Named entity annotation of an 18th-century transcribed corpus: problems and challenges[1]

Helena Freire Cameron[1] [0000-0001-7719-6894], Fernanda Olival[2] [0000-0003-4762-3451], Renata Vieira[2] [0000-0003-2449-5477], and Joaquim Santos[3] [0000-0002-0581-4092]

[1] CIDEHUS, Instituto Politécnico de Portalegre, helenac@ipportalegre.pt,
[2]CIDEHUS, Universidade de Évora, mfo@uevora.pt, renatav@uevora.pt
[3] Escola Politécnica, PUCRS joaquimneto04@gmail.com

Abstract. This paper reviews a stage of the process of annotating named entities in 18th-century texts to enrich historical research sources and link them to other bases. The categories in question are person, location and organisation, valid categories for historian analysis. We discuss the difficulties observed in the process and point eventual solutions.

Keywords: Named entity, *corpus* annotation, 18th-century Portuguese.

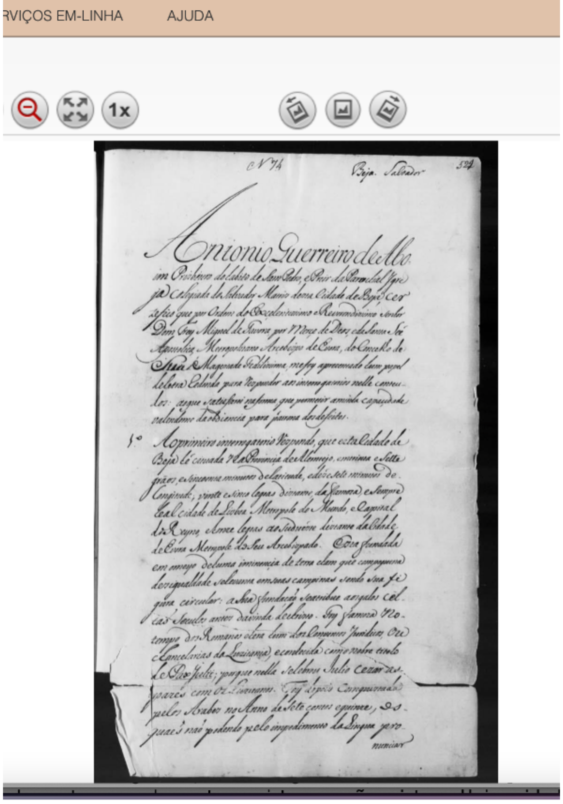
1 Introduction

This paper describes a stage of the process of annotating named entities in 18th-century texts from a transcribed *corpus*, *Memórias Paroquiais* [*Parish Memories*]. We studied a *subcorpus* regarding the biggest region in the south part of Continental Portugal, Alentejo (PM_A). These texts are the answers to a survey sent in January of 1758 to the bishops asking them to resend it to the parish priests to respond, aiming: 1) to obtain feedback about the state of the territory after the big earthquake of 1755; 2) to gather information to create a Geographical Dictionary of Portugal. The inquiry, equal for all the country, had 60 questions, organised into three points: land, mountain and river. It asked about almost everything locally and it finished with an open question about other pertinent topics of the parish not examined in the previous questions.

We began working in the *Parish Memories* in 2007-2008, using the digitised copies of the microfilms available on the Portuguese National Archive of Torre do Tombo website (Figure 1). At first, we just had the goal of transcribing them and posting the results on the web. Next, we tried to make the texts more easily

This work is funded by national funds through the Foundation for Science and Technology, under the project UIDB/00057/2020

searchable, introducing a manual system of tags and making it available online in the CIDEHUS Digital Web Portal[2]. Later on (2020), other possibilities opened up; annotating named entities was one way to advance to information extraction. We aimed to organise the information contained in the *corpus* and also make it available as open data and open linked data. Our first goal is to link data from the *Parish Memories* to data from the book *Corografia Portuguesa* (Costa, 1712), also online in the CIDEHUS Digital[3]. As the process of annotation proceeded, many questions arose. We systematised some of them in the present paper to inventory and discuss the difficulties of annotating 18th-century Portuguese historical texts and to reflect on the next annotation phase.



