Scriptum Digital
Vol. 7 (2018), pp. 25-35

EL *OLD SPANISH TEXTUAL ARCHIVE*, DISEÑO Y DESARROLLO DE UN CORPUS DE TEXTOS MEDIEVALES: LEMATIZACIÓN Y ETIQUETADO GRAMATICAL

FRANCISCO GAGO JOVER (*College of the Holy Cross*)
fgagojov@holycross.edu
ORCID-ID: https://orcid.org/0000-0002-8273-8791

F. JAVIER PUEYO MENA (*College of the Holy Cross*)
javier.pueyo@gmail.com
ORCID-ID: https://orcid.org/0000-0001-9067-5666

RESUMEN
Este trabajo expone los aspectos relacionados con el procesamiento de las formas, lemas, análisis gramatical y textos en el *Old Spanish Textual Archive* (*OSTA*), un corpus lingüístico de más de 32 millones de palabras, basado en las más de 400 transcripciones semi-paleográficas de textos medievales escritos en castellano, asturiano, leonés, navarro-aragonés y aragonés realizadas por los colaboradores del Hispanic Seminary of Medieval Studies (HSMS). Se describe además el proceso de etiquetado y lematización mediante el uso de *Freeling*, una herramienta de Procesamiento del Lenguaje Natural, y de *HSMS-app*, una herramienta de análisis textual desarrollada para este proyecto.
PALABRAS CLAVE: diseño de corpus electrónicos, anotación de corpus, corpus digitalizado del castellano antiguo, español medieval

THE *OLD SPANISH TEXTUAL ARCHIVE*, DESIGN AND DEVELOPMENT OF A CORPUS OF MEDIEVAL TEXTS: LEMMATIZATION AND POS TAGGING

ABSTRACT
This paper presents aspects related to the processing of forms, lemmas, grammatical analysis and texts in the *Old Spanish Textual Archive* (*OSTA*), a linguistic corpus of more than 32 million words, based on the more than 400 semi-paleographic transcriptions of medieval texts written in Castilian, Asturian, Leonese, Navarro-Aragonese and Aragonese prepared by the collaborators of the Hispanic Seminary of Medieval Studies (HSMS). It also describes the process of tagging and lemmatization using *Freeling*, a Natural Language Processing tool, and *HSMS-app*, a textual analysis tool developed for this project.
KEY WORDS: electronic corpus design, corpus annotation, digital medieval Spanish corpus, medieval Spanish

1. INTRODUCCIÓN

El *Old Spanish Textual Archive* (*OSTA*), un corpus lingüístico accesible en línea, permitirá efectuar búsquedas lingüísticas complejas en las más de 400 transcripciones semi-paleográficas de textos, escritos en castellano, asturiano, leonés, navarro-aragonés y aragonés, y producidos entre los siglos XIII y XVII, realizadas por los colaboradores del *Hispanic Seminary of Medieval Studies* (HSMS) desde los años 70 del pasado siglo[1].

En la fase inicial del proyecto, tras numerosas reuniones de trabajo y la consulta de otros corpus, establecimos las posibilidades mínimas de consulta que *OSTA* debía ofrecer

[1] El proyecto *OSTA* fue presentado en el *Simposio sobre fuentes digitales e historia de la lengua* organizado por Cilengua en San Millán de la Cogolla en octubre de 2016.

Fecha de recepción: 30/01/2017
Fecha de admisión: 15/07/2017
ISSN: 2014-640X

a los investigadores: forma original, forma normalizada, lema, categoría gramatical, fecha de producción, autoría, lugar de producción y materia (género). Esta decisión determinó tanto el método de procesado de los textos como la arquitectura de la base de datos —resumida gráficamente de la siguiente manera:

