




TEI-friendly annotation scheme for medieval named entities: a case on a Spanish medieval corpus

Elena Álvarez-Mellado¹  · María Luisa Díez-Platas¹ · Pablo Ruiz-Fabo¹ · Helena Bermúdez¹ · Salvador Ros¹ · Elena González-Blanco²

Accepted: 28 October 2020
© The Author(s) 2021

Abstract Medieval documents are a rich source of historical data. Performing named-entity recognition (NER) on this genre of texts can provide us with valuable historical evidence. However, traditional NER categories and schemes are usually designed with modern documents in mind (i.e. journalistic text) and the general-domain NER annotation schemes fail to capture the nature of medieval entities. In this paper we explore the challenges of performing named-entity annotation on a corpus of Spanish medieval documents: we discuss the mismatches that arise when applying traditional NER categories to a corpus of Spanish medieval documents and we propose a novel humanist-friendly TEI-compliant annotation scheme and guidelines intended to capture the particular nature of medieval entities.

✉ Elena Álvarez-Mellado
elena@scc.uned.es

María Luisa Díez-Platas
ml.diezplatas@linhd.uned.es

Pablo Ruiz-Fabo
pablo.ruiz@linhd.uned.es

Helena Bermúdez
helena.bermudez@linhd.uned.es

Salvador Ros
sros@scc.uned.es

Elena González-Blanco
egonzalezblanco@linhd.uned.es

¹ Digital Humanities Innovation Lab (LINHD), School of Computer Science, UNED University, Madrid, Spain

² CoverWallet, Madrid, Spain

Keywords Named-entity annotation · Annotation scheme · Historical NER · Medieval named entities · Medieval Spanish corpus

1 Introduction

Written text is the primary means of access to the past. Political treaties, legal records, commercial transactions, notarial deeds, chronicles: these historical documents constitute the main piece of evidence of historical events. Our understanding of historical developments relies heavily on the textual information they contain. However, this historical data is frequently sparse and fragmentary, scattered among many documents or even unmanageable. As valuable as it may be, data is hardly of any use if it cannot be retrieved effectively. That is why applying Information Extraction (IE) techniques to historical documents once it has been digitalized and transcribed properly is particularly fruitful: automatically extracting the names of places, people or organizations mentioned on historical documents can provide us with valuable historical evidence that otherwise would be extremely cumbersome to obtain or could even go unnoticed under a pile of documents.

Named entity is a widely used term in Natural Language Processing (NLP) that refers precisely to textual information units like person names, location names, organization names, etc. The task of identifying these entities is known as Named Entity Recognition (NER). NER is one of the main subtasks of IE (Grishman and Sundheim 1996) and a key step in text analysis, as it provides information on what and who are mentioned in a given text (Nadeau and Sekine 2007). Consequently, NER has steadily remained a hot topic in NLP for the last twenty years.

NLP has dealt extensively with the mechanisms and difficulties that NER poses (Tjong Kim Sang and De Meulder 2003). Numerous NER systems have been proposed, traditionally trained on easily available large general-domain corpora written in resource-rich languages, such as newspaper articles or newswire (Desmet and Hoste 2014). However, general-domain techniques tend to perform poorly when used for unseen genres (Plank 2016; Poibeau and Kosseim 2001). Historical documents present remarkable differences compared to general-domain texts. In fact, historical language has been described as a type of under-resourced language (Pettersson et al. 2013) that is highly non-canonical both in terms of content (genre and domain) and form (linguistic variation, lack of orthographic standardization) (Sánchez-Marco et al. 2011). As a result, NER annotation schemes designed with modern general-domain text in mind fail to capture the nature of historical named entities.

Although historical documents have traditionally been neglected in terms of text analysis in favor of more accessible and industry-appealing fields (such as the biomedical, social or journalistic domain), a growing number of interdisciplinary projects born within the field of Digital Humanities (which deals with the digitization, preservation and study of historical documents) are promoting a data-driven approach to historical texts (Piotrowski 2012). Previous work has shown the potential of applying and adapting Digital Humanities standards (such as TEI-XML,

the standard for the digital encoding of texts in the humanities) to perform information extraction tasks on historical texts. For example, initiatives like the Pleiades Gazetteer or Pelagios Commons (Isaksen et al. 2014) localize places mentioned in documents from the Antiquity and collect them in an open and interoperable format. And Maraoui et al. (2018) adapted TEI-XML standards to encode Arabic person names in order to annotate the Hadith corpora, a collection of narrations of the deeds of Prophet Muhammad that are relevant for understanding the Qur'an jurisprudence. However, the medieval world still remains mostly unexplored in terms of NER exploitation, in spite of the attractiveness of its vast textual tradition.

Annotating is the process of enriching a collection of text by adding linguistic and interpretative information to it (Hovy and Lavid 2010). Annotation is a fundamental step towards the development and evaluation of a NER system: human-annotated corpora provide us with valuable data that can serve both as training material for machine learning approaches as well as a gold standard against which new algorithms and techniques will be evaluated, Pustejovsky and Stubbs (2012). Consequently, it is key to have a set of shared criteria to follow that describe the nature and format of the annotation process, i.e. an annotation scheme. Annotation schemes contribute to narrowing down the linguistic phenomena to be tackled and produce more reliable and coherent annotations (Bayerl et al. 2003). This is particularly interesting when dealing with highly non-canonical texts, such as historical language, where non-standardized spelling and format can make reaching an agreed consensus on which text sequences should be annotated particularly elusive.

In this paper we present a novel humanist-friendly annotation scheme and annotation guidelines specifically designed to capture the particular nature of medieval named entities. This annotation proposal is an adaptation of traditional name entity classifications tailored to suit the medieval domain and to facilitate the application of NER techniques (which are far from optimal to reflect the essence of medieval reality) to medieval documents. This annotation scheme complies with XML-TEI markup, the XML standard for annotation and representation of texts among digital humanists.

The annotation scheme we present in this paper is conceived for the identification of named entities in Medieval Spanish texts. The motivation to undertake this annotation task was to provide support (both in terms of annotation scheme and gold standard creation) to HisMeTag (Hispanic Medieval Tagger) (Díez Platas et al. 2017, 2020), a NER tool that was developed in collaboration with Pelagios project¹, whose goal is to supply linked open data methods to relate and explore historical places. The purpose of HisMeTag is to support digital humanists in the analysis of medieval texts by detecting, annotating and geolocating historical places on a given document through contextual information. The F1 score obtained by this tool ranges between 0,63 and 0,89 depending on the time period of the document (Díez Platas et al. 2017).

