Primera aproximación para la extracción automática de *Entidades Nombradas* en corpus de documentos medievales castellanos

Mª Eugenia Iglesias Moreno
Depart. Biblioteconomía y Documentación, Depart. Informática, Universidad Carlos III de Madrid
meugenia.iglesias@uc3m.es

Pilar Azcárate Aguilar-Amat
Depart. Biblioteconomía y Documentación, Depart. Informática, Universidad Carlos III de Madrid
pilar.azcarate@uc3m.es

Sonia Sánchez Cuadrado
Depart. Biblioteconomía y Documentación, Depart. Informática, Universidad Carlos III de Madrid
sonia.sanchez.cuadrado@uc3m.es

1. INTRODUCCIÓN

El reconocimiento y clasificación de entidades nombradas (NE – Named Entities)[1] supone una tarea importante para la identificación de nombres propios de personas, lugares y organizaciones. Constituye una subtarea para la extracción y recuperación de información. Los sistemas para el reconocimiento de NE utilizan habitualmente técnicas basadas en métodos estadísticos o en gramáticas lingüísticas. Se han desarrollado diferentes investigaciones para el reconocimiento de NE, sobre todo en inglés aunque en los últimos años se han desarrollado diversos trabajos para otros idiomas como el árabe (Shaalan y Raza, 2007), turco (Metin, Kışla y Karaoğlan, 2012) o español (Galicia-Haro, Gelbukh y Bolshakov, 2004). Para los corpus de documentación medieval los estudios se han centrado, principalmente, en la composición y etiquetado de corpus informatizados con el propósito de elaborar diccionarios y diferentes instrumentos de representación del conocimiento como tesauros u ontologías. Es el caso, a nivel nacional, del *Tesouro Medieval Informatizado da Lingua Galega (TILG)*[2], el *Corpus Diacrónico del Español* (CORDE)[3], *Corpus de Documentos Españoles anteriores a 1700 (CODEA)*[4], *CHARTA*[5] y el *Proyecto Biblia medieval*[6]. A nivel internacional mencio-

[1] Chinchor N.: MUC-7 Named Entity Task Definition (1997). <http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/overview.html>.
[2] Proyecto de investigación realizado en el Instituto da Lingua Galega (ILG) de la Universidade de Santiago de Compostela (USC). <http://ilg.usc.es/tmilg/>.
[3] RAE. <http://corpus.rae.es/cordenet.html>.
[4] Grupo de Investigación de Textos para la Historia del Español (GITHE). Universidad de Alcalá. <http://demos.bitext.com/codea/>.
[5] Corpus Hispánico y Americano en la Red: Textos Antiguos. <http://www.charta.es/>.
[6] <http://www.bibliamedieval.es/>.

naremos el *Corpus Informatizado do Português Medieval (CIPM)*[7], el *Tesoro DEM Informatizado (DEMi)*[8] *y el corpus Hispanic Seminary of Medieval Studies (HSMS)*[9].

El objetivo de este trabajo es analizar los problemas surgidos en la anotación automática de los nombres propios contenidos en documentos medievales en castellano medieval con herramientas de procesamiento del lenguaje natural (PLN). Para ello, utilizaremos la aplicación Freeling para español estándar y la variante diacrónica del español de los siglos XII al XVI (Sánchez-Marco, Boleda y Padró, 2011) en un corpus formado por los documentos que componen el *Libro Becerro de las Behetrías de Castilla* (en adelante LBB), del siglo XIV. En esta primera aproximación, nos hemos centrado en la evaluación manual de los resultados obtenidos en el etiquetado de los antropónimos y topónimos. Proponiendo finalmente una adaptación de los módulos de la aplicación con el propósito de aumentar el nivel de acierto del etiquetado y facilitar el posterior reconocimiento y clasificación de las Entidades Nombradas en un trabajo futuro.

Este artículo se organiza en los siguientes epígrafes: en la Sección 2 revisamos las características del corpus LBB. En la Sección 3 describimos el método seguido para la selección del corpus de anotación, evaluación manual de errores y propuesta de adaptación de Freeling. En la Sección 4 se mostrarán los resultados de la evaluación manual. Finalmente, presentaremos algunas conclusiones y trabajo futuro (Sección 5).

