La edición de corpus históricos en la plataforma TEITOK.
El caso de *Oralia diacrónica del español (ODE)*

Miguel Calderón Campos
Universidad de Granada

This article presents the *Oralia diacrónica del español* (ODE) (*Spanish diachronic oral discourse*) corpus - a continuation of the *Corpus diacrónico del español del reino de Granada, 1492-1833* (CORDEREGRA) (*Diachronic Corpus of Spanish in the Kingdom of Granada),* primarily comprised of declarations by witnesses and inventories of goods. One new feature of the ODE is the transcription of the manuscripts into XML, according to the character coding standard proposed by the TEI consortium, and closely following the model that was successfully carried out in the *P.S. Post Scriptum: A Digital Archive of Ordinary Writing (Early Modern Portugal and Spain*) project. The TEITOK (Janssen 2016) tool, designed for creating, maintaining and publishing linguistically annotated corpora and to provide them with a robust search engine, is being used for the tokenization, standardization and labelling of the corpus.

Keywords: Corpus Linguistics, Diachronic Linguistics, Spanish, XML, TEITOK

1. Introducción

La investigación en lingüística histórica no se concibe en la actualidad sin el recurso a los corpus digitales en línea. Para el español, contamos con grandes corpus de iniciativa académica, como el CORDE y el *Corpus del nuevo diccionario histórico del español*, o extraacadémica, como el subcorpus histórico del *Corpus del español*. Además, en los últimos años, el esfuerzo se está concentrando en la confección de corpus especializados, geográfica, temática o tipológicamente[1].

[1] Una información detallada de los corpus histórico-lingüísticos disponibles puede verse en el *Portal de corpus históricos iberorrománicos* (Torruella & Kabatek 2018); véase también Torruella 2017.

© 2019 The Author(s); distributed under the Creative Commons Attribution License

Entre 2010 y 2014 se elaboró en la Universidad de Granada el *Corpus diacrónico del español del reino de Granada, 1492-1833, CORDEREGRA*, compuesto por documentación manuscrita inédita de las actuales provincias de Granada, Málaga y Almería. Incluye principalmente dos tipos textuales, de tendencia oralizante: declaraciones de testigos e inventarios de bienes (Calderón Campos, 2015).

La experiencia de estos años permitió comprobar algunas carencias y limitaciones técnicas del modelo que se estaba siguiendo, la mayoría de ellas relacionadas con el hecho de haber realizado las transcripciones con un procesador de textos convencional, tipo *Microsoft Word*. Esta decisión de partida dificultó enormemente dos de los objetivos iniciales que se habían planteado:

a. Proporcionar a los usuarios un corpus en línea que permitiera realizar búsquedas complejas.
b. Ofrecer dos ediciones de cada documento, una paleográfica y otra normalizada, la primera destinada a los historiadores de la lengua; la segunda orientada a un público más heterogéneo.

Además, siguiendo el modelo de la red CHARTA (Sánchez-Prieto, 2011), era deseable poder mostrar el facsímil del manuscrito original, con el fin de que los usuarios pudieran contrastar la validez de las transcripciones.

La publicación en 2014 del corpus de cartas privadas *P.S. Post Scriptum: Archivo Digital de Escritura Cotidiana en Portugal y España en la Edad Moderna* ofreció un modelo muy avanzado sobre el que fundamentar los cambios necesarios para convertir el antiguo CORDEREGRA en una útil herramienta de investigación lingüística. El corpus epistolar utilizó como soporte informático la plataforma TEITOK (Janssen 2016)[2], diseñada para crear, mantener y publicar en línea corpus anotados lingüísticamente. En las líneas que siguen expondremos los principales cambios metodológicos requeridos para la implantación de TEITOK. El resultado es un nuevo corpus, *Oralia diacrónica del español* (ODE)[3], en el que se están adaptando las transcripciones antiguas al modelo de *Post Scriptum* y añadiendo nueva documentación, no circunscrita necesariamente al reino granadino. Estos nuevos documentos pertenecen a la misma tipología del CORDEREGRA, esto es, declaraciones de testigos e inventarios, pero proceden de regiones del centro-norte de la Península, con objeto de disponer de un corpus de control con el que comparar la documentación granadina. Está previsto elaborar un corpus de

[2] http://teitok.corpuswiki.org/
[3] http://corpora.ugr.es/ode

medio millón de palabras, el 65% del antiguo reino de Granada, el resto de modalidades norteñas[4].

