo como el de mayor incorporación de arabismos en ámbitos designativos como la milicia, la administración y la economía (García González 2008)[1].

Por otra parte, a medida que los cristianos recuperaban territorios de los musulmanes, los mudéjares fueron abandonando su lengua en favor del romance y, con ello, el número de arabismos fue decreciendo. Este hecho se

[1] Paralelamente, los arabismos comienzan a documentarse en todas las obras del legado alfonsí, ampliando sustancialmente los campos léxico-semánticos en los que aparecen dada la variedad temática que este presenta.

suele periodizar entre los siglos XIV y XV y se ha señalado, además, que esos arabismos fueron, sobre todo, tecnicismos. A partir de finales del siglo XV se establece el momento en el que el hispanoárabe y los arabismos comienzan un proceso de obsolescencia: "desde finales del siglo XV, y sobre todo desde el siglo XVI, se presenta un momento de decadencia en la introducción y uso de léxico de origen árabe […] debida al desprestigio lingüístico de dicha lengua, que se vincula con la religión musulmana y con el enemigo vencido" (González Sopeña 2019: 59). Además, se aducen al respecto de esa decadencia hechos históricos como la conquista final del reino nazarí de Granada, la colonización americana y la llegada de las corrientes renacentistas italianas (Colón Domènech 2002).

Si nos atenemos a todo lo expuesto, se viene afirmando que desde 1500 los arabismos comenzaron un proceso de pérdida y sustitución léxicas. Si durante la Edad Media los arabismos léxicos se insertan en campos léxico-semánticos vinculados a grandes artes y disciplinas, como la medicina, la agricultura o la matemática, así como a ámbitos institucionales y administrativos, se observa que, a partir de la Edad Moderna, los ámbitos designativos en los que siguen apareciendo arabismos se corresponden con artes menores y de la vida cotidiana (cerámica, cuero, objetos culinarios, vestimenta) (Maíllo Salgado 1997: 90)[2]. Por ello, esa afirmación es aceptable en términos generales, pero es, asimismo, matizable cuando se toman en consideración tipos textuales alejados de lo culto y formal, como es el caso de este trabajo.

2.2. *Estudios en torno a los arabismos en lengua española*

En esta sección se plantea un breve recorrido de toda la producción científica en torno al elemento árabe en lengua española. De forma esquemática, se puede establecer una sucinta tipología de los estudios sobre arabismos léxicos en torno a todo lo exhibido:

a) Investigaciones sobre las épocas de mayor adopción de arabismos y sobre los campos léxico-semánticos en lo que se ha incorporado.

[2] Así, por ejemplo, los territorios del antiguo reino de Granada se erigen como un enclave en el que se mantuvieron e incluso se incorporaron términos árabes relativos a múltiples aspectos de la vida cotidiana tras dos siglos desde su conquista (González Sopeña 2019).

b) Investigaciones en relación al decrecimiento en la adopción de léxico de origen árabe y de su competencia o reemplazo por otras voces (sustituciones léxicas).
c) Investigaciones sobre los procesos fonéticos específicos en la adopción de palabras árabes, así como sobre la intensa variación ortográfica que muestran estas palabras en los documentos desde la Edad Media.

Estos tres tipos de investigaciones se concretan en decenas de artículos, monografías y manuales que se erigen sobre las diferentes etapas de la lengua española. Un grueso importante de estudios sobre arabismos se circunscribe en torno a la Edad Media, por ser este el periodo de contacto más o menos directo, y, por tanto, el de mayor incorporación de estas palabras (Oliver Pérez 2004), siempre en relación con los condicionamientos sociales que se desplegaron en la reconquista cristiana. Así, hay estudios que analizan los arabismos de las obras alfonsíes, tratados y crónicas medievales (García González 1998, Maíllo Salgado 1998, Neuvonen 1941, Pocklington 1984, Oliver Pérez 2005-2006) o el papel que jugaron los mozárabes en su integración al español (García González 2007).

Los trabajos sobre arabismos léxicos también se han enmarcado en los inicios de la Edad Moderna y se han centrado en buena medida en aspectos sobre la pérdida u obsolescencia de arabismos por el desprestigio cultural hacia el mundo musulmán que trajo la llegada del humanismo y la vuelta al mundo clásico grecolatino (Walsh 1967), así como sobre su competencia o relevo por otro término de origen no árabe (Giménez-Eguíbar 2015, 2016). Otro tipo de investigaciones sobre arabismos en el Siglo de Oro ha puesto el acento en voces escasamente documentadas o que solo aparecen y se mantienen en campos léxico-semánticos muy específicos, relacionados con alguna actividad económica, la industria textil o los enseres domésticos (Calderón Campos 2010, Serrano Niza 2007, González Sopeña 2017, Morala Rodríguez 2012a, Puche Lorenzo 2012). A todo esto, habría que añadir todos los artículos que ponen el foco sobre arabismos en una región o zona concreta hispanohablante, trabajos de corte dialectal, tanto sincrónicos (Garulo Muñoz 1983) como diacrónicos (Torres Montes 1996). En el periodo áureo también vieron la luz obras de carácter lexicográfico, ligadas a tareas de evangelización, que marcaron importantes precedentes en materia de arabismos en siglos posteriores; sirvan de ejemplo el *Vocabulista arábigo* de Alcalá (1505), el *Compendio de algunos vocablos arábigos* de López Tamarid (1585) o la *Recopilación de algunos nombres arábigos* de Guadix (1593).

Menos estudios sobre arabismos se observan en el siglo XVIII, como el de Torres Martínez (2014). Esta falta se debe, en parte, al auge en la entrada de galicismos durante el siglo ilustrado. No obstante, desde el siglo XIX la lexicografía ha destacado en la elaboración de diccionarios y glosarios dedicados a los arabismos y a otros elementos léxicos de corte oriental. Así, sobresalen los compendios de Dozy y Engelmann (1869) o el de Eguílaz y Yanguas (1886). Este conocimiento tan amplio sobre los arabismos se completa, además, con toda la información lexicográfica que aparece en los diccionarios generales del español (como los de la Academia), etimológicos (DCECH) y específicos para este tipo de palabras (Corriente 1999 y 2008). Se suman, asimismo, las descripciones generales sobre este tipo de voces que se hallan en manuales de historia de la lengua, tanto clásicos (Lapesa 1981 [1942]) como capítulos monográficos muy recientes (Giménez-Eguíbar 2024).

Sin duda, el enorme fondo diacrónico que presenta este tipo de léxico ha ocasionado la descomunal cantidad de estudios descritos. En toda esta línea de investigación sobre léxico de origen árabe, apoyada en bases documentales valiosísimas para el historiador de la lengua, se pierde mucha información específica sobre arabismos léxicos. Esto se debe a las deficiencias tecnológicas que presentan muchos de los corpus lingüísticos (CORDE, CDH, CdE) que han servido de base empírica para muchos de los estudios señalados. Otros de los trabajos mencionados están directamente sustentados en fuentes textuales inéditas no digitalizadas. Actualmente, la filología ha dejado de trabajar de esa forma, dado que, cada vez más, la comunidad científica exige y requiere que los datos de las diferentes investigaciones sean de acceso abierto y en línea.

De forma global, una de las nociones que emana de toda la cantidad de estudios revisados se concreta en la idea de que los arabismos comenzaron a sufrir un proceso de obsolescencia desde el siglo XVI, pero ¿es posible mantener tal afirmación si nos apoyamos en datos de corpus elaborados a partir de documentos que están fuera del circuito formal?, ¿qué nos dicen los datos del corpus ODE?, ¿es posible matizar este estado de la cuestión de forma empírica? La última parte de este estudio mostrará con datos empíricos una periodización de arabismos a lo largo de los cuatro siglos que conforman la Edad Moderna sobre una base documental que no ha recibido toda la atención necesaria para plantear una historia del arabismo más consistente.

3. Metodología

3.1. *Las características del corpus*

Todos los datos aportados en este trabajo están tomados del corpus Oralia Diacrónica del Español (ODE), desarrollado por el grupo DiLEs (Diacronía de la Lengua Española) en la Universidad de Granada. ODE es 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 en un marco temporal que abarca desde 1492 hasta 1900. El corpus, que es de acceso libre y gratuito en la red, está compuesto actualmente por documentación archivística inédita relativa a tres tipos textuales: inventarios de bienes, declaraciones de testigos en juicios criminales y certificaciones periciales de cirujanos sobre personas heridas o fallecidas[3].

Estos tres tipos de texto presentan ciertas características que los acercan a registros más coloquiales por la impronta oralizante que muestran, dado que todos ellos "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 del escribano" (Calderón Campos y Vaamonde 2020: 169). La forma de proceder en la redacción de esos textos es bastante similar: los escribanos van tomando nota de lo que un declarante dice como buenamente lo entendían, sin ningún tipo de pretensión artística. Esto supone la plasmación de hábitos lingüísticos dialectales que desaparecen en los documentos más formales y nos acerca, por tanto, a la realidad lingüística de una época y de un lugar de una forma más fiable. El corpus cuenta actualmente con algo más de un millón de palabras, cuya distribución por tipo textual es la que se recoge en la tabla 1.

ODE supuso una continuación del Corpus Diacrónico del Español del Reino de Granada (CORDEREGRA), compuesto por documentación producida entre 1492 y 1833 en las actuales provincias de Málaga, Granada y Almería. Concretamente, en ODE se amplió la extensión geográfica a otras zonas de Andalucía y del centro y norte peninsulares (Sevilla, Huelva, Cádiz, Madrid, Badajoz, Burgos, etcétera), así como el arco temporal a todo el siglo XIX. Además, ODE ha experimentado actualizaciones tecnológicas con respecto a su predecesor, al haberse incorporado el uso de estándares de codificación (TEI),

[3] La página electrónica del corpus es http://corpora.ugr.es/ode/.

Tabla 1
Distribución de palabras por tipo textual en ODE

Tipo de texto	%
inventarios de bienes	71,74
declaraciones de testigos	20,87
certificados médicos	6,07
otros	1,32

de anotación (EAGLES) y de recuperación de información (CQL), y al haberse centralizado las tareas de gestión y procesamiento lingüístico del corpus en la plataforma TEITOK (Janssen 2016).