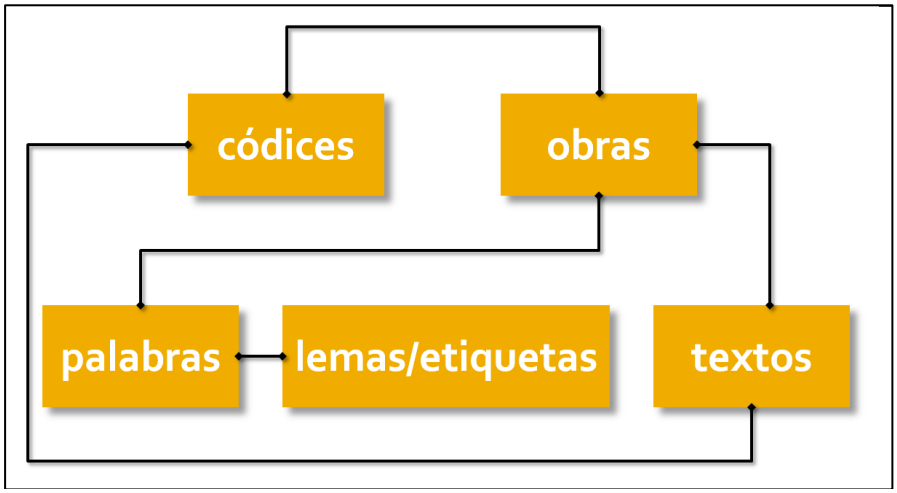

Figura 1. Arquitectura de *OSTA*

Los problemas encontrados y las soluciones propuestas para el procesado de códices y obras son analizados en un artículo previo[2], en el que también ofrecemos un breve panorama histórico del desarrollo de *OSTA*, desde la idea inicial (Nitti 1978) hasta la *Biblioteca Digital de Textos del Español Antiguo* (Gago 2015), su más reciente precedente. En este artículo nos centraremos, por lo tanto, en los aspectos relacionados con las formas, lemas, análisis gramatical y textos.

2. *OSTA* EN CIFRAS

En la actualidad *OSTA* contiene más de 32 millones de palabras (agrupadas en 38.500 lemas diferentes), procedentes de 350 códices (manuscritos e impresos) distribuidos por siglos como sigue:

[2] «El *Old Spanish Textual Archive*, diseño y desarrollo de un corpus de textos medievales: el corpus textual» (en preparación).

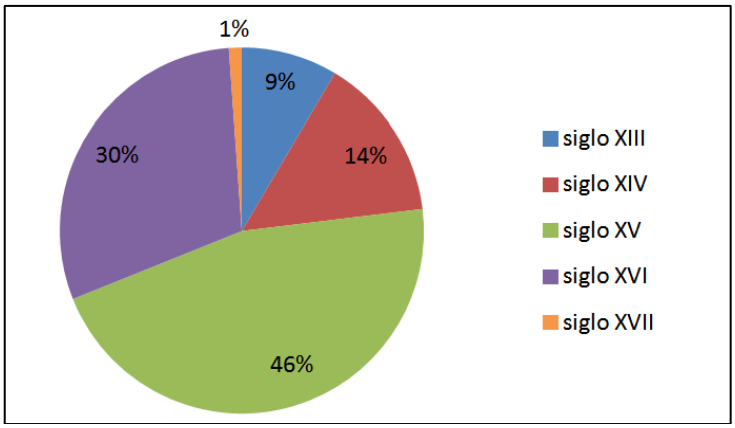

Figura 2. Distribución de códices por siglo

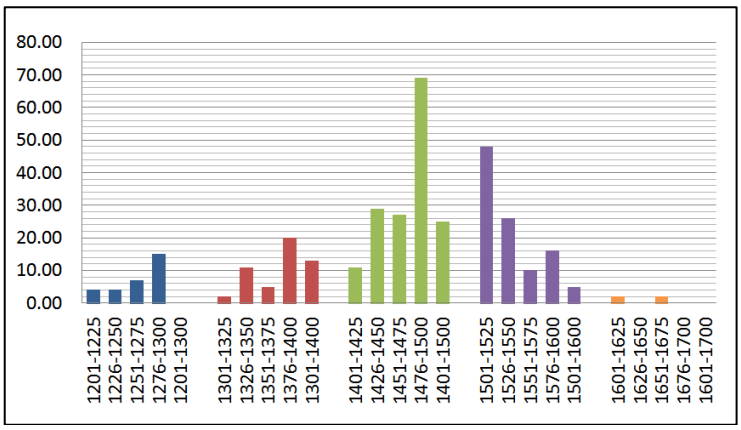

Figura 3. Distribución de códices por 25 años

En número de formas la distribución general del corpus por siglos sería la siguiente:

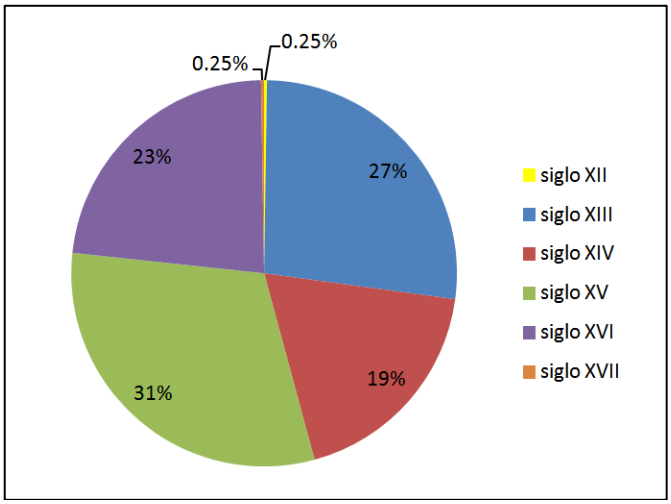

Figura 4. Distribución de palabras por siglo[3]

[3] La discrepancia en el número de siglos entre las figuras 2 (siglos XIII-XVII) y 4 (siglos XII-XVII) es debida a la datación utilizada en cada una de las tablas de las que se extrae la información. Los códices (figura 2) son datados según su fecha específica de producción (la fecha de la copia de un manuscrito o de la impresión

Hasta el momento se ha identificado un total de 1.645 obras (1.433 con diferente título) y 297 autores. De estas, 1.105 son textos en verso (con 1.107.919 palabras en total), y 540 son textos en prosa (31.298.618 palabras).

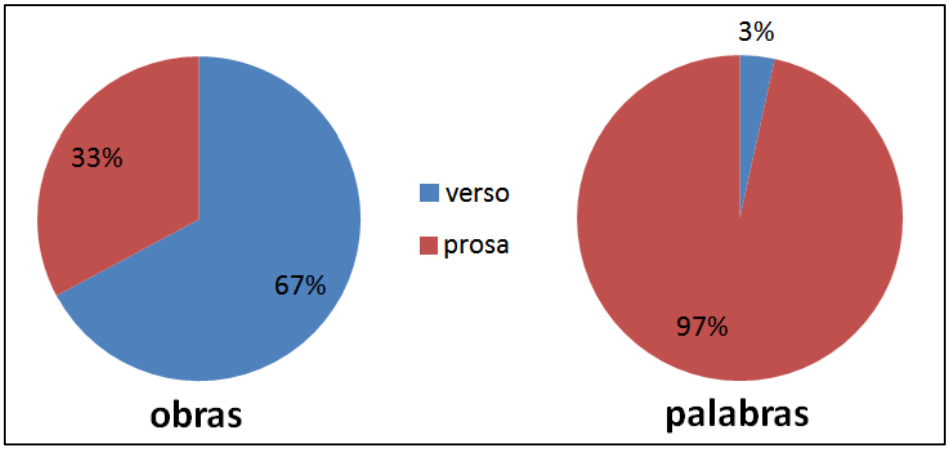

Figura 5. Distribución de verso y prosa

Finalmente, por lo que respecta a las lenguas, se han identificado 14 combinaciones de lenguas en el corpus:

lengua(s)	n.º obras
aljamiado árabe	1
aragonés	25
aragonés + castellano	4
castellano	1538
castellano + gallego	5
castellano + latín	3
castellano + leonés	2
castellano occidental	9
gallego	39
leonés	5
navarro	2
navarro + navarro-aragonés	3
navarro-aragonés	7
riojano + castellano	1