2. Transcripción en XML

Siguiendo el ejemplo de *Post Scriptum*, los documentos del nuevo corpus están siendo transcritos en XML (Extensible Markup Language), según las directrices del consorcio TEI (Text Encoding Initiative). Este modelo de edición está ampliamente generalizado en el mundo de las humanidades digitales, aunque es muy poco frecuente todavía en el de la compilación de corpus históricos (Marttila 2014: 203, Díaz Bravo 2015: 389, Martín Aizpuru 2016: 152)[5].

La principal razón de este rechazo a XML entre los filólogos parece estar en el convencimiento, veremos que equivocado, de que etiquetar los documentos presenta demasiados inconvenientes: ralentiza el proceso de transcripción, ya de por sí engorroso, y además plantea la dificultad añadida de tener que buscar ayuda informática adicional para que la documentación se visualice sin las etiquetas. Efectivamente, los documentos transcritos con procesadores de texto convencionales se formatean directamente para su impresión o lectura. Por el contrario, los documentos etiquetados en XML presentan un aspecto ilegible en el que, entre el contenido textual, se enredan códigos informáticos que deben desaparecer en la visualización final.

El siguiente fragmento, extraído de una declaración de testigos de ODE, sirve para ejemplificar lo que venimos diciendo:

> <lb/>que se alla que a|=ha bisto y curado a Manuel Martin, hijo de <lb/>Fern<add place="above">do</add>||Fernando Martin que esta|=está en el hospital de s<add place="above">or</add>||señor s<add place="above">n</add>san Ju||Juan de Dios de

[4] En la fase de pruebas actual (25/10/2019), el corpus ODE, ya disponible en la red (http://corpora.ugr.es/ode), cuenta con un total de 109 147 palabras, distribuidas de la siguiente forma: a) Almería: 23 inventarios del siglo XVIII (26 643 palabras); 28 inventarios del XIX (24 222 palabras); b) Madrid: 21 inventarios, s. XVIII (27 543 palabras); 21 inventarios, s. XIX (30 739 palabras). Todas estas transcripciones se han hecho directamente en XML y no formaban parte del antiguo CORDEREGRA. Lo que queda por realizar, hasta llegar al medio millón de palabras, es adaptar la antigua documentación de CORDEREGRA al nuevo estándar XML-TEI, para posteriormente incorporarla a ODE-TEITOK. Para ello contamos con un *script* de conversión elaborado por Gael Vaamonde.

[5] Un análisis de los siete principales corpus de base documental del mismo periodo que ODE permite comprobar que solo *Post Scriptum* utiliza el estándar XML-TEI. Se han analizado CHARTA, CODEA +2015, *CorLexIn*, *Cíbola Project*, CORDIAM, COREECOM y *Post Scriptum*.

<lb/>donde es zirujano de una herida en la p<add place="above">te</add>||parte alta de la <lb/>cabeza sobre la comisura sajittal|=sagital, la qual al parezer fue <lb/>hecha con ynstrum<add place="above">to</add>||ynstrumento contundente como palo, piedra o seme<lb/>jante, la qual rompio cuero, gordura, menbrana car<lb/>nosa hasta aberle hecho vna subitraz<add place="above">on</add>||subitrazion|=subintración tota|=total en la misma <lb/>comisura, la qual es peligrosa así p||por su esenzia como la prin<lb/>zipalidad de la p<add place="above">te</add>||parte y por los aszidentes|=accidentes que pueden sobre<lb/>benir tiene peligro de la bida.

Para que este fragmento transcrito en XML se pueda leer sin etiquetas, en ODE se utiliza la herramienta TEITOK, gracias a la cual el texto se visualiza de la siguiente forma:

que se alla que a bisto y curado a Manuel Martin, hijo de
Fern^{do} Martin que esta en el hospital de s^{or} s^{n}san Ju de Dios de
donde es zirujano de una herida en la p^{te} alta de la
cabeza sobre la comisura sajittal, la qual al parezer fue
hecha con ynstrum^{to} contundente como palo, piedra o seme
jante, la qual rompio cuero, gordura, menbrana car
nosa hasta aberle hecho vna subitraz^{on} tota en la misma
comisura, la qual es peligrosa así p su esenzia como la prin
zipalidad de la p^{te} y por los aszidentes que pueden sobre
benir tiene peligro de la bida. Y que esto es la berdad so cargo

Figura 1. Presentación de la transcripción paleográfica en ODE

Obviamente, la visualización del documento no es la principal ventaja de usar XML, sino la posibilidad de vincular el etiquetado con un motor de búsquedas muy avanzado, del que se tratará en el apartado 6.