Pueden encontrarse explicaciones detalladas sobre el uso de estos lenguajes y herramientas y, en general, sobre el proceso de construcción de ODE en Calderón Campos (2019), Calderón Campos y Vaamonde (2020) y Vaamonde (2024). En este apartado, nos limitaremos a explicar brevemente las características tecnológicas y metodológicas del corpus que resultan de especial interés en el contexto de la investigación que planteamos en este estudio, esto es, con el objeto de tener una mejor comprensión del proceso de anotación y posterior análisis de los arabismos documentados en ODE. Estas características pueden resumirse en al menos tres puntos esenciales: (i) la marcación de los datos en XML-TEI, (ii) la integración de diferentes capas de información al nivel de la palabra y (iii) la recuperación de información mediante sintaxis CQL.

Partimos del hecho de que todos los datos contenidos en ODE han sido codificados con el lenguaje de marcado XML y aplicando los estándares propuestos por el Consorcio internacional TEI (Text Encoding Initiative) para la representación de textos en formato electrónico. 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 2010: 226-227; Allés-Torrent 2015: 19). Conviene señalar, no obstante, que hasta hace relativamente poco tiempo el empleo de este estándar para la anotación de corpus históricos era una práctica más bien excepcional, como nos recuerda Kytö:

> The searchability of a corpus is crucially dependent on how the corpus has been annotated. Again, there is a lack of consensus on this point, and compilers of historical corpora have been slow or even reluctant to apply standards such as the Text Encoding Initiative (TEI) Guidelines (P5). Many of the better known corpora are annotated for the main textual features but not all, and not as exhaustively as could have been the case (Kytö 2011: 437).

A esta reticencia al uso de estándares de codificación en la construcción de corpus históricos debemos sumar, además, el hecho de que en España el uso de TEI ha sido en general bastante minoritario. En un trabajo de 2010, apuntaba Fradejas Rueda que "no se ha desarrollado en España esta nueva tecnología" (Fradejas Rueda 2010: 225) y cinco años más tarde sostenía Allés-Torrent que "[e]n España, el uso de TEI se remonta a algunas décadas, aunque no está todavía extendido" (Allés-Torrent 2015: 19). En la actualidad, esta situación parece estar cambiando y la comunidad de TEI en español empieza a dar muestras de cierta vitalidad (Del Rio Riande y Allés-Torrent 2023). En lo relativo al diseño de corpus históricos, en los últimos diez años han ido surgiendo algunas propuestas, herramientas y proyectos de investigación relacionados con el modelado de datos en TEI, como la publicación de la *Guía para editar textos CHARTA según el estándar TEI* (Isasi Martínez *et al.* 2014); la construcción del corpus P. S. Post Scriptum (CLUL 2014), cuyos datos pueden ser descargados enteramente en la versión P5 de este estándar; o la herramienta TEITOK (Janssen 2016), diseñada precisamente para crear corpus que combinan marcación en TEI con anotación lingüística.

Siguiendo la estela y la metodología establecidas por estos recursos se sitúa el corpus ODE, que también aprovecha las ventajas inherentes a este estándar. Cada documento de texto en ODE constituye un archivo XML compuesto por dos bloques claramente diferenciados (figura 1): por un lado, la cabecera (<teiHeader>), que contiene la marcación estructurada de diferentes metadatos, como son el título del documento (<title>), la ubicación del manuscrito (<msIdentifier>), la tipología del texto (<textClass>) o su contextualización espaciotemporal (<setting>); por otro lado, el texto propiamente dicho (<text>), cuya transcripción se ha llevado a cabo siguiendo un enfoque paleográfico: se respetan la disposición, la ortografía y las particularidades visuales del manuscrito, incluyendo cancelaciones (<del>), adiciones (<add>) o cambios de línea (<lb/>), entre otros aspectos, creando así un edición académica digital que permita preservar la autenticidad de los manuscritos seleccionados para el corpus.

```
<teiHeader>
   <fileDesc>
      <titleStmt>
         <title>Inventario post mortem de los bienes que quedaron a la muerte
         de doña Francisca Javiera de Salamanca y Miranda</title>
         [...]
      </titleStmt>
         [...]
      <sourceDesc>
         <msDesc>
            <msIdentifier>
               <country>España</country>
               <settlement>Burgos</settlement>
               <institution>Archivo Histórico Provincial de Burgos</institution>
               <repository>Protocolos Notariales</repository>
               <idno>AHPBU 7096</idno>
            </msIdentifier>
         [...]
         </msDesc>
      </sourceDesc>
   </fileDesc>
         [...]
   <profileDesc>
         [...]
      <settingDesc>
         <setting>
            <name type="place" subtype="Burgos">España, Burgos, Burgos</name>
            <date when="1764" when-custom="XVIII">1764</date>
         </setting>
      </settingDesc>
      <textClass>
         <catRef target="inv"/>
      </textClass>
   </profileDesc>
</teiHeader>
<text>
   <body>
         [...]
      <lb/> Siette almuadas de terziopelo
      <del rend="overwritten">c</del><add place="inLine">c</add>arme<lb/>si
      guarnezidas con galon de oro y bor<lb/>las de
      seda en quattrocientos sesenta y dos r<add place="above">s</add>.
      <lb/> Ottra de cañamazo con borlas de seda <lb/> en doze rreales.
      <lb/> Vna alfombra grande de lana de differen<lb/>ttes colores en
      seiscienttos rr<add place="above">s</add>.
         [...]
   </body>
</text>
```

Figura 1
Ejemplo simplificado de documento XML-TEI en ODE. Elaboración propia

Como ya se ha mencionado, una de las ventajas fundamentales de contar con datos debidamente marcados en XML-TEI es que este lenguaje no solo articula el contenido del documento, sino que añade la posibilidad de su poste-

rior recuperación. Por ejemplo, a partir de un conjunto de datos que haya sido codificado de acuerdo con la estructura recogida en la figura 1 resulta relativamente sencillo filtrar aquellos documentos que sean inventarios de bienes (<catRef target="inv"/>) o que hayan sido producidos en un lugar determinado (<name type="place" subtype="Burgos">) o en una fecha concreta (<date when="1764">). De modo análogo, y siguiendo con el ejemplo de la figura 1, es posible recuperar aquellos fragmentos de texto sobre los que el escribano haya sobrescrito otra expresión (<del rend="overwritten">) o que hayan sido añadidos por encima del renglón (<add place="above">).

Repárese, no obstante, en que el texto, tal como aparece en la figura 1, contiene anotación textual, pero carece de anotación lingüística. Dicho de otro modo, se trata de una edición diplomática del manuscrito, pero no de un corpus anotado, lo que impide realizar consultas más sofisticadas orientadas al análisis lingüístico de palabras o de expresiones más complejas. La conversión de la edición digital en un corpus anotado se realiza a través de la plataforma TEITOK. Esta plataforma integra diferentes herramientas de procesamiento lingüístico que son ejecutadas secuencialmente sobre el contenido del elemento <text>. Para el caso de ODE, se hace uso de un *tokenizador*, un normalizador ortográfico, un etiquetador morfosintáctico y un lematizador.

Para almacenar la información resultante de cada una de estas tareas de procesamiento, TEITOK se vale de una lógica de marcación muy sencilla: cada *token* —esto es, cada palabra o signo de puntuación— se delimita mediante un elemento <tok> y toda la anotación lingüística correspondiente se representa mediante atributos contenidos en dicho elemento. Por ejemplo, la expresión *Siette almuadas de terziopelo carmesi*, que da inicio al fragmento de texto recogido en la figura 1, devolvería el resultado que se recoge en la figura 2, una vez aplicadas las mencionadas tareas de procesamiento lingüístico en TEITOK.

Como se infiere de la figura 2, el atributo @nform (*normalized form*) indica la forma de la palabra con grafía normalizada; de esta manera, las formas originales *Siette*, *almuadas*, *terziopelo* y *carmesi* han sido normalizadas, respectivamente, a *Siete*, *almohadas*, *terciopelo* y *carmesí* (esta última con tilde). El atributo @pos (*part of speech*) indica la etiqueta morfosintáctica, basada en la propuesta del grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas (Leech y Wilson 1996); por ejemplo, la forma *almuadas* se ha anotado con las siglas NCFP000, que indican Nombre, Común, Femenino y Plural. Por su parte, el atributo @lemma indica la

```
<tok nform="Siete" pos="Z" lemma="7">
   Siette
</tok>
<tok nform="almohadas" pos="NCFP000" lemma="almohada">
   almuadas
</tok>
<tok pos="SPS00" lemma="de">
   de
</tok>
<tok nform="terciopelo" lemma="terciopelo" pos="NCMS000">
   terziopelo
</tok>
<tok form="carmesi" nform="carmesí" pos="AQ0MS0" lemma="carmesí">
   <del rend="overwritten">c</del><add place="inLine">c</add>arme<lb id="e-533"/>si
</tok>
```

Figura 2
Fragmento de texto procesado en TEITOK. Elaboración propia

forma lematizada, esto es, tal como aparece en la entrada de un diccionario. Finalmente, para aquellas palabras que en la edición digital han sido marcadas con algún elemento en XML, se hace uso de un atributo @form, que recoge la forma original correspondiente libre de lenguaje de marcas, facilitando así su recuperación y procesamiento en el corpus; esto último es lo que sucede, por ejemplo, con la forma *carmesi*, cuya transcripción paleográfica ha requerido el uso de los elementos <del> (para marcar una primera grafía "c" que ha sido cancelada), <add> (para marcar una segunda grafía "c" que se ha sobrescrito sobre la primera) y <lb/> (para marcar un cambio de línea entre las grafías "e" y "s"). Toda esta información lingüística, así como la propia estrategia de marcación basada en la adición de pares atributo-valor para cada *token*, serán fundamentales para la anotación de arabismos en ODE. Volveremos sobre ello en el apartado 3.2.