3. ETIQUETADO Y LEMATIZACIÓN: HERRAMIENTAS Y PROCESO

Para el procesamiento lingüístico de las transcripciones semi-paleográficas del HSMS y para poder añadir nuevos niveles de lematización y etiquetado gramatical, decidimos utilizar *Freeling* 3.1 (Carreras 2004; Padró 2011 y 2012), una herramienta de Procesamiento del Lenguaje Natural diseñada para el análisis multilingüe. *Freeling*

de una edición), mientras que las obras (figura 4) son datadas de acuerdo con la fecha original de producción, correspondiente a la fecha de redacción conocida o supuesta del original de cada obra.

encadena diferentes procesos y programas (*tokenizer*, *splitter*, analizador morfológico, anotador lingüístico, desambiguador, etc.) y devuelve textos anotados lingüísticamente mediante el uso de una serie de recursos léxicos o diccionarios (de léxico común, de topónimos y antropónimos, de abreviaturas, de expresiones locutivas, etc.), de reglas (sean reglas de segmentación textual o sean reglas lingüísticas para la identificación de clíticos, análisis de morfemas derivativos, reconocimiento de terminaciones verbales, etc.) y de un modelo probabilístico, creado a partir de un corpus de entrenamiento corregido manualmente, que se aplica durante las etapas de desambiguación de homónimos y de análisis de las formas desconocidas.

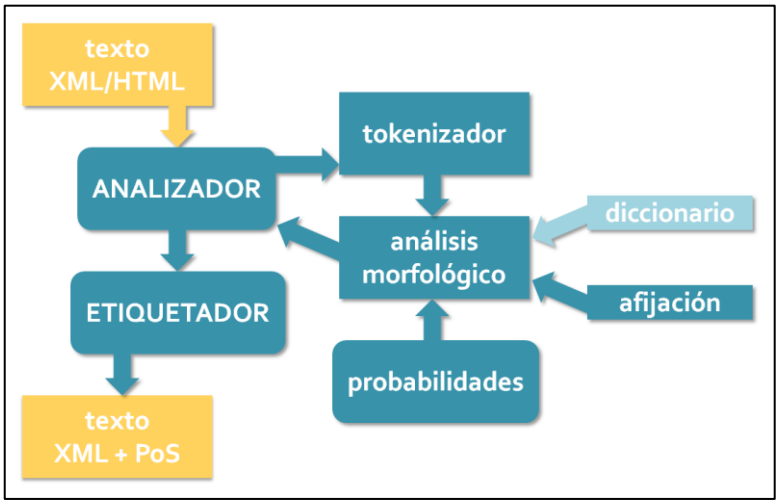

Figura 6. Diagrama de flujo de *Freeling*

Afortunadamente para nuestro proyecto, la adaptación de *Freeling* al español medieval ya había sido realizada, al menos en una fase inicial, por Sánchez Marco, Boleda y Padró (2011) quienes, tras utilizar parte de las transcripciones semi-paleográficas del HSMS para conformar su *golden corpus*, entrenar el programa, y crear y modificar los recursos y las reglas en los ámbitos señalados, demostraron que *Freeling* podía ser utilizado con éxito para la anotación lingüística de textos diacrónicos del español. Por nuestra parte, hemos ampliado considerablemente los recursos lingüísticos y las reglas de *Freeling* para poder dar cuenta de todo el léxico y de todos los fenómenos lingüísticos sistemáticos de la colección textual completa del HSMS. Una vez se corrija manualmente una parte significativa del corpus, volveremos a entrenar a *Freeling* sobre un corpus más amplio y correctamente distribuido, lo que mejorará aún más la precisión de su análisis estadístico. Igualmente se ha creado un módulo de normalización ortográfica, siguiendo en lo posible los criterios de edición desarrollados por la red internacional CHARTA (CHARTA 2013), que dará la opción de realizar consultas mediante formas normalizadas ortográficamente (*avemos*), en lugar de hacerlo sobre las posibles formas paleográficas subyacentes (*avemos*/*auemos*/*havemos*/*hauemos*).