N 74 Beja. Salvador 524

Antonio Guerreiro de Abo-
im Presbitero do habito de Sam Pedro, e Prior da Parrochial Igre-
ja Collegiada do Salvador Matriz desta Cidade de Beja, cer-
tefico que por Ordem do Excelentissimo e Reverendissimo Senhor
Dom Frey Miguel de Tavora, por Merce de Deos, e da Santa Sé
Apostolica, Metropolitano Arcebispo de Evora, do Conselho de
Sua Magestade Fidellissima, me foy apresentado hum papel
de letra Redonda para responder aos interrogatorios nelle conteu-
dos: ao que satisfarei na forma, que permettir a minha capacidade
valendome da obediencia para [illegible] dos defeitos.

1º Ao primeiro interrogatorio respondo, que esta Cidade de
Beja he situada na Provincia de Alemtejo, em trinta e sette
graos, e [illegible] minutos de Latitude, e dez e sete minutos de
Longitude; vinte e sinco legoas distante da famoza, e sempre
leal cidade de Lisboa Metropole do Mundo, e Capital
do Reyno, e treze legoas ao Sudueste distante da Cidade
de Evora Metropole do seu Arcebispado. Esta fundada
em meyo de huma eminencia de terra a qual que com pequena
desigualdade se levanta em suas campinas sendo sua fi-
gura circular: a sua fundação se atribue aos Gallos Cel-
tas Seculos antes da vinda de Christo. Foy famoza no
tempo dos Romanos e hera hum dos Conventos Juridicos, ou
Chancelarias da Luzitania, e conhecida com o nobre titulo
de Pax Julia; porque nella o celebre Julio Cezar as
pazes com os Lusitanos. Foy depois Conquistada
pelos Arabes no Anno de Sete centos e quinze, os
quaes não podendo pelo impedimento da Lingua pro-
nunciar

Fig. 1. Sample of the digital version of the *Parish Memories,* available at Torre do Tombo website[4].

[2] http://www.cidehusdigital.uevora.pt/portugal1758

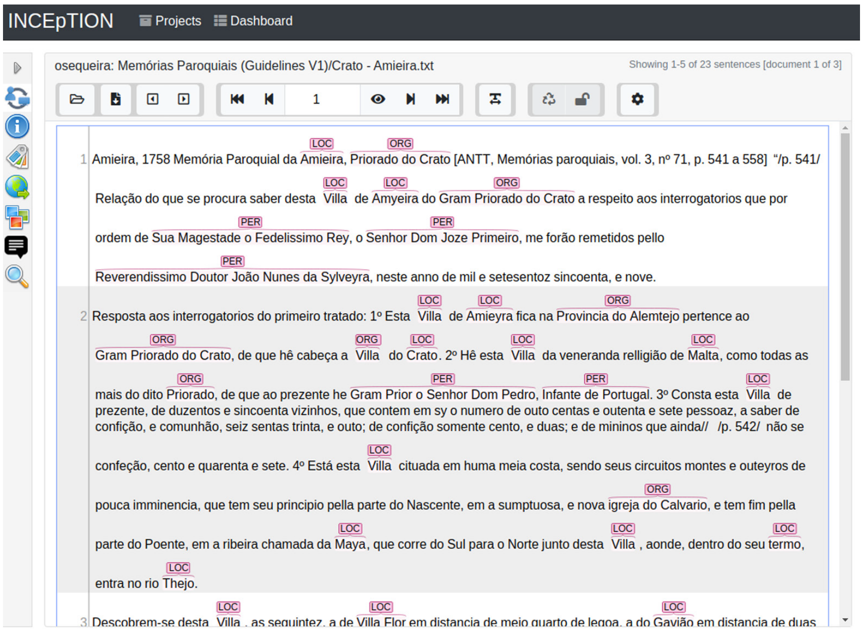
[3] http://www.cidehusdigital.uevora.pt/ophir-restaurada/corografia