Por lo que se refiere al tercer y último aspecto tecnológico, relativo a la recuperación de información, el sistema TEITOK realiza una conversión automática del contenido de los archivos XML en un corpus que sigue el formato CWB (Corpus WorkBench) (Christ *et al.* 1999). La principal ventaja de este formato radica en su capacidad para llevar a cabo búsquedas utilizando la sintaxis CQL (Corpus Query Language), un lenguaje de consulta diseñado específicamente para recuperar información de manera eficiente de un corpus lingüístico. Debido a su amplia gama de funcionalidades, este lenguaje ofrece una gran flexibilidad para realizar consultas avanzadas y sofisticadas, aunque en esencia se basa en una lógica similar a la ya expuesta; esto es, se trata de seleccionar atributos y valores específicos para cada *token* en el corpus.

En ODE, una consulta sencilla que permite ilustrar la mecánica de este lenguaje sería la obtención de todos los usos registrados de una determinada palabra en el corpus. Imaginemos, por ejemplo, que queremos obtener todas las ocurrencias de la palabra *almohada*, con independencia de cómo aparezca escrito este término en su forma original manuscrita. Nótese que lo que queremos obtener, en realidad, es el contenido de todos aquellos *tokens* que hayan sido normalizados ortográficamente a *almohada*, lo que equivale a recuperar todos los *tokens* que contengan un atributo @nform con el valor *almohada*. La sintaxis CQL correspondiente a esta consulta sería, por tanto, la siguiente:

```
[nform="almohada"]
```

También podemos filtrar la búsqueda en función de algún metadato que nos interese especialmente. Por ejemplo, podemos repetir la consulta anterior, pero aplicarla solo al subcorpus de documentos de ODE producidos durante el siglo XVIII:

```
[nform= "almohada"] :: match.text_century = "XVIII"
```

Conviene señalar que el resultado de estas consultas incluirá ejemplos que contienen la palabra *almohada*, pero también otros que contienen *almoada*, *almuada*, *halmoada* o *armuada*, todas ellas voces documentadas en ODE y que, obviamente, han sido normalizadas a la forma estándar actual *almohada*. En otras palabras, los diferentes niveles de anotación con que es enriquecido cada *token* del corpus permiten ejecutar consultas basadas en una relación de uno a varios. Esta relación se establece, por ejemplo, entre forma normalizada y formas originales (*almohada* vs. *almoada*, *armoada*, *halmoada*...), o entre lema y formas normalizadas (*almohada* vs. *almohada*, *almohadas*), o entre etiquetada morfosintáctica y lemas (nombre común vs. *almohada*, *lienzo*, *sábana*...), entre otras posibilidades.

Las ventajas que tiene esta estrategia para la anotación de arabismos son evidentes. Al momento de redactar estas líneas, el arabismo *almohada*, en singular, presenta 214 ocurrencias en ODE, distribuidas en hasta 9 formas ortográficas diferentes; en el caso del plural, las cifras son bastante más elevadas: 1106 ocurrencias distribuidas en 21 formas ortográficas diferentes. Repárese en que estamos hablando de 1320 casos relacionados con un único lema, *almohada*, que puede aparecer escrito de 30 formas diferentes (ver tabla 2).

Tabla 2
Frecuencia absoluta y variación ortográfica del lema almohada *en ODE*

singular	N	Plural	N
almohada	142	*Almohadas*	651
almoada	52	*Almoadas*	339
almuada	12	*Almuadas*	74
almoha	3	*halmohadas*	11
halmohada	1	*halmoadas*	7
almoda	1	*Almojadas*	3
almada	1	*almuhadas*	2
halmoada	1	*Almodas*	2
armuada	1	otros 13 casos	17
TOTAL	214	TOTAL	1106

Esta estrategia de marcación (XML-TEI) y posterior recuperación (CQL), siempre a partir de una lógica de atributos y valores que clasifican la información lingüística de cada *token*, constituye la base sobre la que se ha realizado el proceso de anotación de arabismos que planteamos en este trabajo. Al contar con una ortografía estandarizada de cada *token*, se obtiene una capa de anotación uniforme para identificar y analizar las formas léxicas de las palabras, lo que simplifica el proceso de lematización. Y al contar con un corpus lematizado, se abre la posibilidad de contrastar cada lema con un lemario externo proveniente de un diccionario, lo que facilita la realización de anotaciones automáticas ulteriores. Así, se puede utilizar esta técnica para identificar y marcar los lemas que son arabismos, mediante la comparación con una lista previamente establecida de palabras de origen árabe.

3.2. *La anotación de arabismos*

La idea central que cierra el apartado anterior es que la anotación lingüística de un corpus establece una base sólida para agregar información lingüística

adicional, especialmente en el caso de los corpus históricos construidos sobre ediciones diplomáticas de documentación manuscrita, propensos a contener múltiples realizaciones ortográficas de una misma voz. Siguiendo con el ejemplo del arabismo *almohada*, la ventaja de esta estrategia es que no resulta necesario identificar las 1320 ocurrencias de esta palabra, en sus diferentes variantes ortográficas y gramaticales atestiguadas en ODE, sino que basta con identificar una única forma: el lema, esto es, el valor del atributo @lemma compartido por todos esos *tokens* (es decir, *almohada* o, en lenguaje CQL, [lemma="almohada"]).

Asumida esta idea, el siguiente paso es obtener una lista de arabismos lo más amplia y precisa posible, preferentemente en su forma lematizada, que pueda ser usada para contrastarla con los datos del corpus e ir anotando automáticamente los lemas coincidentes. En lo que atañe a este trabajo, esa lista ha sido extraída usando la herramienta IEDRA (Rodríguez Alberich 2020), que permite generar y descargar listas de palabras que cumplen determinados criterios lexicológicos a partir de la información contenida en la 23.ª edición del *Diccionario de la Lengua Española* (DLE). Así, usando el buscador de IEDRA se han obtenido todos los lemas del DLE en cuya información etimológica se incluye alguna de las tres marcas lexicográficas siguientes: *ár.* (árabe; por ejemplo, la forma *aceituní*), *ár. hisp.* (árabe hispánico; por ejemplo, la forma *naranja*) o *ár. clás.* (árabe clásico; por ejemplo, la forma *alcuza*)[4].

Como resultado de este proceso, se han obtenido cuatro listas de palabras en formato TXT. Usando un entorno UNIX, se han fusionado todos estos datos en un único archivo (*arabismos.txt*), eliminando de paso posibles duplicados; por ejemplo, la palabra *almohada* aparecía recogida tanto en la lista correspondiente al árabe hispánico como en la correspondiente al árabe clásico. La orden ejecutada para lograr este resultado se aporta en la figura 3:

```
cat archivo1.txt archivo2.txt archivo3.txt archivo4.txt | sort | uniq > arabismos.txt
```

Figura 3
Fusión de archivos y eliminación de duplicados en UNIX. Elaboración propia

[4] Se ha optado por excluir del proceso de búsqueda y descarga en IEDRA la marca lexicográfica *ár. marroquí* (árabe marroquí), en tanto que todos los arabismos marcados con dicha etiqueta proceden del árabe o del árabe clásico, lo que crearía redundancias innecesarias.

El contenido del archivo *arabismos.txt* devolvió un total de 1270 lemas. De forma manual, no obstante, se añadieron algunos lemas más a esta lista, bien por tratarse de voces de origen árabe que no aparecían recogidas en el DLE (por ejemplo, *alcofa* o *alguidar*), bien por tratarse de palabras que sí están recogidas en el diccionario, pero no aparecían marcadas como arabismos al momento de redactar estas líneas (por ejemplo, *tabí*)[5].

Cabe apuntar, finalmente, que la lista resultante no resuelve los casos de homonimia, esto es, no incluye el índice numérico diferenciador para el caso de las voces homónimas, puesto que las listas generadas en IEDRA a partir del buscador por marcas lexicográficas no aportan esta información. Como es sabido, el DLE adopta un criterio etimológico para la distinción de homónimos en la macroestructura del diccionario, por lo que la falta de esta información en el lemario utilizado tiene especial repercusión para los intereses de este trabajo. Esta particularidad, de hecho, nos ha llevado a eliminar del archivo *arabismos.txt* algunos lemas, en especial aquellas voces de origen árabe que no están documentadas en el corpus, pero que son formas homónimas de otras que sí lo están y proceden de étimos de otras lenguas.

Un caso paradigmático de esto último es la palabra *real*, que contiene tres entradas diferentes en el DLE (*real1*, *real2* y *real3*). Solo *real3*, con el sentido de "lugar de acampada", es de origen árabe (*rahl*), pero no está documentada en nuestro corpus; en cambio, sí aparece —con una muy elevada frecuencia— *real2*, voz referida al tipo de moneda y derivada del latín *regālis*. Mantener este lema en una lista confeccionada para anotar arabismos en ODE solo produciría, por tanto, un alto número de falsos positivos, de ahí su eliminación. Otro caso similar es el del lema *arar*, también suprimido del lemario base, ya que todas las ocurrencias de esta forma en ODE se refieren al verbo *arar2* ("remover la tierra", del latín *arāre*) y ninguno al sustantivo *arar1* ("tipo de árbol", del árabe clásico *ʻarʻar*).

Finalmente, también se han excluido manualmente de la lista aquellas palabras cuyo último étimo es de origen árabe —y, por tanto, incluyen alguna de las marcas lexicográficas antes mencionadas—, pero que han entrado en la lengua española a través del francés, puesto que entendemos que estas voces han de ser consideradas en realidad como galicismos y no como arabismos. Nos referimos a lemas como *chupa*, *sofá*, *jergón*, *muaré*, *alepín* o *beduino*, entre otros.

[5] Para estos casos se ha tomado como referencia el diccionario de Corriente (2008).