¹ See <http://commons.pelagios.org/>.

This paper is organized as follows: in Sect. 2 we explain which named entity categories have been considered for the present scheme; in Sect. 3 we describe the annotation scheme itself (both the annotation format and the tagset); in Sect. 4 we provide the annotation guidelines that should be followed when applying the annotation scheme to a medieval corpus and how to deal with problematic cases that can occur in medieval documents; Sect. 5 contains the application of the scheme and guidelines to annotate a Spanish corpus of medieval documents along with the obtained results and error analysis; finally, Sect. 6 contains the conclusions and future work.

2 Annotation framework

In this section we introduce the types and idiosyncrasies of named entities in the medieval domain.

2.1 Rationale

Named entities are usually classified according to the type of entity they refer to (such as person, location, etc). These categories are highly domain-dependent: while a journalistic NER scheme will be concerned with categories such as places or people, a biomedical NER scheme will deal with categories such as molecules or drugs. Therefore, it is key to define the set of NER categories and tailor it to the text domain that will be annotated.

In order to assess the need for tailored-made guidelines for the annotation of medieval named entities, several examples of medieval documents in Spanish were inspected, along with some general domain guidelines such as ACE guidelines (Linguistic 2005), MUC-7 guidelines (Chinchor 1998) and the Extended Named Entity Hierarchy (Sekine 2003). Named entity annotation on Medieval Spanish texts presents three specific challenges:

- (i) The diversity of orthographic norm (Sánchez-Marco et al. 2011).
- (ii) The complexity of morphosyntatic structures that can appear with proper nouns during this time period (with named entities occurring inside comprehensive structures that can include geographical origin or lineage of the person being mentioned: *rey Alfonso de Castilla*, “King Alfonso of Castile”). These patterns create a set of dependencies between the entity and its joined attributes that the annotation scheme has to account for.
- (iii) The profusion of person entities that are mentioned exclusively either through a nickname or through their social role within society (with no explicit person name whatsoever, for example: *El Cid*). The use of nicknames and role names as a way to refer to someone is by no means exclusive of the medieval domain and, in fact, can be found in other general-domain texts. However, the abundance of this phenomenon in medieval documents is extremely high: nicknames and role names may appear in modern texts along with the actual name of the person being

Table 1 Examples of named entities for every category: general-domain vs medieval world

Named-entity category	Medieval example	General-domain example
Person name	<i>Alexander the Great, Aphrodite</i>	<i>Elvis Presley, Hillary Clinton</i>
Location name	<i>Kingdom of León, Burgos Cathedral</i>	<i>New York City, Brooklyn Bridge</i>
Organization name	<i>Order of Saint James, the Trojans</i>	<i>The Washington Post, The Beatles</i>
Role name	<i>The King of Castilla, the Pope</i>	–
Miscellanea	<i>Babieca, Tizona</i>	<i>Nessie</i>

mentioned. In the medieval domain, however, nicknames and role names are frequently the only way in which a person is referred to: therefore, not considering role names or nicknames as named entities would mean missing a great amount of entities appearances that will not be referred to in any other way in the text.

2.2 Named-entity categories

In this section, we will describe the categories that have been considered for our annotation scheme. The proposed annotation scheme considers five general named-entity categories: **person names**, **location names**, **organization names**, **role names** and **miscellanea**. Person, location and organization names are inherited categories from the MUC 7 guidelines (Chinchor 1998). Role name is a tailor-made category to annotate medieval titles, positions and social relationships; this category is unprecedented in named entity categories and, to the best of our knowledge, is an original contribution of this work. Finally, miscellanea is an open-ended category that seeks to cover any historically-relevant named entity that does not fit into any of the other categories and that the annotators could decide that needed to be annotated (a precedent of this category can be found on Sekine et al. (2002) under the Name_Other category). The primary objective of this annotation project is to annotate places and people mentioned in medieval documents. Consequently, other entities (such as events, quantities, etc) were considered beyond the scope of the annotation project. The nature and scope of the five named-entity categories considered in this work are as follows (see Table 1 for comparative examples):

- **Person names** are proper names that refer to a person (e.g. *Alexander the Great*). When dealing with medieval documents, person names should be expected to refer to kings and queens, knights, nobility, gods and other mythological figures. Person names can be first names, surnames or a combination of both. Nicknames and aliases (which were quite frequent in medieval society and even work as official names on legal documents) will also be considered a type of person name. Both real people and fictitious characters will be considered as a person entity. Sekine et al. (2002) classify fictitious characters as a product rather than a person (what are fictitious characters but a

type of mental human production?). However, this distinction between reality and fiction becomes rather blurred when dealing with medieval chronicles: historical accounts of true events are frequently sprinkled with mythical characters and legends. Therefore, no NER category distinction will be made between historical figures and literary characters within the medieval domain. Likewise, gods, deities and religious figures names (including biblical nicknames and divine aliases) will also be considered as a type of person entity.

- **Location names** refer to any location, including geopolitical units (countries, regions, towns, kingdoms etc; e. g. *Constantinople*, *Kingdom of León*), geographical names (mountains, rivers, landforms in general; e.g. *Mediterranean Sea*) or facilities (buildings and monuments, like castles, bridges or monasteries; e.g. *Church of Santa Gadea*). Generally speaking, any geolocalizable place (i.e. any given location that can be tracked on a map) will be considered a location entity². Although religious and mythical places are not geolocalizable, they will also be considered a special type of location name (e.g. *Garden of Eden*, *Paradise* or *Styx river*).
- **Organization names** are associations, institutions or any other group of people. While in the journalistic domain organization entities tend to be companies, sports team or music bands, in medieval documents such associations are normally religious orders, armies or governmental institutions (e.g. *Order of St. James*, *Aragonese Courts*). Following (Sekine 2003), the organization category will be considered in a broad sense and will be applied to any coherent group of people that share a name and a sense of unity or cause, like names of peoples or religious and ethnic communities (e.g. *the Greeks*, *the Trojans*, *the Moors*, *the Jews* or *the Christians*).
- **Role names** are names that refer to medieval positions, nobility titles, professions or family relations. In broad terms, role names are any named relationships or states that establish a type of bond between two or more individuals, an individual and a group or an individual and a geopolitical entity. These relationships can work *de facto* as a way to identify someone within the medieval society and include titles like “the queen” or “the Pope” as well as professions like “the baker” or “the priest” and relations like being someone’s parent or someone’s heir. Role names (such as “King of Castilla”) enable the establishment of relationships at three levels: firstly, between the person and their role; secondly, between the role and the place upon which this role has relevance; and finally, by inference, between the person and the geographic location. In traditional general-domain named-entity hierarchies and schemes, role names are not considered named entities. However, previous work has demonstrated that role names are a valuable piece of information to be retrieved for the Digital Humanities community. After all, role names allow to establish and contextualize relationships between people and their social position, their lineage, the places to which they belong, etc. See as an example the work developed by Murray (2017), where relationship dynamics from heterogeneous