2. DESCRIPCIÓN CORPUS LBB

Para esta investigación se ha seleccionado el corpus LBB, compuesto por la transcripción de Gonzalo Martínez Díez (1981) de los 2.109 documentos que integran el *Libro Becerro de las Behetrías de Castilla* de mediados del siglo XIV, en formato electrónico. Es un censo escrito en romance dónde se registran más de dos mil poblaciones castellanas, agrupadas en quince merindades y localizadas en el territorio de Castilla situado al norte del río Duero (Estepa Diez, 2003). El corpus LBB contiene un total de 210.609 tokens, de los que 4.460 son diferentes (Tabla 1).

La elección de este corpus documental se justifica, principalmente, porque posee una alta representación de nombres propios de persona y lugar, objetivo principal de estudio en nuestra investigación.

Tabla 1. Descripción del Corpus LBB

Corpus LBB	
Nº Documentos	2.109
Nº Caracteres	1.085.080
Nº Palabras	208.278
Nº Palabras diferentes	4.268
Nº Tokens	210.609
Nº Tokens diferentes	4.460

[7] Centro de Linguística de la Universidade Nova de Lisboa (CLUNL). <http://cipm.fcsh.unl.pt/>.

[8] Anterior *Diccionario del Español Medieval (DEM)*. Centro de Investigación de la Academia de Ciencias y Letras de Heidelberg. <http://www.adw.uni-heidelberg.de/dem/fichero/ficherolista.html>.

[9] Proyecto del Dictionary of the Old Spanish Language (DOSL). University of Wisconsin-Madison. <http://www.hispanicseminary.org/index-es.htm>.

Los documentos se distribuyen por merindades, división administrativa del reino de Castilla en el siglo XIV, a la que pertenecía cada una de las poblaciones registradas (Tabla 2).

Tabla 2. Distribución de documentos del Corpus LBB

Nº documentos por merindad			
Nº merindad	*Nombre merindad*	*Nº documentos*	*Nº palabras*
I	Cerrato	93	10.854
II	Infantazgo de Valladolid	64	6.848
III	Monzón	93	10.340
IV	Campos	68	5.967
V	Carrión	119	9.579
VI	Villadiego	104	10.007
VII	Aguilar de Campóo	262	23.933
VIII	Liébana-Pernía	129	10.641
IX	Saldaña	194	15.468
X	Asturias de Santillana	179	28.585
XI	Castrojeriz	114	15.436
XII	Candemuñó	73	6.393
XIII	Burgos-Ubierna	117	8.359
XIV	Castilla Vieja	371	34.311
XV	Santo Domingo de Silos	129	11.557
TOTAL:		2.109	208.278

Un análisis manual previo del corpus LBB nos ha permitido determinar su estructura y las principales características respecto a los nombres propios. Teniendo en cuenta que se trata de un censo, el corpus presenta homogeneidad en la estructura de los textos. Los registros de cada población, en la mayoría de los casos, incluyen los siguientes elementos: i) nombre del lugar, ii) obispado al que pertenece (opcional), iii) relación socio-jurídica, acompañada seguidamente del señor (señores) o institución al que pertenece, iv) derechos del rey y v) derechos del señor o institución.

El análisis de las características se ha centrado en la revisión de signos de puntuación, de transcripción, detección de variaciones gráficas y diferentes estructuras de los antropónimos y topónimos propias del castellano medieval. Para evitar problemas en el posterior procesamiento automático de segmentación y etiquetado, se ha procedido a la adaptación manual correspondiente del texto.

Los signos de transcripción incluidos en el texto aparecen entre corchetes o paréntesis y las palabras agregadas en letra cursiva. Se tomaron las siguientes consideraciones:

- Eliminación de los signos que contenían información aclaratoria del transcriptor ([*sic*], [en blanco], [?]) que no formaban parte del texto original.
- Eliminación de paréntesis y corchetes incorporando los caracteres al texto para un primer análisis.