Una vez aplicadas estas modificaciones, el número final de formas contenidas en el archivo *arabismos.txt*, y que hemos utilizado para la anotación de arabismos en ODE, asciende a 1264. Por razones obvias, hemos centrado nuestra atención solo en las categorías léxicas y, más concretamente, en los nombres comunes, los verbos y los adjetivos de nuestro corpus[6]. La anotación se ha realizado de forma automática mediante la creación de un *script* en lenguaje Perl. Básicamente, este *script* busca coincidencias entre las formas contenidas en el lemario externo, por un lado, y los valores contenidos en el atributo @lemma de cada *token* del corpus, por otro lado; en caso de producirse una coincidencia entre ambos conjuntos de datos, el *script* crea un nuevo atributo para ese *token*, que hemos denominado @etags (*etymological tags*), y le asocia el valor *arabism*. De forma algo más detallada, los pasos que ejecuta el *script* se pueden resumir en el pseudocódigo siguiente, que se ha ejecutado de forma iterativa sobre cada uno de los archivos XML contenidos en ODE:

- Paso I. Importa el módulo XML::LibXML para trabajar con documentos XML en Perl.
- Paso II. Abre y lee el contenido del archivo *arabismos.txt* en la variable $txt. Establece el manejador de archivo en modo UTF-8 y luego cierra el archivo. Divide el contenido del texto almacenado en la variable $txt en líneas individuales y luego guarda esas líneas en el arreglo @arabismos.
- Paso III. Abre y analiza el archivo XML proporcionado. Si el archivo no existe o no se proporciona ningún nombre de archivo, imprime un mensaje de error y sale del *script*. Analiza el XML usando el módulo XML::LibXML, establece la codificación en UTF-8 y recupera todos los elementos <tok> del documento XML.
- Paso IV. Itera sobre cada elemento <tok> y extrae los atributos *etags*, *pos* y *lemma*. Omite el procesamiento si el atributo *pos* no empieza por V (verbo), por A (adjetivo) o por NC (nombre común), o si el atributo *etags* ya contiene el valor *arabism* (es decir, si ya ha recibido anotación).
- Paso V. Verifica si el contenido del atributo *lemma* coincide con alguna entrada en el arreglo @arabismos. Si hay coincidencia, actualiza el atributo *etags* en consecuencia (es decir, añade *etags*="*arabism*" dentro del elemento <tok>).

6 Como es sabido, la influencia del árabe en el paradigma de las categorías funcionales del español es prácticamente anecdótica: la preposición *hasta* o las interjecciones *ojalá* y *guay*.

- Paso VI. Abre el archivo de salida y guarda el contenido XML modificado.

El resultado de este *script* se puede observar en el fragmento de texto recogido en la figura 4, idéntico al de la figura 2 pero con la anotación de arabismos ya aplicada. Como se puede ver, el *script* ha incluido el atributo @etags con el valor *arabism* en los *tokens* que contienen el lema *almohada* y el lema *carmesí*, ya que ambas formas engrosan el lemario externo utilizado como base de la anotación.

```
<tok nform="Siete" pos="Z" lemma="7">
   Siette
</tok>
<tok nform="almohadas" pos="NCFP000" lemma="almohada" etags="arabism">
   almuadas
</tok>
<tok pos="SPS00" lemma="de">
   de
</tok>
<tok nform="terciopelo" lemma="terciopelo" pos="NCMS000">
   terziopelo
</tok>
<tok form="carmesi" nform="carmesí" pos="AQ0MS0" lemma="carmesí" etags="arabism">
   <del rend="overwritten">c</del><add place="inLine">c</add>arme<lb id="e-533"/>si
</tok>
```

Figura 4
Fragmento de texto con anotación de arabismos. Elaboración propia

Por último, y una vez aplicado el *script*, hubo que revisar manualmente un número reducido de falsos positivos, que fueron producidos por voces homónimas cuyos diferentes significados aparecen documentados en ODE. Es el caso, por ejemplo, de la forma *zagal*, en que confluyen en español tanto la voz del árabe hispánico *zaḡál[l]* (o quizá la del árabe clásico *zuḡlūl*) como el étimo latino *sagum*, el primero con el significado de "mozo" y el segundo con el de "refajo". Pese a que todas las ocurrencias de este lema fueron anotadas automáticamente como casos de arabismo, la revisión manual permitió corregir casos como los de (1a) o (1b), que se refieren al sentido de "refajo", y mantener la anotación en otros como (1c) o (1d), que se refieren efectivamente al sentido de "mozo":

(1) a. Un zagal de coton nuebo con guarnicion en sesenta r(eale)s.
b. Dos zagales de bayeta de Alconchel encarnados en treinta y cinco r(eale)s.

c. Y estubo este t(estig)o con el d(ic)ho Jua(n) ramirez de Burgos sirbiendole de çagal.
d. Y despues a oido decir que el ttal mozo herido no era dueño ni sagal de d(ic)has yeguas.

Como es fácil de suponer, el proceso de anotación aquí descrito es flexible y, por tanto, extrapolable a otros corpus o a otros campos léxico-semánticos. Si bien el *script* utilizado puede requerir ajustes menores para adaptarse a las características de codificación del corpus que es objeto de análisis o al formato de los datos usados como base de la anotación, el enfoque general sigue siendo el mismo.

4. Resultados

Al momento de redactar estas líneas, el tamaño de ODE asciende exactamente a 1 196 589 *tokens*, que se distribuyen en 53 980 lemas diferentes. Una vez aplicado el *script* y revisado el resultado de la anotación, hemos anotado un total de 11 940 *tokens* como arabismos, que se distribuyen en 309 lemas diferentes. Teniendo en cuenta que el lemario utilizado contiene 1 264 lemas, se han documentado en ODE un 25 % del conjunto de arabismos usados como base de la anotación. Los datos generales se resumen en la tabla 3, que incluye también la suma de los nombres comunes (NC), verbos (V) y adjetivos (A) en ODE:

Tabla 3
Datos generales sobre arabismos en ODE

	tokens	lemas
corpus total	1 196 589	53 980
subcorpus (NC, V, A)	424 139	13 538
arabismos	11 940	309

Los datos relativos al número de *tokens* recogidos en la tabla 3 se han extraído mediante el uso de sintaxis CQL desde la propia interfaz de consulta de ODE: [ ], para el número total de *tokens*; [pos="(NC|V|A).+"] para el

subcorpus objeto de anotación, y [etags=".*arabism.*"] para el número de arabismos. Los datos relativos al número de lemas se extrajeron aplicando una distribución de frecuencia por lemas sobre el resultado de las tres búsquedas anteriores. En lenguaje CQL:

```
Matches = [etags=".*arabism.*"]; group Matches match lemma;
```

Téngase en cuenta que esta última consulta requiere una secuencia de órdenes y, actualmente, su ejecución desde la interfaz solo está disponible para usuarios registrados. No obstante, cualquier usuario puede obtener resultados equivalentes utilizando el menú desplegable "Opciones de frecuencia", que encontrará al final de la página de resultados, tras ejecutar una consulta[7]. Usando esta misma estrategia, se pueden obtener los arabismos con mayor frecuencia de uso en el corpus. En la tabla 4 listamos los 30 arabismos con mayor presencia en ODE.

Tabla 4
Los 30 arabismos más frecuentes en ODE

lema	*tokens*	lema	*tokens*	lema	*tokens*
almohada	1 320	carmesí	223	cenefa	108
maravedí	1 261	aceite	191	guadamecí	102
alcalde	976	bancal	166	cotonía	95
azul	884	algodón	163	taza	95
fanega	797	celemín	161	badana	90
arroba	613	adarme	146	tarima	90
azófar	380	alfombra	132	gasa	86
candil	352	alhaja	129	jarra	83
almirez	302	aljófar	127	marfil	76
alguacil	283	albacea	108	acetre	66

[7] Véase http://corpora.ugr.es/ode/index.php?action=cqp.

También hemos querido analizar en este trabajo la distribución del número de arabismos en diferentes épocas. Para ello, hemos extraído del corpus el total de arabismos por tramos de 25 años. Nuevamente, los datos se han extraído mediante varias consultas en CQL, como la que ofrecemos a continuación para la obtención del primero de los intervalos que se han delimitado. La consulta recogida en (2a) permite obtener el número total de *tokens* entre 1501 y 1525; la recogida en (2b) hace lo propio respecto del número total de *tokens* que fueron anotados como arabismos:

(2) a. [ ] :: int(match.text_year) >= 1501 & int(match.text_year) <= 1525
b.[etags=".*arabism.*"] :: int(match.text_year) >= 1501 & int(match.text_year) <= 1525

La tabla 5 contiene, por cada tramo temporal de 25 años, las frecuencias absolutas (FA) y las frecuencias normalizadas (FN) por millón de casos que se han obtenido a raíz de dichas consultas. Estas frecuencias se han calculado teniendo en cuenta tanto el número de *tokens* como el número de lemas. Por ejemplo, la primera fila de la tabla 5 indica que entre 1501 y 1525 ODE contiene 18 580 *tokens* y 1327 lemas; que, de ellos, se han registrado 80 *tokens* y 34 lemas correspondientes a voces de origen árabe, y que esto supone una frecuencia normalizada de 4306 y 25 622 casos por millón, respectivamente.

Para una mejor interpretación de estos datos, se ha realizado un gráfico de líneas usando la librería *ggplot* de R (figura 5). Se trata de un gráfico de líneas que muestra los datos recogidos en las dos últimas columnas de la tabla 5, es decir, las frecuencias normalizadas correspondientes a los lemas (línea negra en la figura 5) y a los *tokens* (línea gris en la figura 5). Los primeros —los datos referidos a los lemas— nos parecen los más interesantes para poder evaluar la intensidad en el uso de arabismos a lo largo del tiempo, ya que permiten evitar el posible sesgo producido por determinadas palabras que pueden tener un elevado número de ocurrencias en intervalos concretos del corpus. Recogemos, no obstante, ambos conjuntos de datos con el objeto de proporcionar un panorama más completo y contrastante de la distribución de arabismos en el corpus.