² For a report about the annotation of place names in literary texts in a TEI environment see Frontini et al. (2016).

types of Medieval texts are analyzed focusing on tensions between family members, including the study of domestic violence. Therefore, although information related to the social role, lineage, nobility title or geographical origin of the person being mentioned is not usually encoded in traditional NER schemes, it should not be overlooked when working on the medieval domain. Furthermore, role names play a central role in medieval documents: kinship (and family relationships in general) are a fundamental identifier on legal documents (for example, when it involves nobility titles or inheritance); likewise, professions and positions become especially relevant in political documents. In fact, these roles and relationships are the usual way (and some times even the only way) in which certain people are referred to in medieval texts, as it is not the person itself that is central, but the institution or geopolitical entity that their position or title represents: both in legal and literary works, it is often the case that a certain person is exclusively identified through their social role name (e.g. *the prior*, *the king*). As a result, roles in medieval documents meet the defining criteria of referential unicity, referential autonomy, denominational stability and referential relativity established by Fort teal. (2009) for named entities. On the basis of the above, role names will be considered named entities in this annotation scheme and therefore annotated as such.

- **Miscellanea** covers any other named entity that does not fit in any of the previous categories and that may be useful to historical research. Some examples of miscellaneous named entities that can be found in medieval documents are important objects that were significant enough within the medieval society to bear a proper name of their own, like sword names (e.g. *Excalibur*, *Tizona*, *Colada*), horses names (*Bucéfalo* or *Bucephalus*, the horse of Alexander the Great), book names (*the Gospels*) or chants and prayers (*Hail Mary*). This category includes every entity that is personified in the text with a proper noun. It should be considered that the act of naming weapons, animals or religious artifacts comes from the historical or literary role that those named entities fulfill, from a cultural or narrative point of view. In any case, future refinements of this category are expected after evaluating with a group of experts the research questions that this type of information might help address.

3 Annotation scheme: format and tagset

In this section we describe the annotation scheme: first, the encoding format is described; second, a comprehensive description of the tagset itself is provided.

3.1 Named-entity annotation format

Annotation coding formats should be as clear, reusable, unambiguous, semantically adequate, consistent, expressive, platform-independent and self-explanatory as possible (Ide and Romary 2004; Stede and Huang 2012).

In addition to these general criteria, when dealing with named-entity annotation we should bear in mind the nature and scope of the texts that we aim to annotate: after all, named-entity annotation for medieval manuscripts (or any other historical document) concerns mainly scholars working on the field of cultural heritage, social sciences and the humanities, fields which tend to be quite apart from the NLP community and its developments and standards. Fortunately, there is a common ground that enables the exchange between these two apparently unrelated areas of knowledge: the Text Encoding Initiative (Text 2008).

XML-TEI markup scheme has become the prevailing standard for the representation of texts in digital form among digital humanists (Pierazzo 2016). In fact, the suitability of XML-TEI format for NER purposes was pointed out as early as MUC-7 guidelines (Chinchor 1998), although the proposal never gained full traction within the NLP community. XML-TEI meets the requirements in terms of annotation (re)usability and generalization, and its tagset is wide, flexible and expressive enough to account accurately for the historic NER categories within the medieval textual tradition.

Furthermore, XML-TEI format allows to limit the annotation tagset to a few common general tags while adding as many specificities and subcategorization as needed in the attributes. This feature is particularly interesting when addressing a highly specific field like the medieval domain, as it allows the annotation to be as comprehensive as needed while avoiding producing too high a number of different categories in the annotation scheme, which can cause lower interannotator agreement levels (Bayerl et al. 2003). This combination of coarse-grained tags with fine-grained attributes enables a double level where XML-TEI standard tags coexist with more granular attributes, ensuring an interoperable, reliable and yet enriched and comprehensive annotation.

Likewise, XML format facilitates element nesting, which can be extremely useful when annotating medieval entities that contain other entities within, which is usually the case when annotating nobility titles. Shall the annotator encounter a nobility title name such as *king of Castilla and Aragón*, XML nesting will allow annotating both the main entity (the role name *king of Castilla and Aragón*) and the geographical locations linked to the nobility title (*Castilla* and *Aragón*), if a more detailed annotation is preferred.

Due to its interoperability, semantic expressiveness, nesting flexibility and popularity within the Digital Humanities community, we consider XML-TEI the ideal format for medieval named-entity annotation.

3.2 Named-entity annotation scheme

We will now list the proposed XML-TEI tags for every named-entity category described in Sect. 2.2, along with the attributes and values that should be added to cover all subcategories. Every XML-TEI tag will be illustrated with examples extracted from our Spanish medieval corpus, which will be described in Sect. 5. For the sake of clarity, either translations or equivalent examples in English will also be provided. It should be noted that some of the attribute values that we propose have no documented previous use and are (to our knowledge) original contributions of

this work. These undocumented attribute values can be distinguished from the TEI-complying attribute values because the former ones are marked using an underscore (`type = "_value"`). The proposed annotation scheme for medieval named-entities is inspired by the MUC-7 and ACE guidelines (Chinchor 1998; Linguistic 2005) and by Sekine's Extended Named Entity Hierarchy, Sekine (2003). See Table 2 for a comparison between the proposed annotation scheme and Chinchor (1998), Linguistic (2005) and Sekine (2003).