Los nombres propios dentro del texto se destacan gráficamente con mayúsculas, según indican las normas de transcripción de Gonzalo Martínez Díez (1981: 105), *"(…), en la utilización de mayúsculas y minúsculas y en el sistema de puntuación seguimos el criterio actual."*. Las variaciones gráficas más relevantes que presentan los nombres propios de persona y de lugar son:

- Uso de caracteres que no existen en español moderno "Ç" y "ç".
- Variaciones del uso de u y v. Por ejemplo, Val Buena de Duero se escribe como "Ual Buena" y "Val Buena".
- Variaciones con omisión o cambio de un carácter para un mismo nombre (Pedro y Pero, Royz y Ruyz, etcétera).

Los antropónimos presentan diferentes estructuras formadas por un elemento (nombre), dos elementos (nombre + apellido patronímico/toponímico) o tres elementos (nombre + apellido patronímico + topónimo). En algunos casos preceden al nombre otros elementos como la fórmula de tratamiento don o donna (don Pedro, don Fernando Sanchez de Ualladolid, donna Margarita, etcétera).

La elisión de una vocal se indica mediante un apóstrofo, *d'* o *D'*. En los nombres propios de persona compuestos se localiza al comienzo de los apellidos para indicar la pertenencia de individuos a familias o poblaciones, siguiendo la fórmula *artículo* + *d'* + *familia/lugar* (los d'Aça, los d'Escouadas de Yuso, etc.). En los nombres de lugares es utilizado para diferenciar o situar la población a la que hace referencia (Castriel d'Oniello, Castro d'Ordiales). En el corpus LBB se han localizado un total de 185 nombres propios que contienen apóstrofo.

Los errores tipográficos que han dado lugar a la adaptación correspondiente del texto en formato electrónico se pueden agrupar en:

- Omisión de puntos a final de frase.
- Minúsculas después de punto final que corresponden a punto y coma.
- Apóstrofo en d' y D', que indica la elisión de una vocal en los nombres propios, incluido como acento.

3. MÉTODO

El método para la evaluación del etiquetado automático de los nombres propios que aparecen en los documentos del Corpus LBB, se ha diseñado considerando que partimos de un corpus no anotado con las características descritas en el epígrafe anterior. Dicho método es el resultado de la combinación del adoptado en proyectos para enriquecer automáticamente corpus históricos con información lingüística, basado en el etiquetado automático con herramientas existentes seguido de la corrección humana, y el planteado por Sánchez-Marco, Boleda y Padró (2011) para la adaptación de la aplicación Freeling para la variante diacrónica del español de los siglos XII al XVI. Se ha estructurado en las siguientes fases:

- Selección de un corpus de anotación.
- Anotación automática del corpus y evaluación manual de los resultados de etiquetado.
- Propuesta de adaptación de la herramienta de PLN, Freeling.

3.1. Selección de un corpus de anotación

Para la evaluación del etiquetado automático de nombres propios con una herramienta de PLN, en esta primera aproximación, se ha seleccionado una muestra del corpus LBB que hemos denominado corpus de anotación. Este corpus de anotación está formado por 803 documentos, correspondientes a las 7 primeras merindades de las 15 que componen el corpus completo (Cerrato, Infantazgo de Valladolid, Monzón, Campos, Carrión, Villadiego y Aguilar de Compóo) y contienen un total de 78.201 tokens, de los que 2.141 son diferentes. Se localizaron todas las palabras que comenzaban por mayúscula candidatas a ser nombres propios, obteniendo un total de 15.338, siendo 1.250 diferentes (Tabla 3).

Tabla 3. Distribución de tokens en Corpus LBB y corpus anotación

	Corpus LBB	*Corpus anotación*
Nº tokens	210.609	78.201
Nº tokens diferentes	4.460	2.141
Tokens comienzan mayúsculas	38.369	15.338
Tokens diferentes comienzan mayúsculas	2.739	1.250

3.2. Etiquetado automático del corpus de anotación de LBB

Para el procesamiento automático de los documentos del corpus de anotación, hemos tomado como punto de partida la versión 3.0 de Freeling[10]. La elección de FreeLing 3.0 como herramienta de etiquetado se ha basado, principalmente, en que como define Lluís Padró (2011), Freeling es una librería de código abierto para el procesamiento multilingüe automático, que proporciona una amplia gama de servicios de análisis lingüístico para diversos idiomas y, que además, incluye una ampliación para el español antiguo (Sánchez-Marco, Boleda, y Padró, 2011).