Finalmente, hemos querido comprobar si los datos que manejamos permiten establecer segmentos diacrónicos en ODE. Dicho de otro modo, nos preguntamos por la posibilidad de identificar diferentes periodos de tiempo basándonos en la similitud de frecuencias de uso relativas a los arabismos documentados en el corpus. Para tal fin, hemos analizado los datos usando el procedimiento conocido como *variability-based neighbour clustering* o VNC,

Tabla 5
Frecuencias de arabismos por tramos de 25 años en ODE (tokens y lemas)

tramo temporal	FA (ODE)		FA (arabismos)		FN (arabismos)	
	tokens	lemas	*tokens*	lemas	*tokens*	lemas
1501-1525	18 580	1 327	80	34	4 306	25 622
1526-1550	59 585	2 737	1 206	87	20 239	31 786
1551-1575	82 215	3 436	1 160	119	14 109	34 633
1576-1600	85 359	3 651	1 049	103	12 289	28 211
1601-1625	68 547	3 044	535	64	7 804	21 024
1626-1650	96 524	3 444	802	68	8 308	19 744
1651-1675	43 724	2 373	385	58	8 805	24 441
1676-1700	71 095	3 421	536	80	7 539	23 384
1701-1725	171 889	5 450	1 492	111	8 680	20 366
1726-1750	131 726	4 778	1 248	97	9 474	20 301
1751-1775	141 295	5 088	1 386	103	9 809	20 243
1776-1800	120 085	5 578	1 058	106	8 810	19 003
1801-1825	38 919	3 351	275	60	7 065	17 905
1826-1850	46 620	3 652	549	73	11 776	19 989
1851-1875	11 427	1 330	106	35	9 276	26 315
1876-1900	8 999	1 320	125	31	13 890	23 484

propuesto originalmente por Gries y Hilpert (2008). Esta técnica de agrupamiento jerárquico se caracteriza por calcular la similitud entre puntos de datos temporalmente adyacentes y, por tanto, "[It] can be used for categorizing individual temporal data points (e.g. years, decades) into larger historical periods based on the similarity in language use" (Brezina 2018: 237).

Hemos realizado esta técnica usando las frecuencias normalizadas referentes a lemas (última columna de la tabla 5) y el dendrograma resultante es el que se ofrece en la figura 6. La medida de distancia utilizada para la elaboración del dendrograma ha sido la desviación estándar.

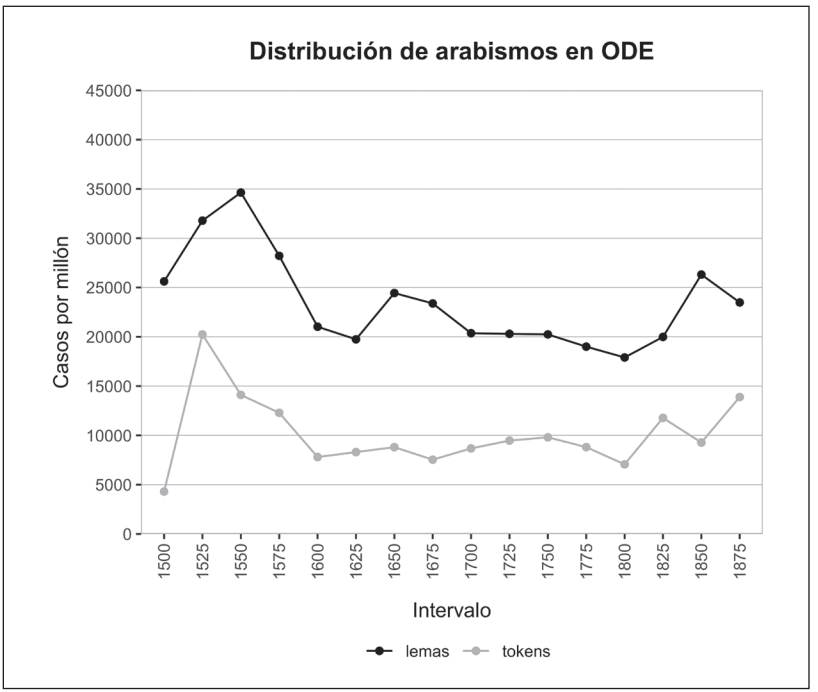

Figura 5
Distribución de arabismos en ODE por tramos de 25 años

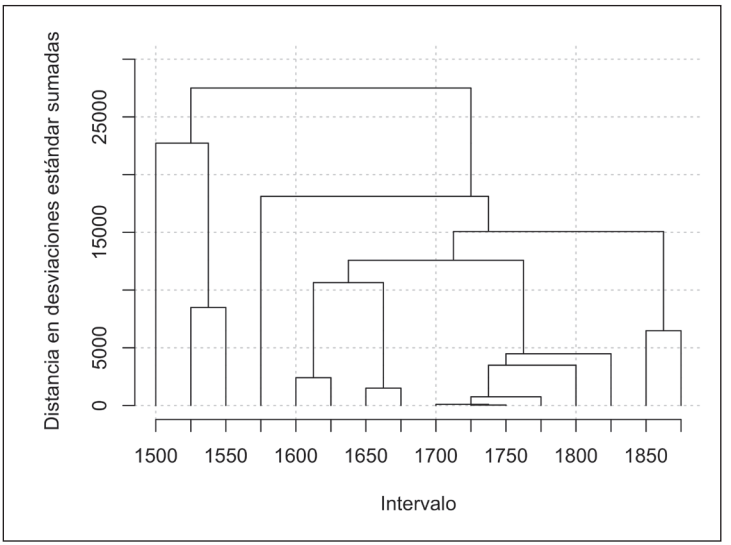

Figura 6
Dendrograma: frecuencia de lemas de origen árabe en ODE

5. Discusión de los resultados

En esta sección se analizan los resultados obtenidos en torno a tres aspectos básicos señalados en el apartado anterior, a saber: (i) los 30 arabismos más frecuentes (tabla 5), (ii) la distribución de los arabismos en tramos de 25 años (figura 5) y (iii) los posibles agrupamientos que se observan en el dendrograma (figura 6).

Con respecto al primer aspecto, despuntan de forma clara dos arabismos por su frecuencia de uso en el corpus: *almohada* y *maravedí*. Ambos se explican por la tipología textual que conforma el corpus y que, de forma mayoritaria, se traduce en relaciones de bienes: inventarios, dotes, testamentos, tasaciones, embargos, etcétera. Así, por lo que se refiere a la voz *almohada*, esta aparece en prácticamente todos los tipos de relaciones citados, junto con su precio tasado en diferentes monedas o valores, como muestran los ejemplos de (3). Acompañamos cada ejemplo del año entre corchetes:

(3) a. Dos **almohadas** de grana en dos mill m(a)r(avedi)s [1586]
b. Dos **almoadas** de rruan labradas con ylo lizado en tres du(cad)os [1620]
c. Quatro **almoadas** con encajes en seis pesos [1708]
d. Quattro **almoad(a)s** de lino desilad(a)s y con encajes en quarentta y cinco r(eale)s [1765]
e. Ytt(em) quatro **almoadas** de crea ancha, q(uaren)ta rr(eale)s v(ellon) [1779]
f. Tres **almohadas** de hilo y una de retor en nueve pesetas [1893]

Por su parte, la alta frecuencia de uso de la voz *maravedí* también es una consecuencia lógica de analizar un corpus compuesto por relaciones de bienes, al incluir estas tasaciones de objetos y ser aquella una medida monetaria de curso legal a lo largo de toda la Edad Moderna. Se documenta, con diferente intensidad, en todos los periodos históricos que abarca ODE:

(4) a. Vna saya amarilla con dos tiras de terçiopelo negro en mill **maravedis** [1550]
b. Un hoçinillo en diezyocho **m(a)r(avedi)s** [1592]
c. Yten, seis mill **m(a)r(avedi)s** de renta de censo 1621]
d. Yt(em) un belon grande vale setenta y cinco rr(eale)s v(ell)on y diez **marauedis** [1769]
e. Dos cuencos en veinte y quatro **m(a)r(avedi)s** [1809]

Continuando con el análisis de los arabismos más frecuentes, cabe señalar que dos de ellos se corresponden con cargos administrativos, muy comunes

también en este tipo de documentación: *albaceas* y *alguaciles*. Los albaceas tienen la función de custodiar los bienes y las últimas voluntades de un testador, mientras que los alguaciles ejercen funciones determinadas por un juzgado o ayuntamiento, siendo sus tareas muy variadas (custodiar tejidos, sembrados u otro tipo de enseres). A excepción del arabismo referido a la moneda del maravedí, hoy desaparecida, los otros tres vocablos citados han llegado hasta nuestros días en el uso general de la lengua.

En la lista proporcionada aparecen, además, otros arabismos que podemos clasificar como pertenecientes a los siguientes ámbitos designativos: colores (*azul*, *carmesí*), objetos de cocina (*jarra*, *almirez*), adornos y joyas (*aljófar*, *alhaja*), telas (*algodón*, *guadamecí*) o unidades de medida (*fanega*, *celemín*), entre otros. Muchos de ellos, como *aceite*, *azul* o *algodón*, son de uso general en la actualidad; otros, en cambio, presentan cierta competencia léxica con otras voces de distinto origen y se encuentran en clara situación de obsolescencia y desuso: *aljófar* (en competencia con *perla*) o *almirez* (en competencia con *mortero*). Podemos, por tanto, volver sobre lo expuesto en el estado de la cuestión en lo referente al mantenimiento de arabismos en campos léxico-semánticos de ámbitos vinculados a la vida cotidiana a partir del siglo XVI. Es esta una cuestión para la que un recurso digital como ODE resulta de especial utilidad, habida cuenta del tipo textual en que este corpus sitúa el foco de atención y, en consecuencia, del tipo de léxico que suele aparecer en sus documentos.