A pesar del "terror" visual que produce la transcripción en XML, la realidad es mucho más sencilla de como parecen ser las cosas al principio. En primer lugar, el número de etiquetas empleadas en el proceso de transcripción se reduce, en el caso de ODE, a una decena, como puede verse en la Tabla 1.

En segundo lugar, manejar estas etiquetas (y otras que puedan añadirse, según las necesidades de cada proyecto) no es más complicado que usar marcas personales, con la ventaja de que se emplea un sistema de anotación estándar compartido por otros muchos equipos de investigación y que tiene, sobre todo, efectos muy destacados sobre las posibilidades de recuperación de información del corpus.

Por ejemplo, en los manuscritos es frecuente encontrar tachaduras legibles, como la que se observa en la figura 2:

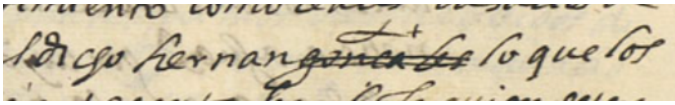

Figura 2. Tachadura legible

A la hora de señalar en la edición que la palabra *gonçales* está tachada en el original, TEI recomienda poner la etiqueta ~~gonçales~~, procedimiento que se sigue en *Post Scriptum* y que se ha adoptado en ODE.

Otros corpus históricos emplean etiquetas particulares con esta misma función. Por ejemplo, en CHARTA se escribe el texto entre corchetes y detrás de la palabra *tachado* en cursiva: [*tachado*: gonçales]; en el proyecto Cíbola se emplea un paréntesis seguido de acento circunflejo: (^gonçales). Los criterios propios, válidos y comprensibles para los lectores, tienen el inconveniente de no ser interpretables informáticamente. Por el contrario, la etiqueta XML-TEI <del> es unívoca y procesable computacionalmente. Además, no es más costoso en términos de tiempo de trabajo elegir una etiqueta XML que una marca particular.

Tabla 1. Etiquetas más frecuentes

Etiqueta XML	Significado
	Párrafo
<pb/>	Inicio de página
<lb/>	Inicio de línea
|| o <ex></ex>	Expansión de abreviatura
<sic></sic>	Así aparece en el original
<add></add>	Adición del escribano
<supplied></supplied>	Conjetura verosímil del editor
<unclear></unclear>	Conjetura dudosa del editor
<gap/>	Segmento sin transcribir
<del></del>	Tachado en el manuscrito

3. Conceptos básicos del etiquetado en XML

Antes de continuar es necesario hacer algunas aclaraciones sobre cuatro conceptos clave del lenguaje XML: etiqueta, elemento, atributo y valor:

<etiqueta nombre="valor">contenido</etiqueta>

El más sencillo de todos es el de *etiqueta*, nombre que recibe cualquier marca delimitada entre paréntesis angulares. En el ejemplo ilustrado en la figura 2 tenemos una etiqueta de apertura <del> y otra de cierre </del>. El contenido es todo aquello que se escribe entre la etiqueta inicial y la final, *gonçales* en el ejemplo.

En algunos casos puede ser relevante añadir alguna información complementaria, de tipo paratextual o codicológico. Por ejemplo, <add> se emplea para indicar que el texto que se transcribe está añadido, bien al margen, bien por encima de la línea. Para especificar que se trata de un interlineado ("mui" en la figura 3) y no un añadido al margen, es posible completar la información genérica que aporta <add> añadiendo dentro de la etiqueta de apertura la indicación de que el texto está escrito por encima de la línea:

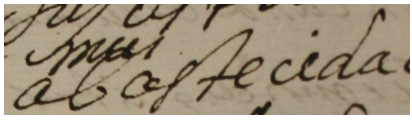

Figura 3. Interlineado

Esta información extra se incluye como *valor* ("above") del atributo *place*. La suma de etiqueta (inicial y de cierre), atributo, valor y contenido recibe el nombre de *elemento*.