El procesado del corpus textual fue agilizado mediante *HSMS-app*, una herramienta de análisis textual propia, desarrollada por F. Javier Pueyo Mena, que integra las librerías de *Freeling* y lo combina con una serie de scripts para que, a partir del texto plano de las transcripciones del HSMS, sea posible obtener un texto en formato XML con toda la información lingüística incorporada, pero manteniendo todas las características textuales

de la obra para facilitar su lectura y su posterior presentación en los resultados de las consultas.

Figura 7. *HSMS-app*

HSMS-app nos permite procesar uno o más textos, procesar un corpus textual completo, o listar los textos ya procesados para seleccionarlos de forma individual y proceder a su edición manual.

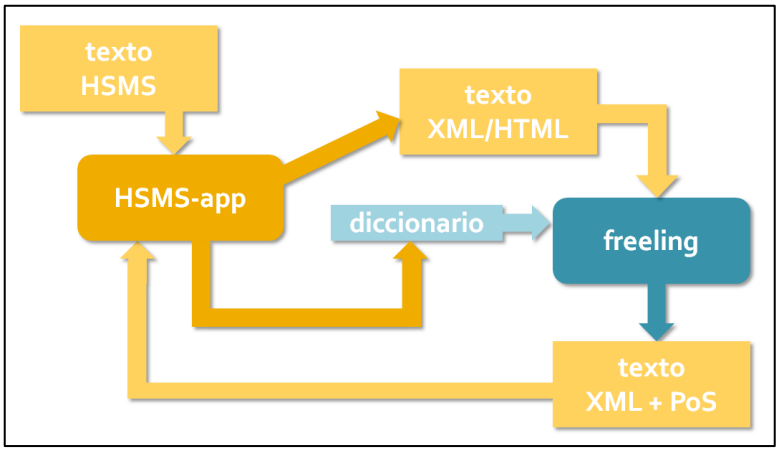

Figura 8. Diagrama de flujo de *HSMS-app*

Después de seleccionar el texto, la transcripción en formato HSMS[4] que queremos procesar, en nuestro ejemplo TEXT.DAC.txt, correspondiente a la *Disputa del alma y el cuerpo* (AHN: Clero: carp. 279, n. 22)

[4] Todos los textos incluidos en *OSTA* están transcritos según las normas establecidas en *A Manual of Manuscript Transcription for the Dictionary of the Old Spanish Language*, preparado originalmente por Mackenzie (1977) y revisado en su quinta edición por Harris-Northall (1997). El complejo sistema de etiquetado allí descrito intenta reproducir el original de la forma más fiel posible, representando ciertos aspectos del formato de la página original y la disposición del texto en ella, así como respetando la

[fol. 1r]
{CB1.
{RMK: HSMS-0345-0001: Disputa del alma y el cuerpo.}
[*s]i q<ue>reedes oyr loq<ue> uos quiero dezir dizre uos loq<ue> ui nol uos i quedo fallir un sabad[*o]
[*e]sient dom[i]ngo amanezient ui una gra<n>t uision en mio leio dormient era<m> asem[*eian]-
t q<ue> so un lenzuelo nueuo jazia un cu[*e]rpo de uemme muerto ell alma era fuera [*e] fuert mientre q<ue> plera ell alma es ent esida desnuda ca non uestida e guisa
[*du]n jfant fazia duelo tan gra<n>t tan gra<n>t duelo fazie al cue[^r]po maldizie fazi[*e]
[*ta]n gra<n>t de duelo e maldizie al cuerpo al cuerpo dixo ell alma de ti lieuo ma[*la] fama tot siemp<r>et maldizre ca por ti penare q<ue> nu<n>ca fe[^ci]st cosa q<ue> seme[^i]as fer[*mo]-
sa ni de nog ni de dia delo q<ue> io q<ue>ria n<un>ca fust a altar por j buena oferda dar ni diez[*mo]
ni prim<i>cia ni buena penite<n>ci[*a] ni fecist oracion nu<n>ca de corazo[n] c[*u]a[n]do iuas all el[*gue]-
si[a] asentauaste a conseia i fazies tos conseios e todos tos(dos) treb[e]ios apostol ni martjr [??]
quisist seruir iure par la tu tiesta q<ue> no curaries fiesta nu<n>ca de nigun santo no
[*cure]st so disa<n>to mas not faran los santos aiuda mas q<ue> a una bestia muda mezquino mal
[??] ta mal ora fuest nado q<ue> tu fu[*este] tan rico agora eres mesquinu di<m> o so<n> tos d<inero>s q<ue> tu mi[??]
estero o los tos m<<o>><rauedis> azaris et melequis q<ue> solies manear et a menudo contar o son [*lo]s pala[*fres]
q<ue> los quendes ie los res te solien dar por to loseniar los cauallos corientes [^las] espuelas
[*pu]nentes las mulas bien amblantes asuueras traina<n>tes los frenos esorados los
[*petr]ales dorados las copas doro fino con q<ue> beuies to uino do son tos bestim<en>tos olos
[*tos] guarnim<en>tos q<ue> tu solies festir e tanbien re[??]}

