Enriching Portuguese Medieval Texts with Named Entity Recognition

Maria Inês Bico, Centro de Linguística, Faculdade de Letras, Universidade de Lisboa, mariainesb1@edu.ulisboa.pt

Jorge Baptista, Faculdade de Ciências Humanas e Sociais, Universidade do Algarve

Fernando Batista, ISCTE, Instituto Universitário de Lisboa

Esperança Cardeira, Centro de Linguística, Faculdade de Letras, Universidade de Lisboa

Abstract

Historical data poses unique challenges to Natural Language Processing (NLP) and Information Retrieval (IR) tools, including digitization errors, lack of annotated data, and diachronic-specific issues. However, the increasing recognition of the value in historical documents has promoted efforts to semantically enrich and optimize their analysis. This paper contributes to this endeavour by enriching the *Corpus de Textos Antigos* through NLP tools and techniques to enhance its usability and support research. The corpus undergoes linguistic annotation, including part-of-speech tagging, lemma annotation, and Named Entity Recognition (NER). Subsequently, the paper delves into the tasks of entity disambiguation and entity linking, which involve identifying and disambiguating named entities by referring to a Knowledge Base. Addressing the challenges posed by factors such as text state, epoch, and the chosen Knowledge Base, the paper presents insights into related work, annotation results, and the linguistic interest of a Medieval annotated corpus for Named Entities. It concludes by

discussing the challenges and providing avenues for future research in this domain.

Keywords: Corpus Analysis, Named Entity Disambiguation, Named Entity Linking, Natural Language Processing, Information Retrieval, Portuguese Medieval Texts.

1. Introduction

Historical textual data has long been relegated to a secondary position within the realms of Natural Language Processing (NLP) and Information Retrieval (IR). It encompasses nearly all the challenges encountered by contemporary data, compounded by the complexities introduced through digitization methods (e.g., OCR errors), an overall scarcity of adequately annotated data, and unique diachronic-specific hurdles (e.g., variations in spelling and grammar), among others. Nevertheless, there is currently a heightened recognition of the value inherent in historical documents. Consequently, concerted efforts have been undertaken over the years to enrich historical textual data and optimize its analysis.

This study aligns with the concerted effort aimed at enhancing the *Corpus de Textos Antigos*[1] by employing a range of NLP tools and techniques to augment its utility and facilitate research endeavours reliant on this repository. In the initial stage, as detailed in Bico *et al*., a portion of the corpus underwent linguistic annotation encompassing part-of-speech (PoS), lemma, and grammatical inflectional values.[2] Subsequently, employing supervised machine learning techniques, automatic annotation of the corpus's texts was executed, assessed, and subsequently subjected to manual correction.

Despite the fact that PoS tagging has not been completed for the entire corpus, subsequent efforts have been initiated to perform Named Entity Recognition and classification (NER, for short), necessitating the manual annotation of named entities (NE) within the corpus. Following the NER task, the subsequent steps of named entity disambiguation (NED) and linking (NEL) naturally ensue. Named Entity Linking constitutes an NLP task that strives to

establish referential connections for NE mentions within an input text, thereby selecting the appropriate entity from a diverse set of possibilities within a Knowledge Base (KB). The NEL task involves annotating NE mentions with the corresponding URI (Uniform Resource Identifier) of the accurate entry, effectively disambiguating it from other mentions. This task presents significant challenges, influenced by various factors (e.g., textual context, timeframe, chosen KB), and warrants multiple approaches for resolution.

The structure of this paper unfolds as follows: Section 2 provides a succinct overview of pertinent related work; Section 3 presents the progress achieved thus far, encompassing the completed work (Section 3.1) and the initial strides taken towards the recognition and classification of NE (Section 3.2); Section 4 showcases the two annotated texts specifically addressing NE, and sheds light on the outcomes of the annotation process; Section 5 elucidates the linguistic significance inherent in a Medieval annotated corpus for Named Entities; Section 6 delves into the complexities associated with NED and NEL; finally, Section 7 concludes with prospective avenues for future research.

2. Related work

Numerous works of relevance can be identified in relation to the present study. This section offers a concise selection of these works.

Brando *et al*. developed a graph-based system dedicated to the task of NEL.[3] This system, known as REDEN (*Referencement et Desambiguisation d'Entités Nommées* or Disambiguation and Reference of Named Entities), effectively performs the two essential aspects of NEL: candidate retrieval and

candidate selection. Leveraging the potential of Linked Data (LD), REDEN traverses Resource Description Framework (RDF) graphs derived from various LD collections, encompassing both broad-coverage resources such as DBpedia and domain-specific datasets like the *Bibliothèque Nationale de France*, seamlessly merging them to construct a non-redundant graph. The system was successfully employed in the analysis of French literary and criticism essays from the 19th and 20th centuries, yielding satisfactory results. REDEN, an open-source resource, aligns with prevailing standards for digital editions (Text Encoding Initiative – TEI) and the semantic web (RDF).