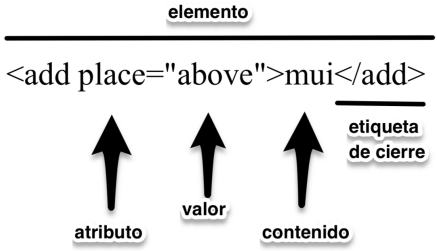

Figura 4. Conceptos clave de XML

XML ofrece instrucciones de procesamiento informático que la plataforma TEI-TOK interpreta adecuadamente para que el documento de la figura 3 se visualice en ODE sin las etiquetas, como se muestra en la figura 5:

mui abastecida

Figura 5. Visualización del interlineado

3.1 Etiquetas para un corpus histórico

De las etiquetas de la tabla 1, cuatro de ellas merecen especial atención. Se trata de <supplied>, <unclear>, <gap> y <del>. Todas se usan, con pequeñas diferencias, cuando el manuscrito presenta algún problema de lectura, por deterioro o por dificultad paleográfica o caligráfica. La primera de ellas, <supplied>, sirve para ofrecer, con argumentos fiables, una propuesta de transcripción para un segmento ilegible del manuscrito.

Figura 6. <supplied>

En la figura 6 se puede hacer una conjetura verosímil ("de caoua herrado"), dado que anteriormente se habían inventariado dos "bufetes grandes de nogal herrado" (ODE, AL1721I0005). En este caso, en el texto en XML se escribe <supplied>herrado</supplied>, indicando precisamente que se está haciendo una conjetura con bastante grado de fiabilidad.

Otras veces el documento está escrito de manera que solo permite hacer una vaga interpretación, con bastantes dudas sobre la certeza de lo transcrito.

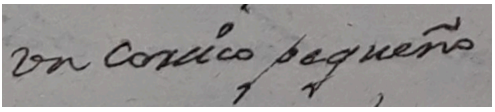

Figura 7. <unclear>

Es lo que ocurre en la figura 7. A pesar de que el manuscrito no está dañado, la caligrafía del escribano solo deja aventurar una lectura a la que no se otorga demasiada certidumbre: "vn corcico pequeño", lo que se etiqueta en XML como <unclear>corcico</unclear>.

Por último, puede ocurrir que el estado del manuscrito o la calidad de la fotografía nos impidan reconstruir un determinado fragmento, como ocurre en el

margen derecho del manuscrito de la figura 8, ilegible por la encuadernación del legajo.

En este caso, la etiqueta vacía </gap> indica la imposibilidad de ofrecer una lectura verosímil detrás de "Diez y nuebe libras".

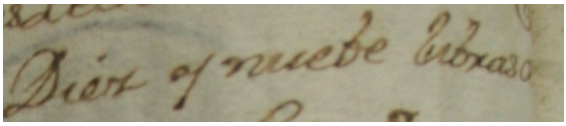

Figura 8. <gap>

Las restantes etiquetas de la tabla 1 son fáciles de interpretar: estructura el texto en párrafos, <lb/> ("line begining") señala el inicio de una línea[6] y <pb/> ("page begining") el de página[7].

Por medio de <sic></sic> se destaca el texto original aparentemente erróneo, para informar de que no se trata de un fallo de transcripción (<sic>Grabiel</sic>). Por último, en ODE son válidas dos estrategias para marcar las expansiones de abreviaturas: una estándar de TEI: <ex> y otra especial de TEITOK, la doble plica ||. De esta forma, la forma abreviada *dho* puede transcribirse como d<ex>ic</ex>ho, o bien como dho||dicho.

4. El problema de la fuente única. Tokenización y normalización

Uno de los objetivos de ODE es presentar dos ediciones de un mismo documento, una paleográfica y otra normalizada. En CORDEREGRA se había intentado resolver este problema editando y marcando por separado la transcripción paleográfica y la edición crítica. El procedimiento generaba dos documentos (es decir, dos archivos de *Microsoft Word*) independientes, de gestión también independiente, lo que complicaba la revisión de lo que en realidad eran dos corpus distintos, no uno solo con distintas ediciones.

Imaginemos, por ejemplo, que en la versión paleográfica se hubiera expandido por error la abreviatura *tpo* como *tipo*, y que en el proceso de revisión se descubre que debería haberse desarrollado como *tiempo*; este fallo obligaba a hacer una doble corrección, primero en el documento que contiene la transcripción paleográfica y luego en el que incluye la normalizada. Se trabajaba, como se ve,

[6] Admite el atributo "n" para numerar las líneas: <lb n="1"/>.