HSMS-app genera un fichero—TEXT.DAC.xml—en el que se han sustituido las etiquetas propias del HSMS por otras en formato XML:

ortografía original, la puntuación y, siempre que sea razonable, los patrones percibidos de separación de palabras. Preservándose también las contracciones, elisiones, apócopes, y aféresis originales, y señalando claramente todos los casos de corrección efectuada por el transcriptor.

```
<FOL>1r
<CB1>
<CB>
<RMK> HSMS-0345-0001: Disputa del alma y el cuerpo.</RMK>
<EDITADD>*s</EDITADD>i q<ABB>ue</ABB>reedes oyr loq<ABB>ue</ABB> uos quiero dezir
dizre uos loq<ABB>ue</ABB> ui nol uos i quedo fallir un sabad<EDITADD>*o</EDITADD>
⋮
⋮
<EDITADD>*tos</EDITADD> guarnim<ABB>en</ABB>tos q<ABB>ue</ABB> tu solies festir e
tanbien re<EDITADD>??</EDITADD></CB>
</CB1>
</FOL>
```

Este fichero—TEXT.DAC.xml—es entonces procesado en *Freeling*, que devuelve un fichero— TEXT.DAC.tagged.xml—al que ha añadido la lematización, normalización y el análisis morfológico de cada una de las palabras:

```
<FOL>1r
<CB1>
<CB>
<RMK> HSMS-0345-0001: Disputa del alma y el cuerpo.</RMK>
si•si+sí•si•CS#si@CS@sí@PP3CNO00@sí@RG@si@NCMS000
quereedes•†quereedes•querer•VMIP2P0 oyr•oír•oír•VMN0000
loque_el•†loque_el•el•DA0NS0 loque_que•†loque_que•que•PR0CN000
uos•vos+vós•tú•PP2CS00P quiero•quiero•querer•VMIP1S0
dezir•dezir•decir•VMN0000#decir@VMN0000@decir@NCMS000
dizre•dizré•decir•VMIF1S0#decir@VMIF1S0@dezir@VMIF1S0 uos•vos+vós•tú•PP2CS00P
loque_el•†loque_el•el•DA0NS0 loque_que•†loque_que•que•PR0CN000 ui•vi•ui•I
nol_no•nol_no•no•RN nol_le•nol_le•le•PP3CSD00 uos•vos+vós•tú•PP2CS00P
i•y+ý•i•VMIS1S0#i@VMIS1S0@i@NCFS000@i@Z
quedo•quedó+quedo•quedar•VMIP1S0#quedar@VMIP1S0@quedo@AQ0MS0@quedo@I
fallir•fallir•fallir•VMN0000 un•un•uno•DI0MS0#uno@DI0MS0@uno@PI0MS000
sabado•sábado•sábado•NCMS000
⋮
⋮
tos•tos•tos•NCFS000#tu@DP2CPS@tos@NCFS000@tos@NCFS000 </EDITADD>
guarnimentos•†guarnimentos•guarnimiento•NCMP000
que•que+qué•que•PR0CN000#que@PR0CN000@que@CS@qué@PE000000
tu•tu+tú•tú•PP2CSN00#tú@PP2CSN00@tu@DP2CSS solies•soliés•soler•VMII2S0
festir•†festir•vestir•VMN0000
e•y+he•e•CC#e@CC@y@CC@haber@VAIP1S0@haber@VMIP1S0@e@NCFS000
tanbien•también•también•RG re•re•re•AQ0CN0#re@NCMS000@re@AQ0CN0 ?•?•?•Fit
?•?•?•Fit </CB>
 </CB1>
 </FOL>
```