La arquitectura modular de FreeLing permite el acceso y modificación de los diferentes recursos lingüísticos que incluye (lexicones, gramáticas, diccionarios, etc.), lo que nos ha permitido extender los datos contenidos en los archivos con vocabulario y formas lingüísticas propias del castellano medieval (siglo XIV), contenidas en el corpus LBB.

El proceso de etiquetado morfológico se ha llevado a cabo en dos iteraciones con la aplicación analyzer incluida por defecto en Freeling 3.0, la primera con los módulos para el español estándar y antiguo facilitados y la segunda con nuestra propuesta de adaptación.

3.2.1. Evaluación manual de errores de etiquetado automático de Nombres Propios

En la primera iteración se realizó un análisis morfológico del corpus de anotación aplicando las herramientas para ambas variantes del español, actual y antiguo, de Freeling. Este analizador está basado en las etiquetas EAGLES (v.2.0)[11]

[10] Freeling 3.0. <http://nlp.lsi.upc.edu/freeling/>.
[11] Etiquetas EAGLES. <http://www.lsi.upc.edu/~nlp/tools/parole-sp.html>.

que asigna NP00000 a los nombres propios. De los archivos resultado de éste primer análisis se han extraído, mediante scripts en Python, los tokens con la etiqueta NP00000 obteniendo un total de 8.939 para el español estándar y 2.254 en el caso del español antiguo. Tras una primera revisión manual del etiquetado se determinó modificar la configuración de análisis para el español antiguo que contemplara las opción de reconocimiento de entidades nombradas (NE) y de palabras compuestas o *multiword* imprescindible para los nombres propios compuestos. Se obtuvo un resultado de 5.731 tokens etiquetados con NP00000.

Los principales errores detectados han sido los siguientes:

- **Nombres de persona con la estructura *donna* + *Nombre Propio(NP).*** En el texto se incluyen nombres de persona que sólo aparecen bajo la forma *donna* + *NP*, determinante para reconocer las variantes de nombres de persona a lo largo del corpus. El analizador reconoce la palabra donna como nombre común femenino singular (NCFS000) y como adjetivo calificativo femenino singular (AQ0FS0). Confirmamos que este problema no surgía con las expresiones de la forma *don + Nombre Propio (NP)*.
- **Números romanos en mayúsculas**. La finalidad principal de los documentos del corpus LBB es definir a quién correspondía percibir los derechos y rentas expresadas con números romanos en mayúsculas. Esto supone un problema en el análisis con Freeling 3.0 para español estándar que asigna una etiqueta de Nombre Propio en unos casos y de Nombre Común en otro.
- **Nombres propios seguidos de número romanos.** Por ejemplo, *don Nunno VI* o *Diego Garçia los XIII,* en este caso el analizador etiqueta como nombre propio la expresión entera al considerarla una palabra múltiple o *multiword*.
- **Formas verbales propias del castellano medieval.** Por ejemplo: Pagauan, Dixieron.
- **Artículos, pronombres, conjunciones, contracciones y adverbios propios del español antiguo.** Reconoce como NP determinados artículos, pronombres, adverbios y conjunciones escritos en mayúsculas al comienzo de frase. Por ejemplo: E, Lon, Otrosi, Deste *(de + este)*, etcétera.

3.2.2. Propuesta de adaptación de la herramienta de PLN, Freeling

Nuestra propuesta ha incluido la adaptación de Freeling 3.0 mediante la ampliación de los siguientes módulos:

- Inclusión de donna en la lista de palabras que deben formar parte como afijo de las Named Entities en el módulo *Basic NER,* sección *<Affixes>* como "donna PRE". El resultado ha sido el etiquetado correcto de todas las expresiones del corpus de anotación.
- Ampliación del diccionario de español estándar con la lista de números romanos en minúscula, formas verbales, artículos, pronombres, conjunciones y adverbios propios del español antiguo.
- Modificación del submódulo del diccionario dónde se establecen las reglas para sufijos (<Suffixes>). Se han sustituido los patrones para que reconozca pronombres enclíticos propios del español antiguo.