[7] Se pueden emplear los atributos *n* y *facs*, para indicar el número de folio y la fotografía con la que se vincula el documento: <pb n="1r" facs="foto.pjg"/>.

con un modelo de fuente doble muy poco operativo, especialmente a medida que el corpus iba alcanzado un tamaño considerable. Además, y este parece ser un problema aún mayor, los usuarios tenían que hacer búsquedas separadas: bien en el corpus paleográfico, bien en el corpus normalizado.

El reto es cambiar el modelo por uno de fuente única, es decir, uno en el que cada manuscrito se edite en un solo documento XML que contenga las marcas correspondientes de las distintas ediciones (Isasi et al., 2014).

La solución planteada en TEITOK parte del concepto de <tok>, una etiqueta creada por Janssen (2016) para resolver el problema de la fuente única y presentar distintas ediciones de un mismo documento de una manera eficiente y dinámica.

Cuando se terminan de transcribir los manuscritos comienza una nueva fase en la edición del corpus, la propiamente computacional. Individualmente, cada documento debe tokenizarse, es decir, dividirse en "tokens" o formas ortográficas, en la plataforma TEITOK. Esta operación se realiza automáticamente y genera en el XML un elemento <tok></tok> con la forma transcrita (*form*) como contenido. Para la palabra *vahil* de la figura 9, por ejemplo, se genera por defecto un *tok* con *vahil*[8] como contenido de la etiqueta: <tok id="w-970">vahil</tok>.

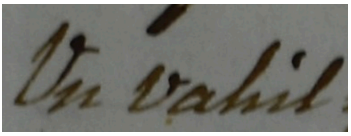

Figura 9. Forma transcrita (form) = vahil

El sistema numera cada token correlativamente (id="w-970") y proporciona una plantilla en la que los editores pueden añadir la forma normalizada correspondiente (figura 10). Cuando se rellena el recuadro de la forma normalizada se crea automáticamente en la etiqueta de apertura de XML un atributo *nform* con el valor de la variante estándar (*badil*):

<tok id="w-970" nform="badil">vahil</tok>

Esta etiqueta equivale a la siguiente orden de procesamiento: cuando se visualicen los documentos, en la edición paleográfica debe aparecer el contenido del

[8] Es decir, *badil* (instrumento para remover las brasas o recoger las cenizas), con pérdida de *d* intervocálica.

elemento tok[9], es decir, *vahil*, y en la edición normalizada el valor del atributo *nform* (*badil*).

En la figura 10 se ilustra la manera de normalizar manualmente el corpus. Esta operación se puede realizar de otras dos formas, en función del grado de complejidad léxica del documento: la primera consiste en escribir directamente en el procesador de XML las versiones paleográfica y normalizada separadas por una plica seguida del signo igual: vahil|=badil. Cuando se tokeniza el texto, TEITOK automáticamente interpreta que *badil* es el valor del atributo *nform* correspondiente a la forma paleográfica *vahil*; la segunda consiste en aplicar el normalizador automático, cuyo porcentaje de acierto depende del corpus de referencia que se tenga como base. Si la palabra *vahil* ya ha aparecido otras veces en el corpus, el normalizador automático la reconocerá; en caso contrario, es preferible hacer el proceso manualmente.

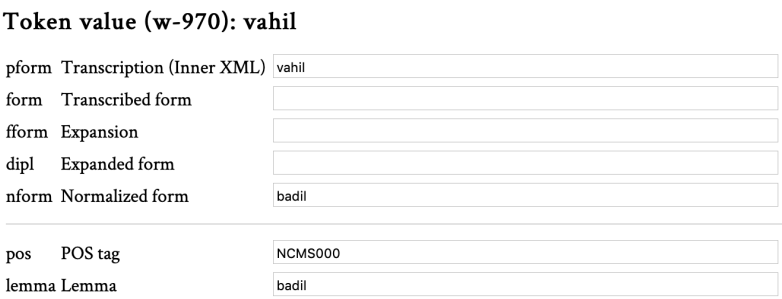

Figura 10. Niveles de edición del token

Obsérvese que este procedimiento puede ampliarse a cuantas ediciones sean necesarias: en el caso de ODE se incluyen dos, pero nada impediría realizar una edición intermedia, de carácter crítico[10] o añadir una traducción al inglés, por ejemplo. Bastaría con añadir nuevos atributos al lado de *nform*.

[9] Que TEITOK llama *form*.