Por otro lado, es necesario señalar el hecho de que mucha de la documentación de ODE está geográficamente adscrita a Andalucía, territorio en el que los musulmanes permanecieron más tiempo y, con ello, su lengua y modos de vida. Cabe recordar que, dentro de Andalucía, la zona oriental prolongó el dominio musulmán hasta 1492, momento el que el antiguo reino de Granada se unió a la corona castellana. Además, los moriscos continuaron viviendo en esa región hasta su expulsión en 1609. Esto ocasiona que los territorios que conformaban el antiguo reino de Granada (Málaga, Granada y Almería) se configuren como una subárea dialectal específica en la que se mantiene bastante léxico de origen árabe (Garulo 1983: 21, Alvar 1996: 256-258). De hecho, a la luz de los datos del siglo pasado contenidos en el ALEA (*Atlas Lingüístico y Etnográfico de Andalucía*), algunos de los arabismos documentados en ODE presentan un reparto geográfico de uso restringido a localidades andaluzas orientales, al lado de otros menos frecuentes. Sirva como ejemplo de estas formas dialectales el caso del arabismo *almocafre*, que aparece repartido

en la zona oriental entre el este de Córdoba, la práctica totalidad de Granada y Málaga, y el sur de Jaén (ALEA, tomo I, mapa 33). Mientras, en el resto de Andalucía el mismo concepto aparece con un término de etimología latina: *escardillo*. Si comparamos el reparto que ambos términos tienen en ODE, observamos que *almocafre* aparece dos veces, una en Jaén y otra en Granada, mientras que *escardillo* presenta 27 casos, 21 de ellos documentados en otras zonas meridionales. Aportamos algunos ejemplos del uso de estas dos voces en (5):

(5) a. Vnas treuedes grandes y otras pequenas. Vn **almocafe**. [1706, Granada]
b. Un escardillo, un **almocafre** y dos picos en seis r(eal)es [1754, Alcalá la Real (Jaén)]
c. Tres asadones y quatro **escardillos**. Una hacha y un martillo [1752, Osuna (Sevilla)]
d. Yttem quatro **escardillos** a cinco rreales vellón cada uno [1786, Medina Sidonia (Cádiz)]
e. Vna azada, un azadón y un **escardillo** [1840, Fiñana (Almería)]
f. Yten vna escubilla de linpiar vieja. Yten vn **escardillo** [1583, Burguillos del Cerro (Badajoz)]
g. Ytt tres tenillas. Ytt vn **escardillo**. Ytt vnas tenazas de errar [1790, Lorca (Murcia)]

Ejemplos similares los observamos en el arabismo *zafa*, que compite con *palangana*; o en el par *almazara* y *molino*. Así, *zafa* se documenta en ODE solamente en la provincia de Almería (11 ocurrencias). En el ALEA (tomo III, mapa 697), el mismo arabismo se extiende solo por el oriente andaluz. Situación idéntica se observa en el caso de *almazara* (ALEA, tomo I, mapa 231). En contraste, *palangana* y *molino* tienen mayor frecuencia de aparición en zonas occidentales de Andalucía.

(6) a. Vna caldera. Vna **safa**. Vn candil [1587, Zurgena (Almería)]
b. Ytt vna **zafita**, barro de Genoba, azul y blanca [1766, Cuevas de Almanzora (Almería)]
c. Dos **zafas** zin baño porcelana, su valor dos pesetas [1896, Canjáyar (Almería)]
d. Una **almazara** con sus artes y pertrechos [1829, Zurgena (Almería)]

Con respecto al gráfico de líneas por tramos de 25 años —y centrándonos en la frecuencia normalizada por número de lemas—, se observa que esta

asciende de forma continuada a lo largo de la primera mitad del siglo XVI, esto es, del tramo 1501-1525 al tramo 1551-1575, momento en el que se documenta el mayor número de lemas de origen árabe por millón de casos. A continuación, la línea comienza a descender de forma ininterrumpida hasta llegar a la primera mitad del siglo XVII (tramo 1626-1650) y, a partir de aquí, puede decirse que se mantiene más o menos estable hasta llegar al siglo XIX, centuria en la que se observa nuevamente un ligero aumento en el número de lemas. Tres cuestiones, por tanto, destacan en este gráfico: el claro descenso de arabismos durante la segunda mitad del siglo XVI, su prolongado mantenimiento durante buena parte del siglo XVII y todo el XVIII, y su ligero aumento a partir del siglo XIX.

La primera cuestión resulta esperable a tenor de la bibliografía especializada, pues son muchas las investigaciones que han señalado que el uso de este tipo de léxico experimentó un progresivo proceso de pérdida a lo largo del siglo XVI (Walsh 1967, Giménez-Eguíbar 2011). Sin embargo, este hecho es matizable, pues durante el siglo XVI la presencia de los moriscos en los territorios del antiguo reino granadino y en zonas del levante propició el mantenimiento e, incluso, la incorporación de arabismos:

> The confrontation of Morisco and Christian elements after the Capitulation results in the transfer of a sizeable number of arabisms in the first half of the sixteenth century. But as privileges are withdrawn throughout the sixteenth century, the disappearance of numerous terms descriptive of religious practices, secular officials in the Morisco social structure and decorative articles and fabrics which were products of the Morisco artisans is imminent (Walsh 1967: 352-353).

Esto explica que en ODE se documenten arabismos vinculados a la moda (*almaizar*, *almalafa*, *zaragüelles*), a la orfebrería (*ajorca*, *alhaite*) y a otro tipo de utensilios domésticos y objetos decorativos (*almofrej*, *alcatifa*, *jaez*) en el corte cronológico que abarca desde el año 1551 hasta 1575 y que, a su vez, estas voces ya no se documenten más en el corpus a partir de la primera mitad del XVII, momento en el que, como se ha anotado, los moriscos fueron finalmente expulsados de la Península. La influencia arábiga no dejó de existir a lo largo de la Edad Moderna, pero sí se vio notablemente empobrecida y arrinconada, en lo lingüístico, a ámbitos de lo cotidiano (Maíllo Salgado 1997).

Teniendo en cuenta esta última idea, la segunda cuestión, relativa al mantenimiento de frecuencias normalizadas de arabismos hasta el siglo XIX, encuentra explicación en el hecho de que muchos de estos arabismos pertenecen

a ámbitos designativos muy cotidianos e, incluso, primarios, como el caso de los colores o de ciertos objetos domésticos. Entre 1626 y 1800 ODE registra 216 lemas diferentes de arabismos. En su mayoría, estos pertenecen a campos léxico-semánticos relacionados con menaje y enseres del hogar (*anafe*, *alacena*, *candil*, *acetre*, *almirez*, *alfombra*, *taza*, *alcuza*, *batea*), telas y vestimenta (*cenefa*, *gasa*, *algodón*, *badana*, *guadamecí*, *tabí*), joyas (*marfil*, *tumbaga*, *aljófar*, *nácar*, *ámbar*, *arracada*), productos culinarios (*azafrán*, *aceite*, *café*, *azúcar*), productos químicos (*latón*, *azófar*), ciertas medidas de capacidad (*arroba*, *fanega*, *adarme*, *celemín*, *tomín*, *marjal*, *azumbre*), oficios (*alguacil*, *albañil*, *alcalde*, *alarife*, *albacea*, *alférez*) y algunas plantas (*anea*, *jazmín*), entre otros. Lo común y cotidiano ha provocado el mantenimiento e integración plena en español de muchos arabismos; otros han experimentado bien un arrinconamiento dialectal como el señalado anteriormente, bien un proceso de competencia y sustitución léxicas.

Sobre el tercer aspecto del gráfico aportado, relacionado con el ligero aumento de lemas correspondientes a léxico de origen árabe, es posible proporcionar algunos datos. Por ejemplo, en ODE observamos arabismos a partir de 1800 que no aparecían desde el siglo XVI en el corpus (*albéitar*, *gabán*, *loco, almazara*). A ellos se suman otros que aparecen por primera vez en ODE, como *álcali*, *canana*, *jaqueca, chivo, sandía* o *toronjil*. La mayoría no son arabismos nuevos en el idioma, dado que el CDH los documenta desde la Edad Media, pero sí muestran grandes lapsos temporales en su documentación desde la época medieval hasta los siglos XVIII y XIX —según indican las estadísticas del corpus académico—, como sucede en el caso de *sandía* o *chivo*. Cabe destacar que otras voces de nueva incorporación en el XIX en ODE no cuentan con ejemplos en el CDH, como *zalona*, la cual está marcada específicamente como andalucismo en el DLE.

Todos estos datos invitan a pensar que el ligero aumento de arabismos observado a partir del siglo XIX en ODE puede estar influido por el horizonte ideológico que se despliega a consecuencia de una de las corrientes culturales del Romanticismo: el orientalismo. Este movimiento puede ser considerado como un fenómeno intelectual y artístico en la Europa decimonónica como resultado del creciente interés por el estudio de las civilizaciones orientales en general, y de las árabes, en particular, en el contexto español (López García 2016). De la mano de esta corriente cultural, comenzó la trayectoria de la novela histórica a principios de dicha centuria, como consecuencia de ciertos acontecimientos políticos y sociales tales como la caída del imperio napoleó-

nico (Mata Induráin 1995: 21). La característica que une a todas estas novelas es su ambientación mayoritaria en escenarios del pasado, más o menos remoto (medievales, de las épocas clásicas griegas y romanas, del antiguo Egipto, del lejano oriente, etcétera). Debido a esto, no debería resultar extraña la recuperación de cierto léxico de origen árabe para ese desarrollo literario específico. No obstante, será necesario contrastar esta interpretación con más datos del siglo XIX, puesto que es el siglo más infrarrepresentado en ODE, siendo, por tanto, imprescindible su ampliación documental.

Por último, con relación al dendrograma presentado en la figura 6, este permite establecer cuatro clústeres bien definidos, que podrían interpretarse como cuatro etapas distintas respecto de la incorporación y frecuencia de uso de arabismos durante el español clásico y moderno. La primera etapa abarcaría desde 1500 hasta 1575 y se caracterizaría por una creciente incorporación de arabismos al español, entre las que se incluyen, como ya hemos señalado, tanto lemas que no volverán a documentarse en el corpus (*alhaite*, *almofrej*) como lemas que solo volverán a aparecer a partir del siglo XIX (*albéitar*, *almazara*). La segunda etapa se correspondería con el tramo de 25 años que va de 1576 a 1600; este corto periodo, que en el dendrograma aparece aislado del resto de tramos posteriores, marcaría el inicio de un descenso en el uso de arabismos, concordando así con la fase de pérdida de arabismos comúnmente señalada en la bibliografía. La tercera etapa, la más dilatada en el tiempo, comprendería dos siglos y medio, desde 1601 hasta 1850. No obstante, el dendrograma muestra aquí a su vez dos agrupamientos jerárquicos claramente contrastables: el que comprende el siglo XVII (1601-1700), de mayor irregularidad debido a que se prolonga el descenso de arabismos iniciado en el tramo anterior, y el que agrupa tanto el siglo XVIII como la primera mitad del XIX (1701-1850), que arroja un panorama bastante más estable. El cuarto y último clúster, correspondiente a los últimos 50 años del siglo XIX (1851-1900), se caracterizaría por un nuevo ascenso que, según revela el gráfico de líneas, se habría iniciado ya a partir de 1800. Este aumento en el número de lemas de origen árabe podría responder, como se ha apuntado, a razones histórico-culturales derivadas de la corriente orientalista decimonónica que pudieron tener un correlato en la vida cotidiana.