3.2.1 Person names

First names, surnames and family names will be annotated using the XML-TEI tag `<persName>`. Although TEI offers specific tags for first names and surnames, this annotation scheme makes no distinction between first names and surnames, as this distinction becomes particularly fuzzy when it comes to medieval and historical names. The entire name shall be annotated using one `<persName>` tag only.

```
<persName>Celestina</persName>
<persName>Rodrigo Díaz</persName>
```

Nicknames and aliases have two different annotation approaches, depending on whether the nickname appears as part of the official name or *in lieu* of it. The reason behind this distinction is to comply with the TEI guidelines³:

1. If the nickname appears as part of the official name, it will be annotated within the `<persName>` tag using the nested TEI tag `<addName>`. This will enable nickname retrieval and classification. The `<addName>` tag should also be used to tag epithets (such as *Alexander the Great* or *Pedro el Cruel*) and regnal numbers (*Alfonso VI*).

```
Yo soy <persName>Ruy Diaz <addName>el Cid Campeador</addName</persName>
[I am <persName>Ruy Diaz <addName>The Warrior</addName</persName>]
```

```
<persName>Pedro <addName>el Cruel</addName</persName>
[<persName>Pedro <addName>The Cruel</addName</persName>]
```

```
<persName>Inocencio <addName>IV</addName>
[<persName>Innocent <addName>IV</addName>]
```

2. Lone nicknames (i.e. the nickname appears by itself, with no official name around) will be annotated using the `<persName>` tag followed by the TEI

³ "Additional name contains an additional name component, such as a nickname, epithet, or alias, or any other descriptive phrase used within a personal name". See <https://www.tei-c.org/release/doc/tei-p5-doc/es/html/ref-addName.html>.

Table 2 Annotation scheme comparison

Named-entity category	Prototypical example	Real example from Spanish medieval corpus	MUC-7 guidelines (Chinchor 1998)
Person	<i>Elizabeth II of England, John Fitzgerald Kennedy</i>	<i>Guayo Julio Cesar (Julius Caesar)</i>	PERSON
Fictional character	<i>Mickey Mouse, Popeye</i>	<i>Celestina</i>	PERSON
Gods and deities	<i>Zeus</i>	<i>Fortuna, Venus</i>	PERSON
Nicknames	<i>Mr. Fix-It</i>	<i>El Cid, El Campeador (The Warrior)</i>	PERSON
Animals (with proper name)	<i>Nessie, Morris the cat</i>	<i>Babieca, Bucefalo (Bucephalus)</i>	PERSON
Organizations, institutions, corporations	<i>Massachusetts Institute of Technology</i>	<i>Orden de Santyago (Order of Santiago)</i>	ORGANIZATION
Cities, towns, regions, countries	<i>New York City, China</i>	<i>Roma, Mesopotamia, Asia</i>	LOCATION
Landforms	<i>Mississippi river</i>	<i>Rio Tormes, los Alpes (Tormes river, the Alps)</i>	LOCATION
Buildings, monuments	<i>Eiffel Tower</i>	<i>Castillo de Ella (Ella castle)</i>	ORGANIZATION
Honorifics	<i>Lord, sir, miss</i>	<i>Don, sennora, maese</i>	No
Group of people regarded as a single unit	<i>The Beatles</i>	<i>Los moros, los santos (the Moors, the Saints)</i>	No
Positions and nobility titles	<i>President of USA</i>	<i>Rey de Castiella (King of Castilla)</i>	No
Family relationships	<i>The lieutenant's daughter</i>	<i>Fioo del Rey Iohan (King Johan's son)</i>	No
Legitimacy expressions	-	<i>Por la gracia de Dios (by the grace of God)</i>	No
Named-entity category	ACE guidelines (Linguistic 2005)	Extended named entity hierarchy (Sekine 2003)	TEI annotation scheme for medieval named entities
Person	Person (PER)	Person	<persName >
Fictional character	Person (PER)	Product: Character	<persName >
Gods and deities	Person (PER)	God	<persName type = }_deity} >
Nicknames	Person (PER)	Person	<persName type = }nickname} > or <addName >
Animals (with proper name)	No	Name_Other	<name >

Table 2 continued

Named-entity category	ACE guidelines (Linguistic 2005)	Extended named entity hierarchy (Sekine 2003)	TEI annotation scheme for medieval named entities
Organizations, institutions, corporations	Organization (ORG)	Organization	<orgName >
Cities, towns, regions, countries	Geo-political Entity (GPE)	Location	<placeName >
Landforms	Location (LOC)	Location: Geological_Region	<geogName >
Buildings, monuments	Facility (FAC)	Facility	<placeName type = }_facility} >
Honorifics	No	Product: Title: Title_Other	<roleName type = }honorific} >
Group of people regarded as a single unit	Geo-political Entity (GPE.PER) or PER.Group	Organization	<orgName type = }_collective} >
Positions and nobility titles	No	Product: Title: Position_Vocation	<roleName >
Family relationships	No	No	<roleName type = }_family} >
Legitimacy expressions	No	No	<addName type = }_legitimacy} >

attribute `type = ``nickname```, which indicates that the entity is not the real name of the person, but an alias.

```
<persName type="nickname">Cid</persName>, en el nuestro mal vos
non ganades nada
[<persName type="nickname">Cid</persName>, you get nothing from
our disgrace]
```

Lone nicknames were annotated differently from additional name components in order to highlight that they are consolidated nicknames that function as official names.

Deities and divine figures (such as gods and saints) will also be considered `<persName>`, but in order to distinguish them from normal person entities they will bear the attribute `type=``_deity``` within the tag.

```
<persName type="_deity">Dios</persName>
[<persName type="_deity">God</persName>]

[<persName type="_deity">Cupido</persName>
<persName type="_deity">Cupid</persName>]
```

Biblical nicknames and other divines aliases (such as *Christ*, *Lord*, *King of Kings*...) are frequent in medieval literature and work *de facto* as indirect names for God and other divine figures. The nature of these names combines notions related to two categories in our annotation scheme, i.e. deity names and nicknames: we formalize deity names with a `type=``_deity``` attribute, whereas we tag nicknames via a `type=``nickname``` attribute. Since biblical nicknames and other divine aliases have a commonality with both deity names and nicknames, we propose an additional type combining both: `type=``_nickname_deity```. This attribute enables the retrieval of all religious nicknames that appear in a given text, which can be of great use when working with biblical texts and other religious sources.

```
Así quiso el <persName type="_nickname_deity">Criador</persName>
[That is how the <persName type="_nickname_deity">Lord</persName>
wanted it]
```

3.2.2 Location names

Geopolitical units (such as countries, cities, towns, villages, regions, etc) will be annotated under the TEI tag `<placeName>`.

```
<placeName>Salamanca</placeName>
<placeName>Castiella</placeName>
```

Facilities such as buildings or monuments that refer to a specific location (castles, monasteries, bridges, etc) should be annotated using the TEI tag `<placeName>` followed by the attribute `type=""_facility""`.

```
<placeName type="_facility">castillo de Ella</placeName>
[<placeName type="_facility">Castle of Ella</placeName>]
```

Landforms and other natural geographical features (such as rivers, mountains, oceans, etc) should be annotated under the TEI tag `<geogName>`.

```
<geogName>río Tormes</geogNames>
[<geogName>Tormes river</geogNames>]
```

This approach is based on the hierarchy defined by Nadeau and Sekine (2007), as well as the TEI guidelines⁴. We are aware of the difficulties that making the difference between human-defined places and geographical accidents poses to automatic processing. However, keeping this difference on the annotation scheme can provide new knowledge to the study of historical place names and on how they came into existence in the first place.