[10] Por ejemplo, en *Post Scriptum* se emplea el atributo *dform* para incluir las variantes estandarizadas de arcaísmos o dialectalismos: *agora*, *naide*, *ansí*, etc. (Vaamonde & Magro 2017:19). Estas formas reducen ya una variación de partida (*hagora* / *agora*, *nayde* / *najde* / *naide*, *ansi* / *ansy* / *ansí*), a favor de un hipotético estándar de la época, anterior a la norma actual. La edición crítica de CHARTA, donde se simplifican los usos gráficos sin trascendencia fonética (por ejemplo, *uino* / *vino*, *soi* / *soy*, *iusto* / *justo*, *hedad* / *edad*, *qumplir* / *cumplir*, etc.) podría también incluirse como ejemplo de edición intermedia entre la paleográfica estrecha y la normalizada.

5. Lematización y etiquetado (POS)

Hasta ahora, la etiqueta *tok* facilita dos cosas: por un lado, la doble visualización de los documentos, en edición paleográfica o normalizada; y por otro, aumenta el rendimiento del buscador, puesto que el elemento *tok* incluye información que vincula todas las variantes ortográficas de una palabra con la misma forma normalizada (*nform*). Por ejemplo, las formas ortográficas (*form*) *vahil*, *vadil* o *badil* apuntan a la misma nform="badil":

<tok id="w-970" nform="badil">vahil</tok>

<tok id="w-55" nform="badil">vadil</tok>

<tok id="w-87" nform="badil">badil</tok>

Como consecuencia, cuando se busca por forma normalizada se obtiene una concordancia como la de la figura 11:

contexto	. | Un tostador. | Un **badil** . | Una reja de arar	1844
contexto	. | .61- Un **vahil** y unas tenazas. | .	1868
contexto	. | n 89. Un **badil** en cincuenta centimos. | n 90	1882
contexto	. | Unas tenazas, un **vadil** y una cuchara de yerro	1754
contexto	en seis rs. | Un **vadil** , unas tenazas, una	1754
contexto	son cuarenta rs. | Un **vadil** de yerro en tres rs	1825
contexto	rl. | 55 ... Un **badil** , dos rs. | | Ropa	1829
contexto	a mo servir. | Un **vadil** de hierro para el fuego	1714

Figura 11. Búsqueda de la nform = badil

Después de que la ortografía ha sido normalizada viene la fase de lematización del corpus. Se trata de un proceso semiautomático, susceptible de revisarse en la plantilla de edición de los tokens (véase figura 10). Cuando se aplica el lematizador automático, TEITOK crea un nuevo atributo con el valor del lema, *recibir*, para una forma transcrita como *reziuio*:

<tok nform="recibió" id="w-97" lemma="recibir">reziuio</tok>

Con esta información se añade una nueva capa al corpus, que lo habilita para realizar búsquedas a un nuevo nivel. Por consiguiente, la búsqueda lematizada del

verbo *recibir* ofrecerá todos los ejemplos, visualizables en edición paleográfica o normalizada, que contengan el atributo lemma="recibir", como se muestra en la figura 12.

Por último, y de manera automática, se realiza el etiquetado morfosintáctico de cada documento. TEITOK utiliza el programa etiquetador Neotag (Janssen 2012), que añade un nuevo atributo al *tok*. En el caso de *reciuio*, en "reziuio juramento por Dios" (ODE, GR1713C2005), el *tok* incorpora la nueva información: pos[11]="VMIS3S0", interpretable como verbo principal (main), indicativo, pasado y tercera persona de singular.

contexto	manda, | otorga que ha	**recivido**	antes de ahora de la	1801
contexto	y en forma haverlos	**recivido**	antes de ahora, y	1801
contexto	dro: otorga que confiesa	**reziuir**	como reziue de la dha	1745
contexto	unos chorros y escopetas qe	**rezibio**	dho Damian Ba | lero, difunto	1775
contexto	biuda, de la que	**resibio**	juramto que hizo | segun dro	1776
contexto	y el menzionado alguacil mr	**recivio**	juramto pr Dios nro sor	1786
contexto	p ante mi el esno	**reziuio**	juramto p Dios y una	1713
contexto	sustentar las cargas matrimoniales,	**recivo**	de la dha mi señora	1709

Figura 12. Concordancia lematizada de recibir

La figura 13 recoge de manera comprensible para un usuario no experto toda la información que está procesada internamente en el elemento tok del documento XML de TEITOK:

form="reziuio"
nform="recibió"
pos="VMIS3S0"
lemma="recibir"

El nuevo atributo *pos* confiere al corpus un nivel extra de posibilidades de búsqueda, que se analizan en el apartado siguiente.