Recognizing the inherent difficulties in locating KBs with sufficient coverage for a cultural heritage resource, Munnelly *et al*. embarked on the endeavour of constructing their own KB specifically tailored for studying the Irish 1641 rebellion depositions.[4] The 17th-century manuscripts presented numerous challenges, including irregular spellings and an extensive array of entities. To establish a solid foundation for the KB, the project merged three distinct primary sources, culminating in a comprehensive inventory of unique entities. Moreover, the project seeks to incorporate temporal information into the KB to enhance filtering capabilities. Although the new custom KB has yet to be integrated with existing semantic web resources, it is anticipated that the entities within the KB will eventually be linked to corresponding entities within widely recognized KBs such as Geonames or DBpedia.

Linhares Pontes *et al*. devised a comprehensive multilingual end-to-end entity linking system, focusing on two historical corpora (CLEF HIPE 2020 and NewsEye), effectively disambiguating entities across English, Finnish, French, German, and Swedish texts.[5] By leveraging entity embeddings derived from

Wikipedia in multiple languages and employing various techniques including match correction and filtering, the system successfully mitigated challenges typically encountered in historical data. Remarkably, the system surpassed the baseline performance for the corpora across all tested languages.

All these different efforts for matching a named entity to a KB entry provide clues for our future work and the challenges the CTA has. Given the particularities of CTA, as will be mentioned, a multilingual approach and/or a custom-built KB might be the solution for a good coverage of Named Entities.

Among corpora dealing with Historical Portuguese, the Tycho-Brahe Corpus (TBC)[6] and Colonia[7] should be mentioned. The TBC is an electronic morphologically and syntactically annotated corpus with 95 texts spanning from the 14th to the 20th century by authors from Portugal, Brazil and African countries. Its main goal is to potentiate the research on the syntactic evolution of the Portuguese Language. The TBC uses *eDictor*, a tool for the digital edition and POS tagging task.[8] Colonia is a corpus comprising more than 5 million tokens form 100 texts. The texts range from the 16th to the 20th century, with each century making a subcorpus. The texts were annotated using TreeTagger with its parameter for Portuguese. Some level of post-processing was done on the texts. To allow search queries, the Corpus WorkBench and Corpus Query Processor were used. The corpus is available to be downloaded or explored through a user-friendly interface.[9]

3. Adopted Approach Detecting Named Entities

3.1. Corpus Processing Pipeline

The *Corpus de Textos Antigos* (CTA) was created by the Centre of Linguistics of the University of Lisbon (CLUL) in 2015. Adopting a standard of high fidelity to the source material, this project has 32 diplomatic editions of 27 different texts (as of March 2023), built within the TEITOK framework.[10] As the corpus was bare of any linguistic annotation, an effort was made to supply it with morphosyntactic information. As of June 2023, the corpus has 6 texts with approximately half a million tokens annotated for PoS, lemma, and grammatical values (inflection). An automatic annotation task was performed using the TreeTagger[11] on two different experiments with different sets of texts, as described in Bico *et al*.[12] Figure 1 shows the general pipeline for this process.

[insert Figure 1 near here]

Figure 1. Named Entity processing pipeline (adapted from Bico et al., 2022).

First, one of the largest manuscripts of the corpus, the *Horto do Esposo* (HdE), dated from 1390-1437), with approximately 155K tokens, and 138K (14,3K different) words was manually annotated by 3 linguists. This was used to build the first Tagger Model 1, which was then used to automatically tag, in a two-step process, a set of other texts. Precision varied from 64.4 to 91%, depending on different factors. The best results were achieved with another witness of the text *Horto do Esposo* (HdE-DCE), so that both texts had thus the same lexicon and the same syntactical structures. Despite the overall satisfying results, expanding

the training dataset can still improve these results. All these new texts were than manually revised, corrected and annotated. Then, the second Tagger Model 2 was trained with the substantially enlarged, PoS-annotated corpus.

For this paper, another text from the CTA, the *Livro dos Mártires* (LdM), was PoS-tagged with Tagger Model 2, and then the tags were revised and annotated. The new text was also added to previous set of PoS-tagged texts from CTA, totalling now approximately half a million tokens.

3.2. NE identification and delimitation Guidelines

Following the experiments on automatic lemmatization and PoS tagging, a portion of the corpus underwent annotation to identify its NE. Two annotators manually annotated two texts, adhering to the specific PoS tag for proper nouns: N+Npr. The chosen annotation scheme followed the BIO format: B denoting the beginning of a named entity, I indicating a token inside the named entity, and O representing other tokens. The annotation process encompassed three distinct entity types: PERSON, signifying individuals' names (e.g., *João*, 'John'); PLACE, representing location names (e.g., *Lisboa*, 'Lisbon'); and PERSON:ORG, a sub-type of person that encompasses collective bodies of people (e.g., *filhas de Sião*, 'daughters of Syon') or organizations (e.g., *a Igreja*, 'the Church').