Las etiquetas usadas en la anotación gramatical siguen el estándar EAGLES para lenguas europeas[5]. El primer carácter en la etiqueta es siempre la categoría gramatical (PoS), que también determina la longitud de la etiqueta y la interpretación de cada carácter en la etiqueta:

forma	lema	etiqueta	
guarnimentos	guarnimiento	NCMP000	nombre/común/masculino/plural
quiero	querer	VMIP1S0	verbo/principal/indicativo/presente/primera/singular
un	uno	DI0MS0	determinante/indefinido/masculino/singular
tanbien	también	RG	adverbio/general

A partir de este momento, una vez procesado el texto, *HSMS-app* permite incorporar en el diccionario las formas desconocidas por *Freeling* (que se muestran en rojo) y revisar tanto los análisis generados automáticamente por las reglas lingüísticas creadas (en azul) como los análisis asignados a formas homónimas potencialmente ambiguas (en verde). Para ello nos servimos de las opciones que aparecen en la columna derecha de la interfaz de *HSMS-app*, donde podemos editar la forma paleográfica, la forma normalizada, el lema propuesto y la etiqueta morfológica asignada. De esta manera, los cambios efectuados en un texto serán utilizados en el análisis de cualquier otro texto, con lo que se facilita enormemente el proceso de etiquetado.

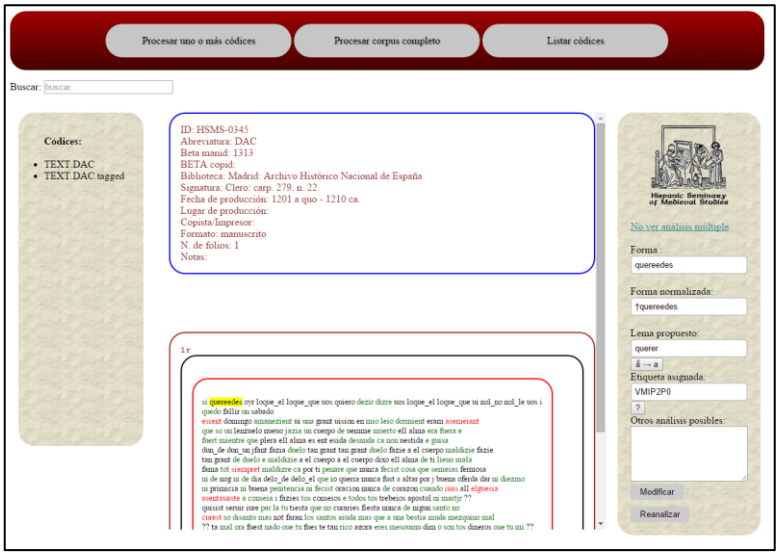

Figura 9. Interfaz de *HSMS-app*

4. PLANES DE FUTURO

A fecha de hoy, el número de formas sin analizar mediante *Freeling* es de unas 900.000 (es decir, algo menos del 3% del corpus). En los próximos meses dedicaremos la mayor parte de nuestro tiempo a ampliar los recursos léxicos de *Freeling*, trabajando en el

[5] Las etiquetas utilizadas por *Freeling* aparecen descritas en https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets/tagset-es.html

reconocimiento de entidades nombradas (topónimos y antropónimos), de variantes ortográficas medievales y de palabras no identificadas por ninguna de las reglas desarrolladas.