Religious, mythological and fictitious places (such as *Garden of Eden*, *Paradise* or *Styx river*) will also be considered `<placeName>` or `<geogName>` (where appropriate) but will be followed by the attribute `type=""_mythological""`. This attribute facilitates making the difference between locations that can be tracked on a map and those that cannot.

```
<geogName type="_mythological">Laguna Estigia</geogName>
[<geogName type="_mythological">Styx river</geogName>]
```

3.2.3 Organization names

Organizations, associations, institutions and other groups of people (such as religious orders, armies or governmental institutions) should be annotated under the TEI tag `<orgName>`.

```
La <orgName>orden de Calatrava</orgName>
[The <orgName>order of Calatrava</orgName>]
```

```
Los <orgName>frayles de guadalupe</orgName>
[The <orgName>monks of guadalupe</orgName>]
```

```
La <orgName>corona de Castilla</orgName>
The <orgName>crown of Castilla</orgName>
```

⁴ “Geographical name identifies a name associated with some geographical feature such as Windrush Valley or Mount Sinai“. See <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-geogName.html>.

Peoples, communities and any coherent group of people that share a sense of unity or cause like political organizations or ethnic groups (such as *the Greeks, the Trojans, the Moors, the Jews or the Christians*) shall be annotated as `<orgName>`, following (Sekine 2003). This `<orgName>` tag will be followed by the attribute `type = ''_collective''`. Likewise, other groups of people that function as a unit (such as *the Three Wise Men or the Twelve Apostles*) will also be annotated under the `<orgName type = ''_collective''>`.

```
1 <orgName type="_collective">Gentiles</orgName>,
2 <orgName type="_collective">judíos</orgName>,
3 <orgName type="_collective"> cristianos</orgName> y
4 <orgName type="_collective">moros</orgName>,
5 todos en esta concordia están.
```

```
1 <orgName type="_collective">Gentiles</orgName>,
2 <orgName type="_collective">Jews</orgName>,
3 <orgName type="_collective">Christians</orgName> and
4 <orgName type="_collective">Moors</orgName>,
5 they all live in harmony.
```

3.2.4 Role names

Positions, nobility titles and professions (such as *Queen of Castile, Bishop of Oviedo or king* appearing in isolation) will be annotated with the TEI tag `<roleName>`.

```
Por mandato del <roleName>rey</roleName>
[By command of the <roleName>king</roleName>]
```

The `<roleName>` tag can be nested within person names or contain other nested tag (like locations or organizations):]

```
El <roleName>rey de <placeName>Castilla</placeName></roleName>.
[The <roleName>king of <placeName>Castilla</placeName></roleName>.]
```

```
<persName><roleName>Rey</roleName> Alfonso</persName>.
[<persName><roleName>King</roleName> Alfonso</persName>.]
```

Saints, angels, archangels, prophets and other religious roles (which usually precede a person name) should also be annotated as role names:

```
<persName type="_deity"><roleName>san</roleName> Miguel</persName>
[<persName type="_deity"><roleName>saint</roleName> Michael</persName>]
```

```
<persName type="_deity"><roleName>arcángel</roleName> Uriel</persName>
[<persName type="_deity"><roleName>archangel</roleName> Uriel</persName>]
```

Legitimacy expressions and other divine right formalisms (such as *by the Grace of God or por la gracia de Dios*) usually appear next to nobility titles and political positions in medieval documents. In fact, this kind of expressions work as a

formal prelude for certain positions and titles and therefore they can be considered part of the role name itself (as they express the legitimacy that provides the ruler with divine right to hold that title). These expressions should be annotated with `<addName>` along with the attribute `type='_legitimacy'`. The `<addName>` tag should be nested within the appropriate `<rolename>` tag.

```
1 <persName>Alfonso
2 <rolename>
3 <addName type="_legitimacy">por la gracia de
4 <persName type="_deity">Dios</persName>
5 </addName>
6 rey de <placeName>Castiella</placeName>
7 </rolename>
8 </persName>
```

```
1 <persName>Alfonso
2 <rolename>
3 <addName type="_legitimacy">by the grace of
4 <persName type="_deity">God</persName>
5 </addName>
6 king of <placeName>Castiella</placeName>
7 </rolename>
8 </persName>
```

Legitimacy expressions are commonly used in historical and legal texts, and they legitimate the roles exhibited while naming a person. The expressions are not an entity in themselves, but they qualify a name, providing historical legitimization.

Family relationships (such as *mother*, *daughter*, *father*...) should also be annotated with the TEI tag `<rolename>` using the specific attribute `type='_family'`.

Vuestras `<rolename type="_family">hijas</rolename>` infantes aún son.
[Your `<rolename type="_family">daughters</rolename>` are still young girls.]

Mi biuda `<rolename type="_family">madre</rolename>`...
[My `<rolename type="_family">mother</rolename>` a widow...]

`<rolename type="_family">fiio de <persName>Johan</perName></rolename>`
[`<rolename type="_family">son of <persName>Johan</perName></rolename>`]

Honorifics (such as *mister*, *sir*, *lady*, *don*, *señor*, *doña*, etc) should also be considered role names and be annotated as nested tags within the appropriate `<persName>` tag. Honorifics will be annotated as `<rolenames>` with the TEI attribute `type='honorific'` attached.

`<persName><rolename type="honorific">don</rolename> Alfonso</persName>`
[`<persName><rolename type="honorific">lord</rolename> Alfonso</persName>`]

3.2.5 *Miscellanea*

Any other name that should be annotated as a named entity but that does not fit in the previous categories can be tagged under the miscellaneous TEI tag `<name>`. This tag can be used by annotators to annotate any additional entities that could bear historical interest but that do not fit in any of the previous categories.

```
El <name>Evangelio</name>
[The <name>Gospel</name>]
```

```
Cinco <name>Pater nostres</name>, y cinco <name>Aue marias</name>
[Five <name>Our father</name>, and five <name>hail mary</name>]
```

4 Annotation guidelines

In this subsection we describe the guidelines to follow and address the most common unclear cases that can be encountered when applying the previous annotation scheme to a medieval corpus.