[4] https://digitarq.arquivos.pt/details?id=4238720

2 Named Entity manual annotation

Our annotation followed the universal categories in NER works: person, location and organisation. These categories are also of significant importance for history research. Historical events are often linked to agents, places and institutions. Many systems were developed to recognise these categories, so we selected them for our first annotation attempts considering the 18th-century texts. We followed guidelines previously proposed by the HAREM project (Santos and Cardoso, 2007), a named entity recognition contest based on contemporary Portuguese. However, we have not yet distinguished different types of mentions. In HAREM subtypes, a person was also marked as being an individual with other singular attributes (like occupation), as in Elisabeth II or the Queen of the United Kingdom, respectively. Capital letters constituted a criterion to identify named entities. Although the use of capital letters is not systematic in PM_A (they were used randomly in common nouns but correctly in the majority of named entities), we followed this directive.

Two historians annotated a representative sample of the *Memories* manually, using the Inception tool (Klie et al., 2018). This annotation platform includes semantic annotation (e.g., concept linking, knowledge base population), essential features for History researchers. The annotation interface is presented in Figure 2.



osequeira: Memórias Paroquiais (Guidelines V1)/Crato - Amieira.txt

1 Amieira, 1758 Memória Paroquial da Amieira, Priorado do Crato [ANTT, Memórias paroquiais, vol. 3, nº 71, p. 541 a 558] "/p. 541/ Relação do que se procura saber desta Villa de Amyeira do Gram Priorado do Crato a respeito aos interrogatorios que por ordem de Sua Magestade o Fedelissimo Rey, o Senhor Dom Joze Primeiro, me forão remetidos pello Reverendissimo Doutor João Nunes da Sylveyra, neste anno de mil e setesentoz sincoenta, e nove.

2 Resposta aos interrogatorios do primeiro tratado: 1º Esta Villa de Amieyra fica na Provincia do Alemtejo pertence ao Gram Priorado do Crato, de que hê cabeça a Villa do Crato. 2º Hê esta Villa da veneranda relligião de Malta, como todas as mais do dito Priorado, de que ao prezente he Gram Prior o Senhor Dom Pedro, Infante de Portugal. 3º Consta esta Villa de prezente, de duzentos e sincoenta vizinhos, que contem em sy o numero de outo centas e outenta e sete pessoaz, a saber de confição, e comunhão, seiz sentas trinta, e outo; de confição somente cento, e duas; e de mininos que ainda// /p. 542/ não se confeção, cento e quarenta e sete. 4º Está esta Villa cituada em huma meia costa, sendo seus circuitos montes e outeyros de pouca imminencia, que tem seu principio pella parte do Nascente, em a sumptuosa, e nova igreja do Calvario, e tem fim pella parte do Poente, em a ribeira chamada da Maya, que corre do Sul para o Norte junto desta Villa , aonde, dentro do seu termo, entra no rio Thejo.

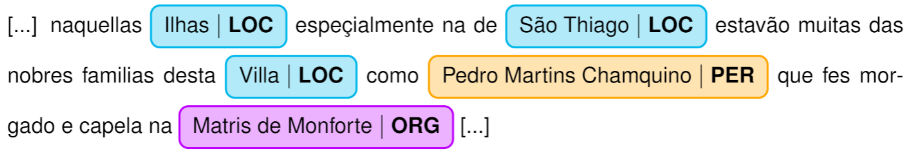
3 Descobrem-se desta Villa , as seguintez, a de Villa Flor em distancia de meio quarto de legoa, a do Gavião em distancia de duas

Fig 2. INCEpTION interface for named entity annotation.

The annotation agreement was measured, resulting in Kappa = 0.71, considered a substantial agreement (McHugh, 2012). Table 1 shows the number of annotated examples for each class.

Table 1: Number of NE manually annotated

Category	Annot1	Annot2
PER	491	474
LOC	693	514
ORG	444	368
Total	1628	1356



[...] naquellas Ilhas | **LOC** espeçialmente na de São Thiago | **LOC** estavão muitas das nobres familias desta Villa | **LOC** como Pedro Martins Champquino | **PER** que fes morgado e capela na Matris de Monforte | **ORG** [...]