[11] Part of speech.

6. El buscador de ODE

En ODE la búsqueda por defecto se realiza sobre la forma normalizada. Entendemos que es la vía más rápida para encontrar todas las variantes formales de una palabra sin necesidad de conocerlas todas. Por ejemplo, al estándar *trébedes* le corresponden en el corpus las variantes *trévedes*, *trébedes*, *trebes*, *treves*, *extrévedes* y *estrébedes*, algunas de ellas muy poco previsibles.

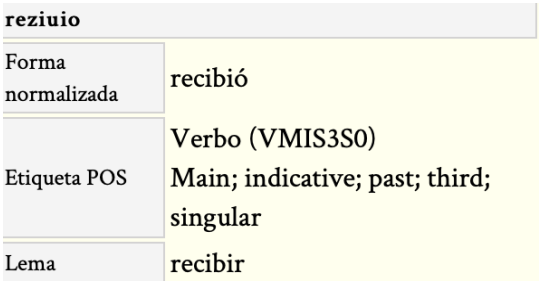

Figura 13. Información sobre el token *reciuio*

Además, TEITOK permite realizar búsquedas avanzadas en CQL (Corpus Query Language). Este protocolo sirve para hacer consultas complejas basadas en la combinación de distintos criterios: por ejemplo, mezclar una búsqueda del atributo *form* con el atributo *pos* para localizar todos los casos de la construcción *en + gerundio*:

```
[form = "en"] [pos = "VMG.*"]
```

O bien obtener todos los ejemplos de artículo determinado seguido de nombre propio, para lo que se recurre a la combinación de dos etiquetas morfosintácticas (POS):

```
[pos = "DA.*"] [pos = "NP.*"]
```

Igualmente, se podrían hallar todos los casos de pronombre *vos* seguido de cualquier forma verbal en segunda persona de plural:

```
[nform = "vos"] [pos = "VM.*2P.*"]
```

Por último, se pueden sumar condiciones mediante operadores lógicos. Por ejemplo, buscar todos los casos de nombres propios terminados en -ico:

[pos = "NP.*" & form= ".*ico"]

Como CQL exige conocimientos avanzados de lingüística de corpus y el manejo de las llamadas "expresiones regulares"[12], TEITOK proporciona a los usuarios un generador de búsquedas ("Query Builder") que facilita la tarea. Así, para buscar todas las formas verbales vinculadas a *haber*, por ejemplo, basta con escribir en el recuadro correspondiente (véase figura 14) el verbo en infinitivo, y automáticamente el buscador genera la secuencia correspondiente en CQL: [lemma = "haber"].

Figura 14. Generador de búsquedas ("Query Builder")

Para realizar búsquedas que incluyan información morfológica o categorial los usuarios tienen a su disposición el etiquetario de ODE, que utiliza una versión ligeramente modificada de las etiquetas de EAGLES[13] propuestas para el español, ya utilizadas en *Post Scriptum* (Vaamonde & Magro 2017). El sistema se basa en las posiciones: la primera letra representa la categoría principal[14], detrás de la cual, cada letra o número identifica determinados rasgos de la categoría. El valor 0 significa que se trata de un atributo no relevante. Así, por ejemplo, VMIP1S0

[12] Expresiones regulares son secuencias de caracteres que permiten encontrar patrones complejos de texto, y no solo palabras concretas. Por ejemplo, la expresión regular .*ico permite buscar todas las palabras de un corpus acabadas en "ico".
[13] The Expert Advisory Group on Language Engineering Standards.
[14] A (adjetivo), R (adverbio), D (determinante), N (nombre), V (verbo), P (pronombre), C (conjunción), I (interjección), Z (numeral), S (preposición), F (puntuación).

identifica a los verbos no auxiliares ("Main"), en indicativo (I), presente (P), primera persona (1) singular (S).

7. Conclusiones

La plataforma TEITOK permite ofrecer en línea corpus históricos normalizados, lematizados y etiquetados morfológicamente, con las enormes ventajas que eso supone para la realización de búsquedas complejas. Desde el punto de vista de la visualización, los documentos, en el caso de ODE, se ofrecen en tres versiones: paleográfica, normalizada y facsímil.