The determination of NEs boundaries adhered to the guidelines initially proposed by the HAREM initiative.[13] In this way, a NE is required to encompass a proper noun, with exceptions being made for *Deus* 'God' and *Igreja* ('Church'), which were initially annotated as common nouns. All entities adopt a flat representation, without the inclusion of NEs within larger NEs. Appositions, epithets, forms of address, and punctuation tokens found within the NE are

included within its boundaries, ensuring the maximal extension of each NE. This approach acknowledges the significance of contextual information for the subsequent task of disambiguating NE referents. For instance, examples such as *Carlos, rei de França* 'Charles, King of France' and *Carlos, rei de Inglaterra* 'Charles, King of England' illustrate the importance of considering contextual details.

The Named Entity annotation was conducted on two extensively annotated texts, namely, the manuscript A of *Horto do Esposo* (HdE-A) and a newly annotated text specifically for this study, the *Livro dos Mártires* (LdM), comprising more than 150,000 and 250,000 tokens, respectively. This approach ensured a broader coverage of NEs. Subsequently, a concise description of these two texts is presented.

Horto do Esposo, an original Portuguese text, is believed to have been authored in the late 14th century, specifically between 1390 and 1437. The author drew inspiration from the medieval Latin sources that were prevalent during that era. Within this text, the author employs the genre of exempla to convey moral messages through concise and memorable sentences.[14] At present, the CTA comprises two editions of the text: ms. A, originating from the scriptorium of the Monastery of Alcobaça (HdE-A); and fragments D, C, and E (HdE-DCE) from the Monastery of Lorvão. Both HdE-A and HdE-DCE have undergone lemmatization and annotation for part-of-speech tags and inflection. Additionally, HdE-A has been annotated for Named Entities.

The *Livro dos Mártires* was authored by Bernardo de Briuega in the 13th century, originally written in Castilian. This work constitutes the third part of a larger five-book project commissioned by King Alfonso X of Castille. Primarily a

hagiographic book, it offers detailed accounts of the lives of saints and martyrs. The content predominantly focuses on Roman martyrs and the Desert Fathers, providing valuable insights into their experience.[15] During the reign of King Denis of Portugal (1279-1325), the text underwent translation into Portuguese. While no extant manuscript from this period has been discovered, this conclusion is drawn based on the identification of archaic 14th century Portuguese forms present in the 1513 edition.[16]

The 1513 edition, by João Pedro Bonhominy de Cremona is the only surviving exemplar of the text. It is annotated with lemma, PoS tag, and for its Named Entities.

4. Annotation Results

4.1. Overall statistics

The annotation endeavour aimed at identifying NEs within the corpus yielded a combined total count of 12,590 occurrences. Table 1 provides a comprehensive breakdown of the types of entities observed across the texts.

Table 1. Entity Types in HdE-A and LdM

	HdE-A		**LdM**		**Combined Total**	
NE Type	**Total**	**%**	**Total**	**%**	**Total**	**%**
PERSON	2,785	84.44	8,507	90.11	11,292	86.69
PLACE	319	10.13	920	9.74	1,239	9.84
PERSON:ORG	45	1.43	14	0.15	59	0.47

Total	3,149	9,441	12,590

The higher number of named entities recorded in LdM is not only due to the text's length but also and mainly because of its genre. Its purpose is described in its opening sentence: *Este he o liuro e legẽda que fala de todolos os feytos e paixoões dos sãtos martires em lingoagem portugues* ('This is the book and legend that speaks of all deeds and passions of the holy martyrs in Portuguese language'). More than a hundred lives of martyrs are described therein and within each one a diverse set of entities are mentioned. While HdE-A shares a similar length with other large texts, it distinguishes itself in terms of genre. As a spiritual and didactic book, its content exhibits a lower frequency and narrower range of references to individuals and locations.

4.2. Part-of-Speech Patterns

For a better comprehension of the corpus and entity referencing, an exploration of annotated entity patterns was conducted. These patterns encompassed the PoS tags and grammatical values assigned to each individual token or group of tokens. Overall, 92 patterns were identified in HdE-A, while a remarkable 508 patterns were recognized in LdM. This notable discrepancy in LdM can be attributed to both the larger number of identified entities and the distinctive writing style employed by the author or translator.

In LdM, there is a frequent occurrence of relation-denoting elements when referring to entities; e.g., *padre* 'father': *são Vidal* padre *de são Gervásio e de são Protásio*, 'Saint Vidal, *father* of saint Gervasius and saint Protasius'.

Moreover, the author or translator of LdM employs various forms to refer to the same entity.