Para la segunda iteración, se ha seguido el mismo método que en la anterior, pero utilizando la herramienta con las modificaciones propuestas. Después de la

extracción de los tokens etiquetados con NP00000, se han obteniendo un total de 4.570 tokens para español estándar y 4.485 para el español antiguo.

4. RESULTADOS

La evaluación manual de los resultados de ambas iteraciones nos ha permitido contrastar el nivel de acierto de la anotación automática de nombres propios de personas y lugares con nuestra adaptación de la herramienta Freeling (Tabla 4). En este primer análisis hemos considerado como acierto el reconocimiento y etiquetado completo de todos los elementos que integran la estructura de nombres propios.

En la primera iteración con Freeling 3.0, se obtuvo un nivel de acierto de un 49,21% para español estándar siendo el principal error de etiquetado los números romanos que se escriben con mayúsculas en los documentos del corpus. En este caso suponen un 46,04% de los errores de etiquetado. Para español antiguo (old-es), el nivel de acierto ha sido de 75,41%.

En la segunda iteración con la adaptación propuesta de la herramienta Freeling los niveles de acierto han sido de 97,81% para español estándar, que supone un 48,6% de mejora frente al obtenido en la iteración anterior (Fig. 1), y 98,23% en el caso del español antiguo (Fig. 2).

Tabla 4. Evaluación de resultados de la anotación automática de nombres propios

	1ª Iteración		*2ª Iteración*	
	FreeLing_Orig_ES	*FreeLing_Orig_OLD*	*FreeLing_Modif_ES*	*FreeLing_Modif_OLD*
NP Total	8939	5731	4570	4485
NP Diferentes	1.769	1.599	1.460	1.403
Aciertos	4.398	4.322	4.470	4.406
Errores	4.541	1.409	100	79

Una vez analizados los resultados de la segunda iteración, podemos concluir que se han solucionado los siguientes errores de etiquetado como nombre propio: i) todos los números romanos del corpus de anotación se le ha asignado una etiqueta de número, ii) se han reconocido las expresiones *donna* + *NP,* iii) formas verbales, artículos, pronombres, conjunciones, contracciones y adverbios propios del español antiguo en el corpus de anotación, se han etiquetado correctamente.

5. CONCLUSIONES Y TRABAJO FUTURO

En este artículo hemos presentado una primera aproximación a la anotación automática de nombres propios en corpus de documentación en castellano medieval. Como primer paso, se ha evaluado el etiquetado de la herramienta de procesamiento del lenguaje natural, Freeling, en dos iteraciones con un corpus seleccionado del corpus LBB: i) Freeling para el español estándar y antiguo por defecto y, ii) Freeling adaptado con nuestra propuesta.