Inicialmente esperamos poder incrementar el número de entradas en el diccionario de *Freeling* mediante el procesamiento de varios de los diccionarios del proyecto *Dictionary of the Old Spanish Language* del HSMS: el *Diccionario español de textos médicos antiguos* (Herrera 1996), el *Diccionario de términos militares del castellano medieval* (Gago 1997), el *Diccionario español de documentos alfonsíes* (Sánchez 2000), y el *Diccionario de la prosa castellana del Rey Alfonso x* (Kasten y Nitti 2002). Un análisis preliminar del contenido de estos diccionarios nos permite afirmar con cierta confianza la posibilidad de añadir unas 30.000 entradas más al diccionario. De esta forma esperamos rebajar el número de formas sin analizar en todo el corpus a 700.000 (un 2% del corpus). En una fase posterior, la edición manual de las 3.000 formas sin identificar más frecuentes solucionaría otros 240.000 casos, dejando el porcentaje de formas desconocidas muy cerca del 1%.

El objetivo final, como no puede ser de otra manera, es poner el corpus a disposición de todos los investigadores mediante la creación de una interfaz online que permitirá combinar búsquedas en todos los niveles de marcación anotados y filtrar los resultados según los campos recogidos en la meta-descripción de cada obra y de cada códice.

REFERENCIAS BIBLIOGRÁFICAS

CARRERAS, Xavier, Isaac CHAO, Lluís PADRÓ y Muntsa PADRÓ (2004): «*FreeLing*: An Open-Source Suite of Language Analyzers», *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. http://nlp.lsi.upc.edu/publications/papers/carreras04.pdf

CHARTA (2013): *Criterios de edición de documentos hispánicos (Orígenes-siglo XIX) de la red internacional CHARTA*. http://www.redcharta.es/criterios-de-edicion

GAGO, Francisco (2015): «La *Biblioteca Digital de Textos del Español Antiguo* (BiDTEA)», *Scriptum Digital*, 4, pp. 5-36.

GAGO, Francisco (1997): *Diccionario de términos militares del castellano medieval*. Madison: University of Wisconsin. Tesis doctoral.

HERRERA, María Teresa, *et al.* (1996): *Diccionario español de textos médicos antiguos*, 2 vols. Madrid: Arco/Libros.

KASTEN, Lloyd A. y John J. NITTI (2002): *Diccionario de la prosa castellana del Rey Alfonso x*. New York: Hispanic Seminary of Medieval Studies.

MACKENZIE, David (1977): *A Manual of Manuscript Transcription for the Dictionary of the Old Spanish Language*. Madison: Hispanic Seminary of Medieval Studies.

MACKENZIE, David y Ray HARRIS-NORTHALL (1997): *A Manual of Manuscript Transcription for the Dictionary of the Old Spanish Language*, 5.ª ed. Madison: Hispanic Seminary of Medieval Studies. http://hispanicseminary.org/manual-es.htm

NITTI, John (1978): «Computers and the Old Spanish Dictionary», *Computers and the Humanities,* 12, pp. 43-52.

PADRÓ, Lluís y Evgeny STANILOVSKY (2012): «*FreeLing* 3.0: Towards Wider Multilinguality», *Proceedings of the Language Resources and Evaluation Conference* (*LREC* 2012). http://nlp.lsi.upc.edu/publications/papers/padro12.pdf

PADRÓ, Lluís (2011): «Analizadores Multilingües en *FreeLing*», *Linguamatica*, 3, 2, pp. 13-20. http://nlp.lsi.upc.edu/publications/papers/padro11.pdf

SÁNCHEZ, María Nieves (2000): *Diccionario español de documentos alfonsíes*. Madrid: Arco/Libros.

SÁNCHEZ-MARCO, Cristina, Gemma BOLEDA y Lluís PADRÓ (2011): «Extending the tool, or how to annotate historical language varieties», *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (Portland, OR, USA, 24 June 2011), pp. 1-9. http://nlp.lsi.upc.edu/papers/sanchezmarco11.pdf