This phenomenon is particularly evident in references to *Jesus* or *Deus* 'God'. In the case of the entity *Jesus Cristo* 'Jesus Christ' alone, LdM presents over 60 distinct forms of reference (e.g., *meu senhor Jesus Cristo Deus que nasceu de Deus* 'my lord Jesus Christ God that was borne out of God'). Furthermore, LdM contains the longest sequences of tokens identified as a single named entity. This consists of 24 tokens:

> *Senhor deos saluador . deos dos anjos . deos de todalas cousas que som pelos logares . pastor dos homẽs . comer dos doentes*, ‘Lord God savior, god of the angels, god of all things everywhere, shepherd of man, food of the sick’.

While the average length of entities in both texts is relatively close to one another (HdE-A: 1,60 tokens; LdM: 2,25 tokens), the median length shows better the difference in length: HdE-A has a median length of 4,5 tokens, whereas LdM almost doubles this value, with a median length of 9,5 tokens. Table 2 presents a comparison of the occurrence of PoS patterns found in each text, providing a distinction based on entity type, and presenting the combined total for both texts. Below each pattern, an example is provided.

Table 2. Most common PoS patterns for NEs in HdE-A and LdM and the corresponding NE types

rank	Pattern: sequence of PoS tags	HdE-A	LdM	Combined Total
1	**A:fs N+Npr:fs** *Santa Teresa* 'Saint Teresa'	-	252 PER	252 PER
2	**A:ms N+Npr:ms** *São João* 'Saint John'	526 PER	2,462 PER	2,988 PER
3	**A:ms N+Npr:ms DET+art+def:ms N:ms** *São Pedro o apóstolo* 'Saint Peter the apostle'	-	103 PER	103 PER
4	**ADV V:Kms A:ms N+Npr:ms** *Bem-aventurado são Paulo* 'Blessed saint Paul'	-	278 PER	278 PER
5	**N:ms** *Deus* 'God'	466 PER	538 PER	1,004 PER
6	**N:ms N:ms** *Deus padre* 'God father'	190 PER	-	190 PER
7	**N:ms N+Npr:ms** *Profeta Jeremias* 'Prophet Jeremias'	186 PER 4 PLA	133 PER 7 PLA	319 PER 11 PLA
8	**N+Npr:fs**	-	242 PER	242 PER

	Paulina ‘Paulina’		401 PLA 3 PER:ORG	401 PLA 3 PER:ORG
9	**N+Npr:ms** *Gregório* ‘Gregory’	845 PER 46 PLA	2,036 PER 79 PLA 1 PER:ORG	2,881 PER 125 PLA 1 PER:ORG
10	**N+Npr:ms DET+art+def:ms N:ms** *Nero o imperador ‘Nero the emperor’*	-	223 PER	223 PER
11	**N+Npr:ms N+Npr:ms** *Jesus Cristo ‘Jesus Christ’*	271 PER	325 PER	596 PER
12	**N+Npr:ms SENT DET+art+def:ms N:ms** *Adriano . O imperador ‘Adriano . The emperor’*	-	155 PER	155 PER
13	**PRO+pos:1psm N:ms N+Npr:ms N+Npr:ms** *Nosso Senhor Jesus Cristo ‘Our Lord Jesus Christ’*	-	409 PER	409 PER
Total		2,484 PER	7,156 PER	9,640 PER

	52 PLA	480 PLA 4 PER:ORG	532 PLA 4 PER:ORG

As anticipated, people's names constitute the majority of occurrences in both texts, accounting for 94.73 per cent of the combined total of entities. Regarding grammatical gender, only a small portion of entities in this selection are identified as feminine (9.28 per cent), with nearly half of them being place names (e.g., *Pérsia* 'Persia', *Índia* 'India', *Gália* 'Gaul').

Patterns 2, 5, 7, 9 and 11 occur in both HdE-A and LdM. These are the less complex patterns. Entities are mentioned just by a single proper/common name (N+Npr:ms, e.g., *Pedro* 'Peter'; N:ms, e.g. *Deus* 'God'), two names (N+Npr:ms N+Npr:ms, e.g., *Jesus Cristo* 'Jesus Christ'), or by the combination of an adjective or a noun with the proper noun (A:ms N+Npr:ms, e.g. *são Pedro* 'saint Peter'; N:ms N+Npr:ms, e.g. *imperador Nero* 'emperor Nero'). It should be mentioned that all NE tagged with N:ms (pattern 5) are of *Deus* 'God'. In HdE, pattern 6 is also specific to references to the deity (e.g., *deus padre*, 'God father'). Additionally, out of all entities referred to by two names, the name of *Jesus* is the most repeated, making for 89.60 per cent of the combined total of entities. Other entities referred by their two names are roman writer *Valerius Maximus*, historian *Pompey Trogue*, roman emperors *Antoninus Pius* and *Marcus Aurelius*, and roman governor *Pontius Pilate*, among few others. The most common adjective preceding a proper noun is *santo* 'saint' in the two Portuguese forms (*são*, *santo*), though LdM also registers the single-time occurrence of adjectives *doce* 'sweet',

grande ‘great’ and *pecador* ‘sinner’. Pattern 13 always refers to *Jesus Christ* and the respective form of address (e.g., *nosso senhor Jesus* Cristo, ‘our lord Jesus Christ’).