Fig. 3. Examples of named entities

In Figure 3, we show some examples of each category occurring in the *Parish Memories*, the text also illustrates some spelling differences from that period regarding the contemporary stage of the Portuguese language (*naquellas, espeçialmente, estavâo, villa, fes*). The Portuguese spelling standard was still not fixed in the 18th-century, and many allographs can occur in the same document. There were rules, but nobody was obliged to follow them (Cardeira, 2006; Cardeira & Mateus, 2008; Marquilhas, 2000).

Variance in MP depends upon not only linguistic phenomena like spelling or others but also the ability of the priest and, eventually, the interpretation of the transcriber, as we are dealing with transcribed texts from handwritten sources.

The manual annotation served two purposes, to evaluate available named entity recognition systems applied to this kind of text (Vieira et al, 2021) and to prepare for a next phase of annotation of a larger portion of the corpus to adapt the current system and improve their performance. The preparation for the next phase includes the creation of new annotation guidelines, and in order to advance towards this goal we reflected on the problems and challenges, as we may see further below.

3 Problems and challenges of an 18th-century *corpus*

To work with 18th-century sources is quite different from dealing with nowadays written registers. In this paper, we describe some of the problems and challenges encountered during the annotation. We organised them into a few categories.

Morphology and Spelling: Variation is a challenge for the 18th century Portuguese language's automatic processing of texts as most existing tools operate in the contemporary Portuguese language (Cameron et al., 2020). In the 18th-century, variation could succeed due to linguistic issues and spelling variation. Also, the graphic registration of uppercase and lowercase letters was not consistent: they were used randomly on the original, and some transcribers kept the randomness; others interpreted it and updated it to current Portuguese. Capital letters are usually relevant for NER identification, but their presence may be less consistent in this kind of *corpus.* Regarding punctuation marks, they do not follow the current pattern precisely. This variance in this stage of the language is a constraint not only for automatic machine processing but also for human readers, even if they are language experts, and in minor questions. For example, regarding the allographs *s/ç* in the initial position of the word, in *Çamora/Samora* [Samora Correia, name of a parish], the form used is *Çamora,* and it is inserted in the alphabetic sequence beginning with C included in the last volume of *Memórias Paroquiais,* what may trouble to find the document.

Observations about each category:

PER: The 18th-century Portuguese society is highly hierarchical, where social category defines people. For that, the so-called pre-nouns are very important. Sometimes they are mixed with forms of treatment or protocol forms. They can help identify the social category or the occupation, as in the following example.

1. *Exmo. e Ilustríssimo Senhor D. Manuel da Gama, Vice-Rei da Índia*

In the annotation process, if we annotate the person's name without the pre-nouns, we lose essential information about that person.

ORG: The distinction between physical localities from political entities is often subtle in these texts, wherein the same name can be a Local or an Organisation (Álvarez-Mellado et al., 2021), as in:

2. *Esta <LOC>Villa de Amieyra</LOC> fica na <LOC>Provincia do Alemtejo </LOC> pertence ao <ORG>Gram Priorado do Crato</ORG>, de que hê cabeça a <ORG>Villa do Crato</ORG>.*

LOC: The entity Loc refers to specific places. However, sometimes the parishes do not specify the local but they indicate a general location, as in:

3. *Esta Vila está a meio da encosta*

“meio da encosta” is not a specific location, so it would not be a LOC entity.
The following issues are not particular to 18th-century texts but are general problems for NER that may affect the annotation process.

Discontinuity: The annotation process usually considers the named entity in a continuous sequence. However, in texts, names may appear in a reduced form or separated in the discourse. In the sequence below, “Irmandade de Nossa Senhora do Rozario” is registered in the text as “Nossa Senhora do Rozario”, and the human reader understands that this name is related to “Irmandades”.The entity itself should be named *Irmandade de Nossa Senhora do Rozario*, where, in the text, “irmandade” is written in the plural, as it introduces an enumeration of organisations of this type.