El dendrograma, por tanto, viene a plasmar y corroborar mediante agrupamientos jerárquicos los diferentes periodos que el gráfico de líneas ya dejaba intuir y sobre los que ya se han realizado las observaciones oportunas. Por su extensión y complejidad, no obstante, el tercero de los clústeres mencionados

puede ser matizado con algunos comentarios adicionales. Se han contrastado los arabismos que se registran en el siglo XVII y desaparecen en el XVIII, y viceversa. En el primer caso, para estos dos siglos se codifican como exclusivos del XVII bastantes arabismos referidos a plantas y arbustos (*acebuche*, *alcaparra*, *alcaravea*, *alquitira*, *cubeba*, *galanga*, *turbit*, *zaragatona* y *zumaque*). Este ámbito designativo se erige como especialmente significativo en el mantenimiento de arabismos debido a "la influencia de la lengua árabe en algunas materias de base alquímica y de medicina tradicional" (Salicio Bravo 2018: 202). En muchas ocasiones, dentro de este campo, términos de origen latino y griego han desplazado a los arabismos que ya existían para designar muchas plantas, ocasionando parejas o dobletes léxicos que establecen una competencia entre ellos en el uso (*alquitira* vs. *astrágalo*) (González Sopeña 2021: 3). El resultado de esa competencia tiene dos posibilidades que se traducen en que o bien el término de origen latino ha sustituido definitivamente al préstamo árabe, o bien el arabismo ha quedado distribuido diatópica o diastráticamente de forma residual en algunas zonas rurales. Otros arabismos exclusivos del XVII son *almofía*, *alfanje* o *alioj*. Estos tres ejemplos también cuentan con otras palabras que han venido a sustituirlos: en el caso de *almofía*, se registra competencia léxica con otro arabismo (*aljofaina*), *alfanje* lo hace con *sable* y, *alioj*, con *mármol*.

En el segundo caso, arabismos que tienen presencia en el XVIII y no en el XVII son, por ejemplo, *alcancía*, *alfarjía*, *almazara*, *batea*, *dula*, *almijarra* o *alcayata*. Para algunos de esos casos ya se ha anotado que coinciden en tener un alcance geográfico más notable en Andalucía, en general (*alfarjía*) y, en el oriente andaluz, de forma particular (*almazara, almocafre*), según los datos del ALEA. Esto nos lleva a pensar que el siglo XVIII resulta determinante para la configuración dialectal en el plano léxico del oriente andaluz. Asimismo, en este segundo corte aparecen como exclusivos ciertos arabismos referidos a productos alimentarios, como *arrope*, *arroz*, *almíbar*, *alifa* o *acemite*. *Alifa* está marcado como andalucismo en el DLE, si bien otros, por ser productos alimenticios básicos para los seres humanos, son generales (*arroz*, *almíbar*). Finalmente, este siglo también destaca en la presencia de arabismos vinculados al ámbito de los nombres de plantas (*alazor*, *arrayán*, *azucena*, *almez*, *algazul*, *anea*) y de ciertos minerales (*albayalde*, *azarcón*, *almáciga*), por las razones ya señaladas. Estos también establecen algunos dobletes léxicos (*azarcón* vs. *minio*).

6. Conclusiones

A pesar de la abundante bibliografía existente sobre las voces de origen árabe en español, faltan investigaciones que adopten una aproximación cuantitativa para estudiar la evolución diacrónica de los arabismos en nuestra lengua. Esta carencia se debe sobre todo a la ausencia de corpus históricos debidamente anotados con información etimológica, lo que dificulta el análisis estadístico y la extracción de frecuencias de uso de términos árabes en diferentes periodos temporales. Conscientes de esta limitación, este trabajo se estructura en torno a un doble propósito: explicar la metodología implementada en el proceso de anotación de arabismos en un corpus histórico y realizar, fruto de dicha anotación, un análisis cuantitativo sobre la evolución de este tipo de léxico entre 1500 y 1900.

Con respecto al primer objetivo, el procedimiento descrito en este trabajo para la anotación de arabismos combina tanto técnicas de procesamiento automático como revisiones de carácter manual. Así, se ha partido de un lemario externo de arabismos —extraído del recurso IEDRA y, por tanto, basado en datos actualizados del DLE—, y se ha desarrollado un *script* para la anotación automática de los lemas del corpus contenidos en dicho lemario. El hecho de trabajar con un corpus previamente normalizado y lematizado ha agilizado este proceso; no obstante, este enfoque automatizado generó falsos positivos, debido principalmente a la presencia de voces homónimas (*zagal*) o de términos árabes adoptados del francés (*sofá*). Ante esta situación, se llevó a cabo una corrección manual de estos errores, para asegurar la precisión de la anotación y garantizar, en definitiva, la calidad de los datos analizados. Se ha partido de un lemario compuesto por 1 264 lemas y se han identificado en el corpus un total de 11 940 arabismos distribuidos en 309 lemas diferentes. Esta propuesta metodológica integra, asimismo, el uso de estándares consolidados tanto en el ámbito de las humanidades digitales (marcación en XML-TEI) como en el de la lingüística de corpus (sintaxis CQL), lo que facilita su adaptación para anotar otros corpus lingüísticos.

Con respecto al segundo objetivo, el análisis propuesto se ha basado fundamentalmente en el cálculo de lemas de origen árabe por millón de casos en tramos de 25 años. La comparación de estas frecuencias ha permitido generar tanto un gráfico de líneas como un dendrograma. Ambos revelan, directa o indirectamente, al menos cuatro periodos diferentes relativos a la intensidad en el uso de arabismos entre 1500 y 1900: un primer periodo de ascenso y

apogeo durante el siglo XVI, un segundo periodo de descenso a finales de ese mismo siglo, un tercer periodo en que se avanza hacia una situación de estabilidad durante el siglo XVII y, sobre todo, el siglo XVIII, y un cuarto y último periodo de cierto resurgimiento en la segunda mitad del siglo XIX. Los datos analizados se cimentan, además, en un corpus compuesto mayoritariamente por inventarios de bienes, un tipo textual especialmente fructífero en el uso de voces propias de la esfera de lo cotidiano. Este último aspecto ha permitido documentar arabismos pertenecientes a un amplio abanico de campos léxico-semánticos (ropa, menaje, joyería, cocina, etcétera) e incluso a tener constancia de algunos que apenas sí aparecen en los corpus históricos de referencia (*zalona*, *alioj*).

Esperamos, en definitiva, que este trabajo proporcione una base sólida para comprender mejor la influencia y la evolución del léxico árabe en el español de la Edad Moderna.

Corpus / Referencias bibliográficas

Alcalá, Fray Pedro (1505): *Vocabulista arábigo en letra castellana*. Granada: Juan Varela.

Allés-Torrent, Susanna (2015): "Edición digital y algunas tecnologías aliadas", *Ínsula. Revista de Letras y Ciencias Humanas*, 822, pp. 18-21.

Álvarez de Miranda, Pedro (2009): "Neología y pérdida léxica", en Elena de Miguel (coord.), *Panorama de la lexicología*. Barcelona: Ariel, pp. 133-156.

Arié, Rachel (1993): *España musulmana (siglos VIII-XV)*. Barcelona: Labor.

Betz, Werner (1949): *Deutsch und Lateinisch: die Lehnbildungen der althochdeutschen Benediktinerregel*. Bonn: Bouvier.

Bloomfield, Leonard (1933): *Lenguaje*. Lima: Universidad Nacional Mayor de San Marcos.

Brezina, Vaclav (2018): *Statistics in Corpus Linguistics. A Practical Guide*. Cambridge: Cambridge University Press.

Burnard, Lou (2014): *What is the Text Encoding Initiative? How to add intelligent markup to digital resources*. OpenEdition Press. DOI: <https://doi.org/10.4000/books.oep.426>.

Calderón Campos, Miguel (2010): "Aspectos de la vida social granadina a través de diez arabismos de las actas del ayuntamiento y de las ordenanzas municipales", *Études romanes de Brno*, 2, pp. 179-192.

— (2015): *El reino de Granada en sus documentos (1492-1833)*. Bern: Peter Lang.

— (2019): "La edición de corpus históricos en la plataforma TEITOK: el caso de *Oralia Diacrónica del Español* (ODE)", *Chimera*, 6, pp. 21-36.

Calderón Campos, Miguel y García Godoy, María Teresa (2023): "ALEA-XVIII. Un corpus lingüístico para cartografiar la Andalucía del Setecientos", *Études romanes de Brno*, 44(2), pp. 153-175.