Patterns 3, 4, 7 and 9 are exclusive to LdM in this selection of patterns with +100 occurrences. In pattern 4, the classifier *são/santo* is preceded by the compound adjective *bem-aventurado* ‘blessed’, e.g., *bem-aventurado são Gregório* ‘blessed saint Gregory’. Patterns 3, 7 and 9 are variations of one another. In pattern 3, the simplest, an apposition is added to the pair classifier+noun (e.g., *são Pedro o apóstolo*, ‘saint Peter the apostle’); patterns 7 and 9 do not exhibit the adjective ‘saint’ (e.g., *Adriano o imperador,* ‘Adrian the emperor’); and pattern 9 adds a punctuation mark (e.g., *António . o imperador,* ‘Anthony . the emperor’), where the use of '.' can be interpreted as the graphic representation of the syntactic function of the apposition.

5. Discussion and major Implications

The recognition and classification of named entities in CTA extend beyond the mere identification and linking of entities. This section addresses the guidelines employed for the lemmatization of proper nouns, the challenges encountered during the lemmatization process, and the linguistic significance of the corpus through the presentation of various case studies.

5.1. Lemmatization guidelines and corresponding challenges

As mentioned earlier, HdE-A underwent manual annotation with lemma and PoS tag, while LdM underwent automatic annotation using TreeTagger in the second

experiment. Guidelines for lemmatization were established for both the initial manual annotation and the revision of automatically obtained data to ensure data coherence. Regarding proper nouns, a specific guideline was applied: all proper nouns were annotated with their modern lemma. This guideline encompassed modernized spellings (e.g., *Theofilo* as *Teófilo*, the regular modern word) as well as older forms of names that have fallen out of use (e.g., *Jesus* remains as *Jesus*; *Mafomede* becomes *Maomé* 'Mohammed'; *Bonifaz* corresponds to *Bonifácio* 'Boniface'; and *Cibrão* and other multiple graphic variations are annotated as *Cipriano* 'Ciprian').

The decision to lemmatize older forms under a modern lemma does not undermine their study or eliminate these forms. It was implemented to enhance the usability of search queries, enabling researchers and users to input the modern lemma and effortlessly access forms that may not have been considered before. Since these variants of the name are no longer prevalent in person (anthroponomy) or place (toponomy) naming, conducting a search using the modern lemma streamlines the process of locating older name variations within the corpus.

5.2. Linguistic relevance and case studies

During the late 14th century, Portuguese politics and power underwent significant transformations, leaving a lasting impact on the culture and language of the country. It was during this period that the Portuguese language solidified several linguistic changes that had already been set in motion, shedding Old Portuguese characteristics while introducing new changes.[17] One such novelty was the

process of re-Latinization, whereby the vocabulary was enriched through the incorporation of Latinisms.

Due to this process, during the transitional period spanning the 14th to the 16th century, two variant forms of the same name were used. Take the Latin etymons ADRIANUS or JULIANUS: in Old Portuguese these etymons would have evolved to *Adrião* and *Julião*, through a series of different linguistic phenomena. With the reintroduction of Latinisms into the vocabulary, forms such as *Adriano* or *Juliano* started to occur more frequently. These new forms are closer to the Latin etymon and did not undergo as many linguistic changes as the older forms. Therefore, during this period there was a variant that had come from popular tradition (*Adrião*/*Julião*), and other which was introduced by erudite tradition (*Adriano/Juliano*). Both pairs of words are still used and recognised today in anthroponomy and toponomy (e.g., *Póvoa de Santo Adrião* and *São Julião do Tojal*, two villages in the outskirts of Lisbon; *Adriano Moreira*, a Portuguese politician; *rei Juliano*, the translated name of the famous lemur king from the *Madagascar* movies). Recognizing this distinction, it has been deemed appropriate to assign two distinct lemmas for each pair, differentiating between the complete form and the abbreviated form.

A similar process happened to the etymon CIPRIANUS: in Old Portuguese, this etymon evolved to *Cibrião* or *Ciprião*; later, the name was reintroduced in the language as *Cipriano*. However, unlike the pairs *Adrião/Adriano* and *Julião/Juliano*, the older forms of *Cipriano* have fallen into disuse and are no longer recognised in onomastics. As the older, shorter forms are no longer used in Modern Portuguese, both complete and shorter forms are lemmatized with the modern lemma.