4.1 Medieval aliases

As it was mentioned on the previous section, nicknames are so common in medieval texts that they even become the standard person name on official documents; accordingly, nicknames have their own tag format within the annotation scheme. However, only those nicknames that are fully established and work *de facto* as person names should be taken into account. In other words, only nicknames that answer the question *What is the name of this person?* should be annotated. Poetic phrases that refer indirectly to people or characters on literary medieval texts and that are not the established name of the person (such as *the one from Troy* to refer to *Paris*), will not be considered as real person names for our annotation purposes.

4.2 Designators and specifiers

Geopolitical names are frequently preceded by a locative designator or specifier (i.e. structures such as *kingdom of* or *town of*). These designator are usually a part of the official medieval name itself and therefore should be annotated as part of the named entity:

```
<placeName>reyno de Murcia</placeName>
[<placeName>Kingdom of Murcia</placeName>]
```

```
<placeName>ciudad de Toledo</placeName>
[<placeName>Town of Toledo</placeName>]
```

```
<placeName>tierras de Egipto</placeName>
[<placeName>Lands of Egypt</placeName>]
```

4.3 Nested tags

Person names, location names and role names may have other entities within them. For example, role names and person names can have location names within (*Elizabeth of England, Duke of Edinburgh*) or be nested within other named entities. In these cases, we will seize the nesting possibilities of XML-TEI and applied nested tags when necessary. This granularity allows a more complete annotation and therefore ensures a more informative entity retrieval. Some examples of nested tags follow:

```
<persName>Alfonso de <placeName>Castilla</placeName></persName>  
[<persName>Alfonso of <placeName>Castilla</placeName></persName>]
```

```
<persName>Rodrigo Díaz de <placeName>Vivar</placeName></persName>  
[<persName>Rodrigo Díaz of <placeName>Vivar</placeName></persName>]
```

```
<roleName>rey de <placeName>Aragón</placeName></roleName>  
[<roleName>king of <placeName>Aragón</placeName></roleName>]
```

```
<persName>Lázaro de <geogName>Tormes</geogName></persName>  
[<persName>Lázaro of <geogName>Tormes</geogName></persName>]
```

```
Lo firma <persName>Alfonso <roleName type="_family">fiio del  
<persName><roleName>rey</roleName> Johan</perName></roleName></persName>  
[Signed by <persName>Alfonso <roleName type="_family">son of  
<persName><roleName>king</roleName> Johan</perName></roleName></persName>]
```

4.4 Combined named entities (personName+roleName+placeName)

In medieval legal documents, complex named entities that refer to a person may combine first names, nobility titles, positions and kinship (*Alfonso, king of Castilla*) meaning that several `persName`, `roleName` and `placeName` can be found. As a rule of thumb, whenever a person name is explicit (`<persName >`), that will be the main and most external tag, while the rest of tags (`<roleName >`, `<placeName >`, etc) will be nested inside the main `<persName >` tag.

Por mandato del <persName><roleName>rey</roleName> Alfonso</persName>.
 [By command of <persName><roleName>king</roleName> Alfonso</persName>.]

```
1 <persName> <roleName type="honorific">Don</roleName> Odoart
2 <roleName type="_family">fiio primero</roleName> e
3 <roleName>heredero del
4 <persName><roleName>rey</roleName> Henrric de <placeName>Anglatierra</placeName></persName>
5 </roleName>
6 </persName>
```

```
1 <persName> <roleName type="honorific">Don</roleName> Odoart
2 <roleName type="_family">first son</roleName> and
3 <roleName>heir of
4 <persName><roleName>king</roleName> Henrric of <placeName>England</placeName></persName>
5 </roleName>
6 </persName>
```

4.5 Ambiguity between family names and birthplaces

Medieval person names can have a second element that appears after the first name and is preceded by the preposition *de* (“of”, “from”). This second element resembles a surname but could either be a family name (and therefore should simply be annotated as part of the <persName >, as in example 1) or a birthplace working as a patronymic (and therefore should be annotated as <placeName >, as in example 2):

1. <persName>Diego Sánchez de Bustamante</persName>
 <persName>Alfonso de Molina</persName>
2. <persName>Diego Sánchez de <placeName>Bustamante</placeName></persName>
 <persName>Alfonso de <placeName>Molina</placeName></persName>

In order to ensure a precise annotation, the most informative solution for these cases requires looking these names up in historical records and checking the existence of the town or region in question at the time when the character was alive. If no solid information backs up the hypothesis of that name being in fact the birthplace of the person, it should simply be annotated as part of the name (as was done in example 1).

4.6 Kingdoms: place names or surnames?

Kings and queens tend to have particularly long and complex names that combine titles, names and kingdoms, especially on legal documents. This complexity can make the annotation process tricky.

Let’s take the case of *rey Alfonso de Castilla* (*King Alfonso of Castilla*), that is apparently composed of the person name *Alfonso* and the role name *rey de Castilla* (*king of Castilla*). At first sight, annotating *king of Castilla* as an entire <roleName > would seem like the most natural approach. But in that case the annotator would have to choose between two rather unsatisfactory options: either making a discontinuous tag between *rey* (*king*) and *de Castilla* (*of Castilla*) in order

Table 3 Annotation Error Summary for the error type ‘Choice of markable’

Tag	Annotator 1	Annotator 2
roleName	34	71
placeName	7	10
orgName	27	5
name	9	9
persName	27	20
addName	3	1
geogName	3	0
Total	110	116
Percentage from total errors	42%	44%
Percentage from total annotations	5.16%	5.44%

to leave *Alfonso* out of the `<roleName >` tag, or consider everything as a `<roleName >` (including person name *Alfonso*), which would be against the guideline that states that whenever a person name is explicit, `<persName >` should be the main and most external tag (see Sect. 4.4). Our proposed solution for this type of situation is to apply the `<roleName >` tag only to *rey* (*king*), leaving *Castilla* as part of the `<persName >` tag. This decision implies that *Castilla* will not be directly attached to the role name *king*, but would work somehow as a kind of surname.

```
1 <persName>
2 <roleName>Rey</roleName> Alfonso de <placeName>Castilla</placeName>
3 </persName>
```

```
1 <persName>
2 <roleName>King</roleName> Alfonso of <placeName>Castilla</placeName>
3 </persName>
```

At first sight, this workaround can seem quite unorthodox, but it solves several issues:

1. It avoids discontinuous tags.
2. It complies with the guideline that states that whenever an entity refers to a person and the person name is present, `<persName >` should be the main tag (Sect. 4.4).
3. It is consistent with other NER annotating systems (see example *Elizabeth II of the United Kingdom* under Person category in Sekine (2003)).