4. *As <ORG>Irmandades</ORG> que existem, dentro della [Vila], são sinco, a saber, a do <ORG>Sanctissimo Sacramento</ORG> a que esta anexa a das <ORG> Quarentas Horas</ORG>; a de <ORG>Nossa Senhora do Rozario</ORG>; a de <ORG>Nossa Senhora da Graça<ORG>, a das <ORG>Almas<ORG>, a do <ORG>Appostolo Sam Pedro advincula</ORG>, administrada pelo clero desta Villa.*

A special attention must be given to the surrounding discourse context to consider *Nossa Senhora do Rozario, Appostolo* or *Sam Pedro advincula* rightfully as Organizations and not Persons.

Embedded entities: The great majority of work in NER does not consider embedded entities. One must choose between annotating constituents or the more extensive sequence. In the following examples, we can see a mixture of entities. Whereas it is a general assumption to consider the complete mention, in (5), we can identify both the college name and the university name. In (6), we have a person's name embedded in the organisation’s name.
5. *Collegio de Sao Paullo da Universidade de Coimbra*
6. *Companhia da Infante Dona Brites Pereira Saboya*

4 Next steps

Although the basic categories, Person, Location and Organisation, are relevant, there are significant improvements to incorporate to better capture or reproduce the society of the time and its structural differentiation. Inspired by the experience of prosopographical datasets regarding persons, we consider adopting the subcategories name, occupation, and social category. A proper name may refer to a person, but also, in some cases, references are made or composed by distinct occupations or social status that constitute essential information in historical analysis, as briefly mentioned above. To deal with the relation between persons (family, patronage relations and others) is an issue to be developed shortly.

Location and organisation are challenging to distinguish when geopolitical entities are involved, such as names of countries, states and cities. Therefore, in some previous work, we adopted the GPE category to avoid this ambiguity problem. Besides GPE, we may find other Location and Organisation subcategories such as rivers, mountains for places and churches for organisations. Also, time is another crucial tag missing.

To better describe the work to be done and prevent ambiguities in annotation when using several annotators, we are now working on new guidelines adjusted to the 18th-century Portuguese. Annotating is a way of enriching a *corpus*, and this effort must handle the past complexity and not misrepresent it with anachronistic simplifications. It is essential to look at a word in the economy of the text and to look at the text in its global context. In this sense, the presence of a social historian is significant in a team dealing with distant past *corpora*, be they composed by literary texts or administrative ones. Once we have the new annotated *corpus*, we will retrain existing NER models to annotate this kind of *corpus* for history research automatically.

References

1. Álvarez-Mellado, E., Díez-Platas, M. L., Ruiz-Fabo, P., Bermúdez, H., Ros, S., González-Blanco, E. (2021) TEI-friendly annotation scheme for medieval named entities: a case on a Spanish medieval *corpus*», *Language Resources and Evaluation*, vol.55, nº 2, 525–549.
2. Cardeira, E. (2006) *O essencial sobre a História do Português*. Alfragide: Editorial Caminho.
3. Cardeira, E., Mateus, M. H. (2008). *Norma e Variação*. Alfragide: Editorial Caminho.
4. Cameron, H. F., Gonçalves, M. F., & Quaresma, P. (2020). Linguistic and orthographical classic Portuguese variants challenges for NLP. *Proceedings of the 14th International Conference on the Computational Processing of Portuguese*, (pp. 43–48).
5. Costa, A. C. (1712). C*orografia Portugueza e descripçam topografica do famoso reyno de Portugal: com as noticias das fundações das cidades, villas, & lugares, que contem, varões illustres, genealogias das familias nobres, fundações de conventos, catalogos dos bispos, antiguidades, maravilhas de natureza, edificios & outras curiosas observaçoens*. Lisboa: na officina de Valentim da Costa Deslandes, Tomo primeyro [-terceyro], vol. 1-2-3.
6. Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational*

Linguistics: System Demonstrations, pages 5–9. Association for Computational Linguistics

7. Marquilhas, R. (2000). *A faculdade das letras: leitura e escrita em Portugal no séc. XVII*. Lisboa: IN-CM.
8. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3), 276–282.
9. Santos, D. & Cardoso, N. (edits) (2007). Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. *Linguateca*: digital print.
10. Vieira, R., Olival, F., Cameron, H., Santos, J., Sequeira, O., Santos, I. (2021): Enriching the 1758 portuguese parish memories (alentejo) with named entities. Journal of Open Humanities Data 7, 20.