Regarding the medieval texts HdE-A and LdM, they provide evidence of the concurrent use of these forms. The breakdown of the relevant data is presented in Table 3.

Table 3. Occurrences of long and short forms of the pairs Adrião/Adriano, Julião/Juliano, Cibrião/Cipriano

	HdE-A	**LdM**
Adrião	0	12
Adriano	0	15
Julião	0	26
Juliano	14	36
Cibrião	0	66
Cipriano	3	1

It is immediately evident that both texts exhibit a preference for using the complete form of names, except for the pair *Cibrião*/*Cipriano*, where the shorter version is favored in LdM. This preference can likely be attributed to the translation of the text from the original Castilian into Portuguese. In fact, among the ten different graphic forms for *Cibrião*, the occurrences of *çebriã* (44), *cebriam* (10), *çebreã* (2), and *cebrian* (1) can be seen as influenced by the Castilian text and the name *Cébrián*.

The evolution of the name *João* 'John' can be observed in the data from HdE-A and LdM. Derived from the Latin etymon JOHANNES, during the 14th and 16th centuries, the name exhibited two variations: *Johane* and *Joham*. The latter

was primarily used when preceding names starting with a consonant, e.g., *Johan Baptista* 'John the Baptist'.[18] It is from this form that the modern *João* 'John' originated. Both *Johane* and *Joham* coexisted for a period until the latter gradually supplanted the former. In HdE-A and LdM, this shift can already be observed. Table 4 provides a breakdown of all occurrences of these two forms, clearly indicating a preference for *Joham* before names not starting with a consonant.

Table 4. Occurrences of forms Johane and Joham in HdE-A and LdM

	HdE-A	**LdM**
Johane	6 *Papa Johãne*, *irmão Johãne*, *Johãne*	1 *Sã Johãne o preste*
Joham	38 *Sam Joham/Johãm/Ioham, Johãm Damaceno*, *sam Ioham de Leteram*, *sam Joham de Nápoles*	37 *Joham Gerson*, *Johã Pedro Bonhominy, sam Joham de Letrã, Johã/Joham/Johan, Joam/Joã*

The linguistic relevance of a Medieval annotated corpus for Named Entities becomes evident. Such a corpus not only enables the tracing of the evolution of specific proper names, like *João* 'John', but also serves to illustrate a significant characteristic of the Middle Portuguese phase: the transition between older forms, such as *Julião*, and modern forms, like *Juliano*. Additionally, while etymological dictionaries typically document the term *ad quo*, or the origin of a name's usage, acquiring knowledge of its term *ad quem*, or the endpoint of

its development, proves highly valuable in constructing a timeline for accurately dating a given toponym. For instance, understanding that the toponym *Póvoa de Santo Adrião* reflects the anthroponym *Adrião* rather than *Adriano* provides crucial insights into the historical dating of the toponym.

6. Towards Disambiguation: challenges and possible limitations

Manual entity disambiguation and entity linking, although feasible for smaller texts with limited entities, become impracticable when dealing with large volumes of text. The manual verification and disambiguation of entities, coupled with the need to consider contextual information and search for the correct entity, would constitute an arduous and time-consuming endeavour.

The automatic approach is deemed more advisable, although certain precautions need to be taken. As previously mentioned, the availability of specific resources for historical data is limited, and smaller or less studied languages face additional challenges. Regarding older stages of the Portuguese language, existing corpora either lack named entity annotations or do not align with the period of our project. For instance, the *Post Scriptum* project (2014), represents a linguistically annotated corpus at the morphosyntactic and syntactic levels, albeit with minimal annotation and identification of individuals' names. However, this corpus primarily focuses on private letters written in Portugal or Spain during the Early Modern Ages.[19]

Part of the Named Entity Linking task involves the crucial process of mapping each entity to a corresponding entry within a KB, following an initial automated selection of potential candidates. The CTA corpus, by virtue of its inherent characteristics, encompasses entities that not only pertain to specific

temporal epochs (which may not necessarily coincide with the date of the text) and geographic regions but also exhibit close associations with the historical trajectory of the Church.

It is evident that certain named entities possess an unambiguous and readily identifiable referent (e.g., *Bede*, *Eve*, *Jesus Christ*). However, challenges arise when attempting to identify other entities due to factors such as graphical variations (e.g., *Çerimõ* corresponds to *Queremão* 'Chaeremon'), the prevalence of a name within specific contextual frameworks (e.g., *Papa Clemente* 'Pope Clement' - with 14 popes bearing this name, which one is being referred to?), or even their relative obscurity in the eyes of contemporary readers and researchers (e.g., the identity of *Chudião* 'Chudion').

To achieve optimal entity linking results, the selection of an appropriate knowledge base is of utmost importance. Recent studies have presented various options, including using a single KB, integrating multiple KBs, or developing a custom ontology specifically tailored for the project, which can later be linked to more comprehensive KBs. Regardless of the chosen approach, numerous scenarios can arise when dealing with lesser-known or ambiguous entities.