4.7 Organization names vs place names: metonymies

A common problem that general-domain NER faces is how to deal with metonymy (Poibeau 2006). Metonymy is the linguistic phenomenon in which a part of an object or its location is used to refer to the entire object. A classic metonymy

example is when the name of a country or region is used to refer to its population or government (Markert et al. 2002).

The medieval domain is not an exception to this phenomenon. In medieval documents, metonymy occurs frequently when a place name or facility name (such as a country, a monastery, a city, etc) is used to refer to an institution, in other words, the name of a place represents both a physical location and an administrative entity. When annotating this type of named entities the annotator can be unsure on whether these entities should be annotated literally (i.e. under `<placeName >`) or figuratively (i.e. as `<orgName >`).

For these borderlines cases, the TEI guidelines criteria should be followed, Text (2008): if the physical location corresponds to a district, region, locality or any other geopolitical unit (such as cities, countries, etc) it should be annotated as `<placeName >`. If the named entity is a different type of location (such as a building, institution, etc), the tag `<orgName >` will be preferred.⁵

5 Annotation process and results

The annotation guidelines we have just presented were implemented on a corpus of medieval Spanish. In this section we describe the composition of the annotated corpus and we discuss the obtained results.

5.1 Composition of the corpus

The annotated corpus is a compilation of digitized manuscripts in medieval Spanish edited by the Hispanic Seminary of Medieval Studies⁶ (Jover 2015). The corpus contained 26,200 words and 2,054 named entities. The annotated texts were paleographic editions, which means that the texts reflected the exact writing that appeared on the original text source, including errors, omissions and non-standardized spelling. The manuscripts ranged from the 13th century to the 16th century and included chronicles, legal documents and literary texts (narrative, poetry and theatre).

With regard to the genre distribution of the corpus, the majority of the corpus contained literary texts (70%), including some of the most prominent literary works in Spanish from that time period, such as *El Cantar de Mio Cid*, *Libro de Buen Amor* or *La Celestina*. In this genre, *El Cantar de Mio Cid* deserves a special mention, since a great amount of entities from all categories in our annotation scheme were identified on it. Legal texts represent 20% of the corpus and were particularly rich in complex person names and nobility titles. Historiographical texts (such as chronicles) account for 10% of the corpus and they were crucial to test the annotation of historical people, places, organizations (such as religious orders) and names associated to divinities and saints.

⁵ It should be noted that this guideline is compliant with TEI, but it diverts from ACE guidelines, where names of places are annotated as organizations when they refer to sports team, etc.

⁶ <http://www.hispanicseminary.org/>

5.2 Results and discussion

The corpus was annotated twice by two independent annotators that were trained on linguistic annotation and were provided with the presented annotation guidelines prior to the annotation process⁷. In order to establish the coherence between both annotations, the kappa coefficient was measured (Carletta 1996; Artstein and Poesio 2008). The resulting inter-annotation agreement was 0.802 (N=2054, K=2). This can be seen as an indication of the reliability of the annotation scheme, as Landis and Koch (1977) and Krippendorff (1980) consider a coefficient above 0.8 as indicating reliable annotations.

Regarding the annotation disagreements, 258 issues out of 2,132 annotations were identified (12.10% of all the annotations). The source of disagreement may be classified into three main groups: choice of markable, tag selection and tag swap.

1. **Choice of markable** The choice of markable disagreements are text spans that were tagged by one annotator but not the other. This was by far the most frequent source of disagreement, accounting for 86% of all disagreements (226 out of 258 total disagreements: 110 issues attributable to annotator 1; 116 attributable to annotator 2). These disagreements may reflect different perceptions by annotators of nuances in the texts or in the medieval reality (such as annotating a given location as `placeName` or a given relation or position as `roleName`). See Table 3 for a quantification of the choice of markable disagreements between annotators grouped by tag.
2. **Tag selection** Tag selection disagreements concern spans that were identified as named entities by both annotators but were annotated using different tags. This type of disagreement accounts for 8,5% of all disagreements (22 issues out of 258 total disagreements). Half of these differences were related to the selection of either the `<orgName>` tag or the `<roleName>` tag, specifically when annotating certain religious orders. These errors arose from the subtlety of considering the mention of a religious order as a reference to the members that belong to that religious order (a role) or a reference to the order itself (an organization):


```
1 <roleName> frayles' de la trenjdat</roleName>
2 <orgName> frayles' de la trenjdat</orgName>

1 <roleName> Trinitarian friars</roleName>
2 <orgName> Trinitarian friars</orgName>
```
3. **Tag swap** Tag swap disagreements consists of human errors derived from the application of the annotation scheme to complex entities and consist mostly of nesting errors (see example below). This source of disagreement accounts for 6.5% of all disagreements (17 cases out of 258 total errors).

⁷ Annotation guidelines in Spanish available at <http://postdata.linhd.uned.es/guia-de-anotacion-de-entidades-nombradas-medievales/>

Annotator 1

```
1 <persName> <roleName> Don </roleName>  
2 Alfonso Tellez </persName>
```

Annotator 2

```
1 <roleName> <persName> Don </persName>  
2 Alfonso Tellez </rolesName>
```

6 Conclusions and future work

We have described a NER annotation scheme for medieval texts intended to capture the nature of medieval documents. The proposed category set is an extension of traditional journalistic NER hierarchies and takes into account the key role that nobility titles, positions and family relations played in medieval society (and, consequently, on medieval documents). The proposed annotation scheme tagset is encoded in XML-TEI to ensure interoperability and facilitate its use and exploitation among humanist researchers. Finally, explicit annotation guidelines were introduced and discussed in order to ensure a coherent and reliable annotation and clarify the possible obscure cases that may arise from the special nature of medieval reality.

The presented annotation scheme and guidelines were implemented and tested on a corpus of medieval Spanish, producing an inter-annotator agreement of 0.8 (kappa), which can be considered a solid annotation. Therefore the proposed annotation scheme can be considered a reliable scheme if applied to future NER tasks that involve the annotation of medieval manuscripts and other historical documents.