Todo esto se consigue porque los documentos se transcriben en XML. Las reticencias iniciales respecto de la dificultad y lentitud del proceso de etiquetado se diluyen, especialmente tras comprobar que marcar los textos siguiendo el estándar de XML-TEI no es más lento que hacerlo de acuerdo con criterios particulares. Además, los nuevos documentos son procesables informáticamente y los datos, estructurados y codificados de forma estándar, pueden compartirse con proyectos similares realizados en el entorno de TEI.

Por último, TEITOK resuelve ágilmente el problema de la fuente única, es decir, ofrece la ventaja de trabajar con un solo documento que contiene en su interior toda la información necesaria para visualizar distintas ediciones de un mismo manuscrito. La solución viene dada por la creación del elemento <tok>, en cuyo interior se encuentran, por un lado, todas las capas de visualización del documento (form, nform), y por otro, todos los niveles de búsqueda posibles (form, nform, lemma, pos). La semiautomatización de todo el proceso permite a los historiadores de la lengua crear corpus muy avanzados tecnológicamente sin ser expertos en lingüística computacional.

Agradecimientos

Este trabajo se inscribe en el marco del proyecto de investigación «Hispanae Testium Depositiones, HISPATESD», de referencia FFI2017-83400-P (MINECO / AEI / FEDER, UE).

Bibliografía

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

CHARTA. http://www.corpuscharta.es [21/05/2019].

Cíbola. The Cíbola Project. Editing the Documents of the Hispanic Southwest in the 16th and 17th Centuries. https://escholarship.org/uc/rcrs_ias_ucb_cibola [30/05/2019].

CORDE. *Corpus diacrónico del español*. http://www.rae.es [23/05/2019].

CORDIAM. Corpus diacrónico y diatópico del español de América. http://www.cordiam.org/ [04/05/2019].

COREECOM. http://www.corpuscharta.es/grupos.html [12/05/2019].

CorLexIn. Corpus léxico de inventarios. http://web.frl.es/CORLEXIN.html [12/05/2019].

Corpus del español. http://www.corpusdelespanol.org [15/05/2019].

Corpus del Nuevo Diccionario Histórico del Español (CDH). http://www.rae.es [23/05/2019].

Díaz Bravo, R. 2015. Herramientas computacionales aplicadas al estudio de la Historia de la lengua española. En *Temas, problemas y métodos para la edición y el estudio de documentos hispánicos antiguos*, Tirant lo Blanc: Valencia, 377-393.

Isasi, C., Spence, P., Lobo Puga, A., Martín Aizpuru, L., Pérez Isasi, S. & Pierazzo, E. 2014. *Guía para editar textos CHARTA según el estándar TEI: una propuesta* [02-05-2019].

Janssen, M. 2012. NeoTag: A POS Tagger for Grammatical Neologism Detection. In *Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC-2012). http://maarten.janssenweb.net/Papers/neotag-lrec.pdf. [05-06-2019].

Janssen, M. 2016. TEITOK: Text-faithful annotated corpora. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 4037-4043.

Martín Aizpuru, L. 2016. Algunos recursos informáticos al servicio de la edición de textos: la edición en XML-TEI. En *Hispanica Patavina. Estudios de historiografía e historia de la lengua española en homenaje a José Luis Rivarola*. Padua: CLEUP, 139-154.

Marttila, V. 2014. *Creating Digital Editions for Corpus Linguistics. The case of Potage Dyvers, a family of six Middle English recipe collections*. Helsinki: University of Helsinki.

Post Scriptum. Archivo digital de escritura cotidiana en Portugal y España en la Edad Moderna. http://ps.clul.ul.pt/es/index.php? [15/05/2019].

Sánchez-Prieto Borja, P. 2011. *La edición de textos españoles medievales y clásicos. Criterios de presentación gráfica*. San Millán de la Cogolla: CILENGUA.

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

Torruella, J. & Kabatek, J. 2018. *Portal de corpus históricos iberorrománicos (CORHIBER)*. http://www.corhiber.org/ [02/06/2019].

Vaamonde, G. 2015. P.S. Post Scriptum: Dos corpus diacrónicos de escritura cotidiana. *Procesamiento del lenguaje natural* 55: 57-64.

Vaamonde, G. & Magro, C. 2017. *Manual de edición y anotación en TEITOK de los materiales de P. S. Post Scriptum*. http://ps.clul.ul.pt/es/index.php?action=papers [13/05/2019].