Firstly, there may be cases where the entity has no reference in any knowledge base, e.g., *Erão*, ‘Eron’, resulting in a NIL result from the system. Secondly, the KB might have limited coverage on certain topics, lacking sufficient information to accurately identify the correct entity from a pool of potential candidates. For instance, in LdM, the mention of *Papa Celestino* ‘Pope Celestine’ lacks additional details, making it unclear which of the five Popes named *Celestine* is being referred to. To disambiguate the reference, it becomes necessary to examine the surrounding text and consider the context:

P*or ende o papa Celestino renunciou e leixou o papado. E este Celestino antes que fosse papa havia nome Pedro e vivia no ermo santamente*

'Therefore Pope Celestine renounced and left the papacy. And this Celestine, before he was Pope, had the name of Peter and lived devoutly in the wilderness'.

Thirdly, the system's performance can be affected by language dependencies, particularly when processing and identifying Portuguese names within resources of different languages. Users should be aware that certain information may only be available in specific language versions of Wikipedia, as exemplified by Bernardo de Brihuega, the author of LdM, who is exclusively documented in the Spanish version of Wikipedia.

Recent research has addressed these challenges through various approaches. For instance, Linhares Pontes *et al.* have developed a multilingual end-to-end named entity linking system that employs techniques such as match correction, entity embeddings, and filters to tackle these challenges.[20] Mossalan *et al*. chose a monolingual approach, linking French entities to a French knowledge base. [21] Agarwal *et al.* developed a system that incorporates temporal information to aid in entity disambiguation.[22] Brando *et al.* designed a graph-based entity linking system that integrates multiple knowledge bases to create a unified list of entities.[23]

Given the similar challenges encountered in texts from CTA and the aforementioned projects, we are eager to explore the development or adaptation of a system tailored to address the disambiguation and linking of entities in Old and Middle Portuguese texts, considering their unique characteristics and

challenges. Our future work will incorporate considerations such as multilingualism, spelling variations, and temporal context.

7. Conclusions and Future Work

This paper presents our endeavour to enhance the *Corpus de Textos Antigos* through the application of linguistic annotation and Named Entity Recognition techniques in Portuguese Medieval texts. The corpus underwent extensive annotation, encompassing crucial linguistic features such as part-of-speech tagging, lemmatization, and manual identification of Named Entities. The annotation process yielded a comprehensive collection of 12,590 NEs, comprising person names, place names, and person:organization entities.

The annotated corpus offers valuable resources for the examination and analysis of Portuguese Medieval texts, while also serving as a foundation for the development of specialized NLP and IR tools tailored to historical data. Identifying named entities in historical texts presents distinct challenges arising from digitization errors, limited annotated data, and diachronic-specific factors such as spelling and grammatical variations. By tackling these obstacles and delivering a linguistically annotated corpus, this study contributes to the semantic enrichment and enhancement of historical document analysis.

Future endeavours in this domain can direct attention towards multiple aspects. Primarily, an extension of the linguistic annotation across a broader range of texts is warranted to augment the training dataset, thereby enhancing the accuracy and precision of NER. Furthermore, the exploration of diverse approaches and techniques for NED and NEL can be pursued, aiming to bolster the effectiveness of these endeavours. Such pursuits may involve the

incorporation of external knowledge bases with comprehensive coverage of cultural heritage resources, or the creation of customized knowledge bases tailored specifically to Medieval Portuguese texts.

Moreover, the corpus can be utilized for diverse research objectives, including examining the linguistic characteristics and patterns of NEs in Portuguese Medieval texts, exploring the historical and cultural context embodied in these entities, and investigating the interconnections and associations among entities within and across texts. Such efforts can foster a more profound comprehension of the language and society during the Medieval period.

The work presented in this paper serves as an initial step towards advancing research and development in the domain of NLP and IR as applied to historical documents. By tackling the unique challenges inherent in Portuguese Medieval texts, including entity disambiguation and linking, this study enables the analysis and exploration of these invaluable historical resources. Consequently, it opens up avenues for fresh insights and discoveries in the fields of historical linguistics and cultural studies.

Endnotes

---

[1] CLUL (ed.) *Corpus de Textos Antigos*. (2015). URL: http://teitok.clul.ul.pt/postscriptum/index.php?action=home (last access: 10/11/2023)

[2] M. I. Bico, J. Baptista, F. Batista, and E. Cardeira, 'Early experiments on automatic annotation of Portuguese medieval texts', in G. Silvello, O. Corcho, P. Manghi, G. M. Di Nunzio, K. Golub, N. Ferro, and A. Poggi, eds., *Linking Theory and Practice of Digital Libraries* (2022), 442–49. https://doi.org/10.1007/978-3-031-16802-4_44.