The enriched annotation we have presented provides valuable source materials for researchers in the Digital Humanities. The annotation of medieval role names is, to the best of our knowledge, a novel contribution from this work that provides new and valuable information in terms of historical contextualization of a document: although role name positions are traditionally not considered named entities in off-the-shelf named-entity hierarchies, the annotation of role names in medieval documents allows to establish the relation between people, places and lineages mentioned in a text, enhancing the contextualization of the document, the identification of relation between entities and the setting of a chronological framework for the text. Thanks to the annotation of role names, the tool HisMeTag (which implements the annotation system we have just presented) can provide such relation graphs by implementing the nested tagging as dependency trees.

In terms of future work, it would be interesting to assess the suitability of the presented scheme on documents from other historical periods of time (from the Renaissance onwards). In addition, an interesting avenue for future work would be

to address the coverage of the presented scheme in other historical named entities and to explore the possibility of subdividing the general name category into more fine-grained categories related to specific subdomains.

From a wider perspective, medieval documents are a rich source of valuable information that await being exploited in terms of natural language processing. Further work is needed towards the textual exploitation of information contained on historical archives, a task that requires the collaboration between both the NLP and Digital Humanities communities.

Acknowledgements The presented work is part of the POSTDATA project, funded by the European Research Council (grant agreement number 679528). This work is a continuation of the Pelagios microgrant that was awarded the Digital Humanities Innovation Lab (LINHD) at UNED University towards the creation of a Medieval Iberian gazetteer.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596.
- Bayerl, P. S., Lüngen, H., Gut, U., Paul, K. I. (2003). Methodology for reliable schema development and evaluation of manual annotations. In *Proceedings of the Workshop on Knowledge Markup and Semantic Annotation at the Second International Conference on Knowledge Capture (K-CAP 2003)*, pp. 17–23.
- Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2), 249–254.
- Chinchor, N. A. (1998). Proceedings of the Seventh Message Understanding Conference (MUC-7) named entity task definition. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, page 21 pages, Fairfax, VA. version 3.5, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.
- Desmet, B., & Hoste, V. (2014). Fine-grained dutch named entity recognition. *Language Resources and Evaluation*, 48(2), 307–343.
- Díez Platas, M.L., Ros Mu noz, S., González-Blanco, E., Ruiz Fabo, P., Álvarez Mellado, E. (2020). Medieval spanish (12th–15th centuries) named entity recognition and attribute annotation system based on contextual information. *Journal of the Association for Information Science and Technology*.
- Díez Platas, M.L., Tobarra, L., Ros Mu noz, S., González-Blanco García, E., Robles-Gómez, A., Caminero, A., Rio Riande, G. d. (2017). Hispanic medieval tagger (hismetag): una aplicación web para el etiquetado de entidades en textos medievales. <http://doi.org/10.5281/zenodo.1123416> [Accessed 29/05/2019].
- Fort, K., Ehrmann, M., Nazarenko, A. (2009). Towards a methodology for named entities annotation. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP '09*, pages 142–145, Stroudsburg, PA, USA. Association for Computational Linguistics.

- Frontini, F., Brando, C., Riguet, M., Jacquot, C., Jolivet, V. (2016). Annotation of toponyms in TEI digital literary editions and linking to the web of data. *MATLIT: Materialidades da Literatura*, 4(2):49–75.
- Grishman, R., Sundheim, B. (1996). Message understanding conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Hovy, E., & Lavid, J. (2010). Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1), 13–36.
- Ide, N., & Romary, L. (2004). International standard for a linguistic annotation framework. *Natural Language Engineering*, 10(3–4), 211–225.
- Isaksen, L., Simon, R., Barker, E.T., de Soto Ca narnares, P. (2014). Pelagios and the emerging graph of ancient world data. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pp. 197–201, New York, NY, USA. ACM.
- Jover, F. G. (2015). La biblioteca digital de textos del español antiguo (bidtea). *Scriptum digital. Revista de Corpus Diacrònics i Edició Digital en Llengües iberoromàniques*, 4, 5–36.
- Krippendorff, K. (1980). *Content analysis: An introduction to methodology*. Beverly Hills, CA: Sage.
- Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Linguistic Data Consortium (2005). *ACE (Automatic Content Extraction) English annotation guidelines for entities*.
- Maraoui, H., Haddar, K., & Romary, L. (2018). Encoding prototype of al-hadith al-shareef in tei. *CoRR*, 782, 3–26.
- Markert, K., Nissim, M., Place, B. (2002). Towards a corpus annotated for metonymies: the case of location names. In *In Proc. of the 3rd International Conference on Language Resources and Evaluation; Las Palmas, Canary Islands*, pp. 1385–1392.
- Murray, J. (2017). Family life in the middle ages. <https://doi.org/10.1093/obo/9780195396584-0236> [Accessed 29/05/2019].
- Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1), 3–26.
- Petterson, E., Megyesi, B., Tiedemann, J. (2013). An smt approach to automatic annotation of historical text. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*, number 087, pp. 54–69. Linköping University Electronic Press.
- Pierazzo, E. (2016). *Digital scholarly editing: Theories, models and methods*. London: Routledge.
- Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies: Morgan & Claypool Publishers.
- Plank, B. (2016). What to do about non-standard (or non-canonical) language in NLP. *CoRR*, abs/1608.07836.
- Poibeau, T. (2006). Dealing with metonymic readings of named entities. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.
- Poibeau, T., Kosseim, L. (2001). Proper name extraction from non-journalistic texts. *Computational Linguistics in the Netherlands*, 144–157.
- Pustejovsky, J., Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. “ O'Reilly Media, Inc.”.
- Sánchez-Marco, C., Boleda, G., Padró, L. (2011). Extending the tool, or how to annotate historical language varieties. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCh '11, pp. 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Sekine, S. (2003). *Sekine's Extended Named Entity Hierarchy*. Retrieved April 14, 2018, from <http://nlp.cs.nyu.edu/ene/>.
- Sekine, S., Sudo, K., Nobata, C. (2002). Extended named entity hierarchy. In *Proceedings of Third International Conference on Language Resources and Evaluation (LREC-2002); Las Palmas, Canary Islands*.
- Stede, M., & Huang, C.-R. (2012). Inter-operability and reusability: The science of annotation. *Language Resources and Evaluation*, 46(1), 91–94.
- Text Encoding Initiative Consortium. (2008). *TEI P5: Guidelines for electronic text encoding and interchange*. Retrieved April 14, 2018, from <http://www.tei-c.org/Guidelines/P5/>.

Tjong Kim Sang, E. F., De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 142–147. Association for Computational Linguistics.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.