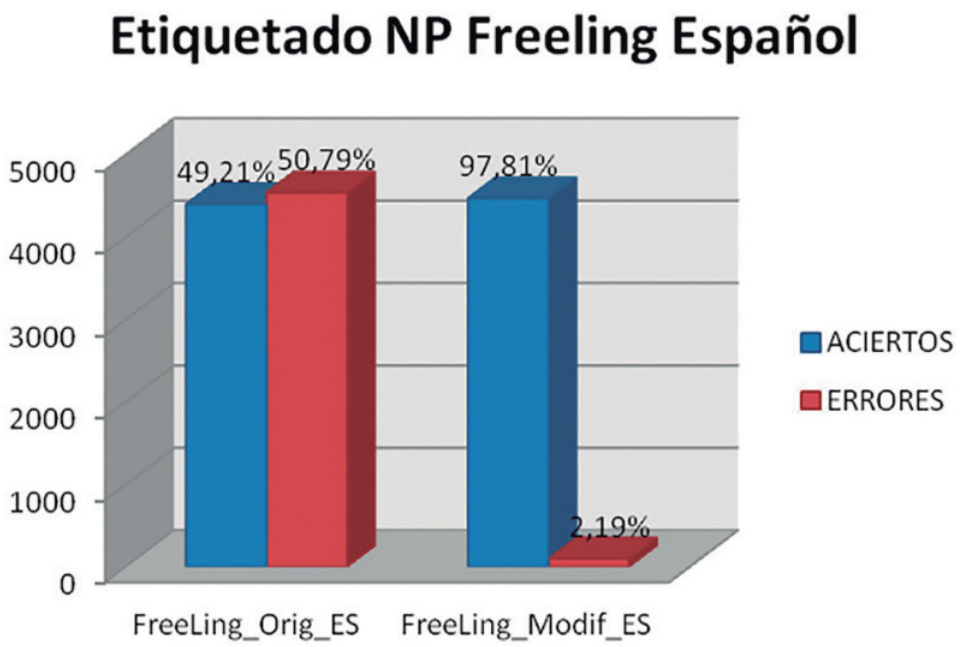

Fig. 1. Nivel de acierto de etiquetado nombres propios con Freeling para español estándar.

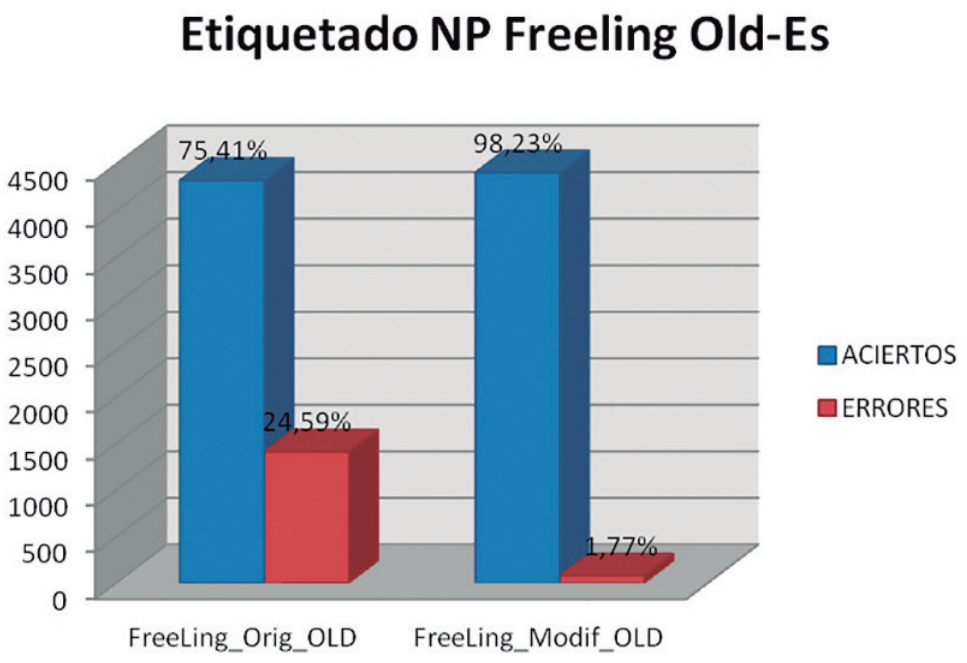

Fig. 2. Nivel de acierto de etiquetado nombres propios con Freeling para español antiguo (old-es).

La revisión manual de los resultados nos ha permitido detectar y clasificar los errores, diseñando nuestra aproximación con el objetivo principal del etiquetado correcto de las estructuras que forman los nombres propios, los números romanos, artículos, pronombres, conjunciones y adverbios, utilizados en el español antiguo basándonos en los que aparecen en el corpus.

Este trabajo, comparando el nivel de acierto de la anotación automática, nos indica que existe un margen para la mejora en los resultados de etiquetado de nombres propios de personas y lugares con una herramienta de procesamiento del lenguaje natural. Así, el nivel de acierto alcanzado en la segunda iteración, 97,81% para el español estándar y 98,23% para español antiguo, nos hace pensar que supone un resultado aceptable para el etiquetado de este tipo de documentos, teniendo como referencia el nivel de precisión de 94,5% obtenido por Sánchez-Marco *et al*. (2011).

Como trabajo futuro nos planteamos repetir el método presentado con el corpus LBB completo con la posterior evaluación para confirmar estos primeros resultados obtenidos. Así como, la extracción y clasificación de las Named Entities de personas, lugares e instituciones del corpus.