[3] C. Brando, F. Frontini, and J. Ganascia, 'REDEN: named entity linking in digital literary editions using linked data sets', *Complex Syst. Informatics Model. Q. 7* (2016), 60–80.

[4] G. Munnelly, H. J. Pandit, and S. Lawless, 'Exploring linked data for the automatic enrichment of historical archives', in A. Gangemi, A. L. Gentile, A. G. Nuzzolese, S. Rudolph, M. Maleshkova, H. Paulheim, J. Z. Pan, and M. Alam, eds., *The Semantic Web: ESWC 2018 Satellite Events*, (2018), 423–33.

[5] E. Linhares Pontes, L. A. Cabrera-Diego, J. G. Moreno, E. Boros, A. Hamdi, N. Sidère, M. Coustaty, and A. Doucet, 'Entity linking for historical documents: challenges and solutions', in E. Ishita, N. Lee San Pang, and L. Zhou, eds., *Digital Libraries at Times of Massive Societal Transition*, (2020) 215–31.

[6] C. Galves, A. L. Andrade, and P. Faria, *Tycho Brahe Parsed Corpus of Historical Portuguese*, (2017). URL: https://www.tycho.iel.unicamp.br/corpus/index.html, last accessed 10 november 2023.

[7] M. Zampieri, *Colonia*, 2013 URL: https://www.linguateca.pt/acesso/corpus.php?corpus=COLONIA, last accessed 10 november 2023.

[8] C. Galves, 'El corpus Tycho Brahe: un corpus sintácticamente anotado del português histórico', *Revista Binacional Brasil-Argentina: Diálogo Entre As Ciências*, 8(1), (2019), 181-204. https://doi.org/10.22481/rbba.v8i1.5585

[9] M. Zampieri, & M. Becker, 'Colonia: corpus of historical Portuguese', *Non-Standard Data Sources in Corpus-based Research. ZSM-Studien Series*, 5 (2013).

[10] G. Vaamonde, M. Janssen, 'Da edición dixital á análise lingüística. A creación de corpus históricos na plataforma TEITOK', (2020) 271–292. https://doi.org/10.17075/cbfc.2020.008

[11] H. Schmid, 'Probabilistic part-of-speech tagging using decision trees', in *Proceedings of International Conference on New Methods in Language Processing*, (1994) 154–163. H. Schmid, 'Deep learning-based morphological taggers and lemmatizers for annotating historical texts', in *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, (2019) 133–137.

[12] M. I. Bico, J. Baptista, F. Batista, and E. Cardeira, 'Early experiments on automatic annotation of Portuguese medieval texts', in G. Silvello, O. Corcho, P. Manghi, G. M. Di Nunzio, K. Golub, N. Ferro, and A. Poggi, eds., *Linking Theory and Practice of Digital Libraries*, (2022) 442–49. https://doi.org/10.1007/978-3-031-16802-4_44.

[13] D. Santos, N. Seco, N. Cardoso, and R. Vilela, 'HAREM: an advanced NER evaluation contest for Portuguese', in N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik, D. Tapias, eds., *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, (2006). C. Freitas, P. Carvalho, H. G. Oliveira, C. Mota, and D. Santos, 'Second HAREM: advancing the state of the art of named entity recognition in Portuguese', in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias , eds., *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, (2010).

[14] L. Rossi. *A Literatura Novelística Na Idade Média Portuguesa*, 1979.

[15] M. C. A. Lucas. *Hagiografia Medieval Portuguesa*, 1984.

[16] C. Sobral, and E. Cardeira. 'O *Livro Dos Mártires* de Bernardo De Brihuega: dois séculos de leitura em português', *Estudos De Lingüística Galega* 10, (2018), 129-41. https://doi.org/10.15304/elg.10.4613.

[17] E. Cardeira. *O Essencial sobre a História do Português*, 2006.

[18] J. L. Vasconcellos. *Opúsculos III – Onomatologia*, 1931.

[19] CLUL (Ed.). 2014. P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna. URL: http://ps.clul.ul.pt, last accessed 10 November 2023.

[20] Linhares Pontes et al., 'Entity linking for historical documents: challenges and solutions'.

[21] Y. Mosallam, A. Abi-Haidar, and J. G. Ganascia, 'Unsupervised named entity recognition and disambiguation: an application to Old French journals', in P. Perner, ed., *Advances in Data Mining. Applications and Theoretical Aspects*,(2014), 12–23. https://doi.org/10.1007/978-3-319-08976-8_2.

[22] P. Agarwal, J. Strötgen, L. Corro, J. Hoffart, and G. Weikum, 'DiaNED: time-aware named entity disambiguation for diachronic corpora', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), (2018) 686–93. https://doi.org/10.18653/v1/P18-2109.

[23] Brando et al., 'REDEN: named entity linking in digital literary editions using linked data sets'.