Calderón Campos, Miguel y Vaamonde, Gael (2020): "*Oralia Diacrónica del Español*: un nuevo corpus de la Edad Moderna", *Scriptum Digital*, 9, pp. 167-189.
CdE = El Corpus del Español. Davies, Mark (2016). <http://www.corpusdelespañol.ogr>.
CdH = Real Academia Española: *Corpus del Nuevo Diccionario Histórico del Español*. <www.rae.es>.
Christ, Oliver, Schulze, Bruno 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.pdf>.
Colón Domènech, Germán (2002): "De arabismos interhispanos", en Albert Soler y Nuria Mañé (eds.), *Para la historia del léxico español*. Madrid: Arco/Libros, pp. 45-54.
Corde = Real Academia Española: Corpus Diacrónico del Español. <www.rae.es>.
Corominas, Joan y Pascual, José Antonio (1980-1991): *Diccionario crítico etimológico castellano e hispánico* (DCECH). Madrid: Gredos.
Corriente, Federico (1977): *A Grammatical Sketch of the Spanish-Arabic Dialect Bundle*. Madrid: Instituto Hispano-Árabe de Cultura.
— (1999): *Diccionario de arabismos y voces afines en iberorromance*. Madrid: Gredos.
— (2004): "El elemento árabe en la historia lingüística peninsular", en Rafael Cano Aguilar (coord.), *Historia de la lengua española*. Barcelona: Ariel, pp. 185-205.
— (2008): *Dictionary of Arabic and Allied Loanwords*. Leiden: Brill.
Del Rio Riande, Gimena y Allés-Torrent, Susanna (2023): "¿Quién conforma la comunidad de la TEI en español? Análisis de los datos de una encuesta", *Journal of the Text Encoding Initiative* [*online*], 16, DOI: <https://doi.org/10.4000/jtei.4927>.
Dle = Real Academia Española: *Diccionario de la lengua española*, 23.ª ed., [versión 23.7 en línea]. <https://dle.rae.es>.
Dozy, Reinhart Pieter Anne y Engelmann, Willem Herman (1869): *Glossaire des mots espagnols et portugais dérivés de l'árabe*. Leiden: Brill.
Dworkin, Steven (2012): *A History of the Spanish Lexicon. A Linguistic Perspective*. Oxford: Oxford University Press.
Eguílaz y Yanguas, Leopoldo (1886 [1974]): *Glosario etimológico de las palabras españolas de origen oriental*. Granada: La Lealtad.
Fradejas Rueda, José Manuel (2010): "La codificación XML/TEI de textos medievales", *Memorabilia*, 12, pp. 219-247.
García González, Javier (1998): "Clases de arabismos en los textos alfonsíes", en Claudio García Turza *et al.* (eds.), *Actas del IV Congreso Internacional de Historia de la Lengua Española*, vol. 2. Logroño: Universidad de La Rioja, pp. 127-136.
— (2007): "Una perspectiva sociolingüística de los arabismos en el español de la alta Edad Media (711-1300)", en Inmaculada Delgado Cobos y Alicia Puigvert Ocal (eds.), *Ex admiratione et amicitia: homenaje a Ramón Santiago*, vol. 1. Madrid: Ediciones del Orto, pp. 523-548.
— (2008): "Viejos problemas desde nuevos enfoques: los arabismos en el español medieval desde la perspectiva sociolingüística", en José Luis Blas Arroyo *et al*. (coords.), *Discurso y sociedad II: nuevas contribuciones al estudio de la lengua en contexto social*. Castellón de la Plana: Universitat Jaume I, pp. 671-684.

Garulo Muñoz, Teresa (1983): *Los arabismos en el léxico andaluz (según los datos del Atlas lingüístico y etnográfico de Andalucía)*. Madrid: Instituto Hispano-Árabe de Cultura.

Giménez Eguíbar, Patricia (2011): "Algunas cuestiones respecto a la pérdida de arabismos en español peninsular", *Romance Philology*, 64, pp. 185-195.

— (2015): "Dos casos de sustituciones léxicas: los arabismos alfayate y alfajeme", en Francisco Javier de Cos Ruiz y Mariano Franco Figueroa (coords.), *Actas del IX Congreso Internacional de Historia de la Lengua Española*. Madrid/Frankfurt am Main: Iberoamericana/Vervuert, pp. 1413-1427.

— (2016): "Attitudes toward Lexical Arabisms in 16th Century Spanish Texts", en Sandro Sessarego y Fernando Tejero-Herrero (eds.), *Spanish Language and Sociolinguistics Analysis*. Amsterdam/Philadelphia: John Benjamins, pp. 363-380.

— (2024): "La contribución del árabe al hispanorromance", en Steven Dworkin *et al.* (eds.), *Lingüística histórica del español*. London: Routledge, pp. 362-371.

Gómez Capuz, Juan (2004): *Los préstamos del español*. Madrid: Arco/Libros.

González Sopeña, Inmaculada (2017): "Arabismos y fiscalidad", *Dicenda*, 35, pp. 109-130.

— (2019): *Los arabismos en el reino de Granada a través de la documentación archivística (finales del siglo xv-siglo xvii)*. Granada: Universidad de Granada.

— (2021): "Arabismos en el campo léxico de los nombres de las plantas y minerales a través de la documentación administrativa del reino de Granada (siglos xvi y xvii)", *Tonos digital*, 40, pp. 1-24.

— (2021): "Sustituciones léxicas en los arabismos del reino de Granada", en Christopher Pountain *et al.* (ed.), *New Worlds for Old Words*. Delaware: Vernon Press, pp. 157-169.

Gries, Stephan Th. y Hilpert, Martin (2008): "The Identification of Stages in Diachronic Data: Variability-Based Neighbour Clustering", *Corpora*, 3(1), pp. 59-81.

Guadix, Diego de (1593 [2005]): *Recopilación de algunos nombres arábigos que los árabes pusieron a algunas ciudades y a otras muchas cosas*, en Elena Bajo Pérez y Felipe Maíllo Salgado (eds.). Gijón: Trea.

Haugen, Einar (1950): "The Analysis of Linguistic Borrowing", *Language*, 26, pp. 210-231.

Isasi Martínez, Carmen (coord.); Lobo Puga, Ana; Martín Aizpuru, Leyre; Pérez Isasi, Santiago; Pierazzo, Elena y Spence, Paul (coord.) (2014): *Guía para editar textos CHARTA según el estándar TEI: una propuesta*. <https://www.redcharta.es/investigacion/>.

Janssen, Maarten (2012): "Neotag: A POS Tagger for Grammatical Neologism Detection", *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Estambul, Turquía, pp. 1-7.

— (2016): "TEITOK: Text-Faithful Annotated Corpora", *Proceedings of the Language Resources and Evaluation Conference*. Portoroz, Eslovenia, pp. 4037-4043.

Kytö, Merja (2011): "Corpora and Historical Linguistics", *Revista Brasileira de Linguística Aplicada*, 11(2), pp. 417-457, DOI: <https://doi.org/10.1590/S1984-63982011000200007>.

Lapesa, Rafael (1981 [1942]): *Historia de la lengua española* (9.ª edición). Madrid: Gredos.

Leech, Geoffrey y Wilson, Andrew (1996): *EAGLES. Recommendations for the Morphosyntactic Annotation of Corpora*. <http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html>.
López Tamarid, Francisco (1585 [1737]): *Compendio de algunos vocablos arábigos*. Madrid: Juan de Zúñiga.
Lörinczi, Marinella (1969): "Consideraciones semánticas acerca de las palabras españolas de origen árabe", *Revue Roumaine de Linguistique*, 14, pp. 65-75.
Maíllo Salgado, Felipe (1997): *La huella arábiga en el idioma español*. Salamanca: Universidad de Salamanca.
— (1998). *Los arabismos del castellano en la Baja Edad Media*. Salamanca: Universidad de Salamanca.
Mata Induráin, Carlos (1995): "Retrospectiva sobre la evolución de la novela histórica", en Kurt Spang, Ignacio Arellano y Carlos Mata (eds.), *La novela histórica. Teoría y comentarios*, *Anejos de RILCE*, 15, pp. 13-64.
Morala Rodríguez, José Ramón (2012a): "Arabismos en textos del siglo xvii escasamente documentados", *Revista de Investigación Lingüística*, 15, pp. 77-102.
— (2012b): "Léxico e inventarios de bienes en los Siglos de Oro", en Gloria Clavería Nadal, Margarita Freixas, Marta Prat Sabater, Joan Torruella Casañas (coords.), *Historia del léxico: perspectivas de investigación*. Madrid/Frankfurt am Main: Iberoamericana/Vervuert, pp. 199-218.
Neuvonen, Ereo Kalervo (1941): *Los arabismos del español en el siglo xiii*. Helsinki: Sociedad literaria finesa.
Ode = Oralia Diacrónica del Español. Calderón Campos, Miguel y García-Godoy, María Teresa (2010-2019): En línea <http://corpora.ugr.es/ode>.
Oliver Pérez, Dolores (2004): "Los arabismos dentro de la historia del español: estudio diacrónico de su incorporación", en Manuel Cecilio Díaz *et al.*, *Estudios dedicados a José María Fernández Catón*, vol. 2. León: Centro de Estudios e Investigación San Isidoro, pp. 1073-1095.
— (2005-2006): "Los arabismos del *Libro Conplido* y otras huellas árabes", *Anuario de Lingüística Hispánica*, 21/22, pp. 67-118.
Pocklington, Robert (1984): "Nuevos arabismos en los textos alfonsíes murcianos", *Miscelánea Medieval Murciana*, 11, pp. 261-295.
Puche Lorenzo, Miguel Ángel (2012): "Léxico de la vida cotidiana en la Murcia áurea", *Cuadernos del Instituto de Historia de la Lengua*, 7, pp. 343-360.
Rodríguez Alberich, Gabriel (2020): *IEDRA*. <https://iedra.es/>.
Salicio Bravo, Soraya (2018): *Los materiales de las técnicas preindustriales en el renacimiento hispano*. Salamanca: Universidad de Salamanca.
Serrano Niza, Dolores (2007): "Arabismos relacionados con el léxico de la seda", *Revista de Filología de la Universidad de La Laguna*, pp. 559-566.
Steiger, Arnold (1932): *Contribución a la fonética del hispanoárabe y los arabismos en el iberorrománico y siciliano*. Madrid: Centro Superior de Investigaciones Científicas.
Torres Martínez, Marta (2014): "Notas sobre el léxico documentado en cartas de dote almerienses del siglo xviii", en José Ramírez Luengo (coord.), *Historia del español hoy: estudios y perspectivas*. Lugo: Axac, pp. 217-256.
Torres Montes, Francisco (1996): "Nombres de medidas agrarias en la costa del antiguo Reino de Granada", en Juan de Dios Luque Durán y Antonio Pamies Ber-

trán (eds.), *Segundas Jornadas sobre estudio y enseñanza del léxico*. Granada: Método, pp. 265-282.

Vaamonde, Gael (2024): "Diseño y explotación de un corpus histórico de textos oralizantes para el estudio del español clásico y moderno", *Revista de Humanidades Digitales*, 9, pp. 41-70.

Walsh, John (1967): *The Loss of Arabisms in the Spanish Lexicon* (tesis doctoral inédita). Virginia: University of Virginia.

Weinreich, Uriel (1953): *Languages in Contact: Findings and Problems*. Den Haag: Mouton.