AGRADECIMIENTOS

Para este trabajo se ha utilizado la transcripción del *Libro Becerro de las Behetrías de Castilla* de Gonzalo Martínez (1981), contenida en la base de datos cedida por el Grupo de Investigación *"QUAESTIO. Sociedades medievales: marcos, redes y procesos",* al que pertenecen los doctores Cristina Jular y Julio Escalona, ambos investigadores titulares del Instituto de Historia del CSIC.

BIBLIOGRAFÍA

Estepa Díez, Carlos, *Las behetrías castellanas,* Valladolid, Junta de Castilla y León, Consejería de Cultura y Turismo, 2003.

Galicia-Haro, Sofía N.; Gelbukh, Alexander y Bolshakov, Igor A., "Recognition of named entities in spanish texts", en Raúl Monroy *et al*. (eds.), *MICAI 2004: Advances in artificial intelligence*, *Third Mexican International Conference on Artificial Intelligence,* Mexico City, Mexico, April 26-30, Springer, 2004, pp. 420-429.

Marrero, Mónica *et al*., "Named Entity Recognition: Fallacies, Challenges and Opportunities", *Computer Standards & Interfaces,* 35 (2013), pp. 82-89.

Martínez Díez, Gonzalo, *Libro Becerro de las Behetrías,* León, Centro de Estudios e Investigación San Isidoro, 1981.

Metin, Senem Kunoba; Kışla, Tarik y Karaoğlan, Bahar, "Named Entity Recognition in Turkish Using Association Measures", *Advanced Computing: an International Journal,* 3 (2012), pp. 43-49.

Padró, Lluís, "Analizadores multilingües en freeling", *Linguamatica*, 3 (2011), pp. 13-20.

Padró, Lluís y Stanilovsky, Evgeny, "FreeLing 3.0: Towards Wider Multilinguality". *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*, Estambul, European Language Resources Association (ELRA), 2012, pp. 2473-2479.

Shaalan, Khaled y Raza, Hafsa, "Person name entity recognition for Arabic", *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources,* Praga, Association for Computational Linguistics, 2007*,* pp. 17-24.

Sánchez-Marco, Cristina; Boleda, Gema y Padró, Lluís, "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, USA, Association for Computational Linguistics, 2011, pp. 1-9.

Xavier, María Francisca, "Variação e mudança lexical no Português Medieval o caso dos verbos", *DOMÍNIOS DE LINGU@GEM – Revista Eletrônica de Lingüística*, 3-2 (2009), pp. 224-242.

RESUMEN

En este artículo presentamos los resultados de una evaluación de la anotación de nombres propios de forma automática en un corpus de documentación medieval castellana. Dicha evaluación se ha realizado sobre el etiquetado obtenido con la herramienta de procesamiento de lenguaje natural, Freeling, en dos iteraciones. La primera, con la versión para español estándar y antiguo facilitadas y la segunda con una adaptación propuesta, basada en la solución de los problemas de anotación debidos a las características y variantes que presentan los nombres propios de personas y lugares en español antiguo. Para ambas iteraciones, se ha seleccionado un corpus de anotación de los documentos que componen el *Libro Becerro de las Behetrías de Castilla* (LBB), del siglo XIV. El nivel de acierto obtenido en la anotación automática de nombres propios con la adaptación propuesta ha sido de 98,23% para el español antiguo, que puede considerarse aceptable para repetir, en un trabajo futuro, el método en el corpus completo.

Palabras clave: Lingüística de Corpus, Anotación de corpus, Documentación medieval, Reconocimiento y Clasificación de Entidades Nombradas.

ABSTRACT

This paper presents the results of evaluating the automatic recognition and annotation of proper names in a corpus of Castilian medieval documents. The evaluation has been done by adapting Feeling, an existing tool for natural language processing. This paper describes the two iterations of this evaluation: the first iteration, using the version for standard and old Spanish, and the second iteration, using an adaptation that has been created based on the problems found in the first iteration. Such problems were mainly caused by the inherent characteristics and variants of proper names and names of places in old Spanish. For that purpose, a corpus of 14th century documents of the *Libro Becerro de las Behetrías de Castilla* (LBB) was used. The proposed adaptation for old Spanish leads to a 98.23% level of success, which indicates that it can be used in the future evaluation of the entire corpus.

Keywords: Natural Language Processing, Medieval Documents, Corpus Annotation, Named Entity Recognition and Classification.