Analysing entity distribution in an annotated 18th century historical source

Daniel De Los Reyes[1], **Renata Vieira**[2], **Fernanda Olival**[2], **Helena Freire Cameron**[3], **Fátima Farrica**[2]
[1]Pontifical Catholic University of Rio Grande do Sul, PUCRS
[2]CIDEHUS - University of Évora, [3]CIDEHUS - Portalegre Polytechnic University, Portugal
daniel.reyes@edu.pucrs.br, renatav@uevora.pt,
mfo@uevora.pt, helenac@ipportalegre.pt,fatimafarrica@sapo.pt

Abstract

This paper presents a distribution analysis of named entities in a historical source, an 18th century Portuguese text collection. The source has been transcribed, revised, normalised and annotated manually with the help of an annotation tool. The distribution analysis was carried out automatically with the help of an extraction parser applied to the annotated texts. The central question of this text is to analyse the meaning of this distribution.

1 Introduction

Named entity recognition (NER) for history research is becoming a trend. In (Ehrmann et al., 2023), we find a survey on named entity recognition and classification in historical documents that considers a variety of other languages. Among others equally pertinent, we may refer to recent studies for NER for historical Portuguese (Grilo et al., 2020; Aguilar et al., 2017; Zilio et al.).

This study introduces a novel exploration of a set of historical Portuguese texts referred to as the "Parish Memories", which were created during the period from 1758 to 1761. This collection holds significant cultural and historical value, comprising the responses to a survey containing 60 questions and distributed in 1758. The answers, originally handwritten by parish priests across the entire kingdom of Portugal, have been meticulously transcribed and normalised for analysis (Olival et al., 2023a).

Our earlier research (Vieira et al., 2021) conducted experiments involving three primary categories (PERSON, LOCAL, ORGANISATION). Later, we performed a *corpus*-based study to define the extension of these categories (Cameron et al., 2022), which subdivide them into more detailed classes as presented in Section 2.

This paper seeks to provide an overview and discussion of the distribution of annotated entities within these more refined categories in the historical collection under consideration.

2 NE categories customised to historical research

The annotation process of this work endeavours to capture the intricacies conveyed in historical sources from past ages, recognising their distinctions from contemporary expressions. We started by delineating five primary categories: PERSON, PLACE, ORGANISATION, TIME, and AUTHOR WORK. The initial four categories seek to address historical queries related to Who, Where, What, and When, while the last category enables us to analyse the textual sources referenced in the *corpus*.

Owing to their intricacy and significance for the source study, the primary categories PERSON and PLACE were divided into several subcategories.

2.1 Sub-categories of Person

The society of the 18^{th} century was characterised by the inequality of individuals before the law and numerous markers of social differentiation. Often, titles and occupational positions were integral to a person's name and identity. For the annotation to be helpful to historians, it must replicate this reality.

Therefore, the category person (PER) considers references by name (PER_NAM), occupation (PER_OCC), or social category (PER_CAT) - in that order of preference if more than one appears in the expression - and group of persons (PER_GROUP).

Examples of mentions of persons by occupation, social category, and groups of persons are:

- Arcebispo de Évora [Archbishop of Évora]
- Conde da Torre [Count of the Tower].
- Sequeiras [the Sequeira family]

There are also specific subcategories for mentions of saints (PER_SAINT), divinities

(PER_DIV), and authors (PER_AUT).

2.2 Sub-categories of Place

Concerning places, we generalised the usual category location (LOC) to place (PLC). This category is subdivided into geopolitical entities (PLC_GPE), aquifers (PLC_AQU), mountains (PLC_MOUNT), facilities (PLC_FAC), and one extra subcategory for other locations (PLC_LOC). References to geographical points, such as rivers and mountains, are essential for geo-references.

2.3 Other categories

Regarding time expressions (TIM), we only annotated specific references to dates, for instance, the year 1755 [the year of 1755].

Organisation (ORG) includes all typologies of organisations, like, for example:

- Convento de Santo António [Santo António Monastery]
- Confraria de São Pedro [São Pedro Fraternity]

Written documents mentioned in the memories were attributed to the category AUTWORK.

3 Distribution analysis

The tool used for the manual annotation was the INCEPTION platform[1]. We worked on the output files generated with the annotation information. We have different files for each parish of each municipality, this organisation allows the analyse the entities across parishes and municipalities.

The annotated subset gathers 71 parishes of Alentejo, corresponding to 17% of parishes of this region, the largest in Portugal. However, qualitatively, they belong to the most important municipalities of the region: Beja, Évora, Portalegre, and Vila Viçosa. The first three are currently the district capitals. Vila Viçosa, in the past, was the headquarters of the Duke of Bragança, the manor house that served as the birthplace of the Portuguese ruling dynasty in 1758-61.

Municipality	Parishes	Texts	NEs
Beja	29	**695**	1895
Évora	22	**879**	1836
Portalegre	14	**312**	855
Vila Viçosa	6	**210**	474
Total	71	**2096**	5060

Table 1: Distribution of NEs by parish

[1] https://inception-project.github.io

Analysing named entities (NEs) extracted from historical texts of parishes in Portugal is essential to understanding the vast range of information present in the documents. In this detailed analysis, we present explanatory graphs showing the categories of named entities by the municipality, the general distribution of these categories, and the main named entities highlighted by the municipality and globally.

We created a text parser to simplify the extraction, analysis and organisation of named entities extracted from Inception's output file. This parser can be found in the following repository[2]. After applying the parser to the annotated texts, we explored the results of annotation and categorisation in general and also by the municipality. Table 1 shows that we analysed more than 2000 texts, covering 71 parishes across four municipalities. Also, we can see in Table 1 that we have 5060 annotated NEs as a result of the manual annotation.

3.1 Distribution of named entity by categories

Table 2: Distribution of NE categories

Category	Distribution (%)
PLC	42.66
PER	40.43
ORG	7.96
TIM	6.22
AUTWORK	2.73

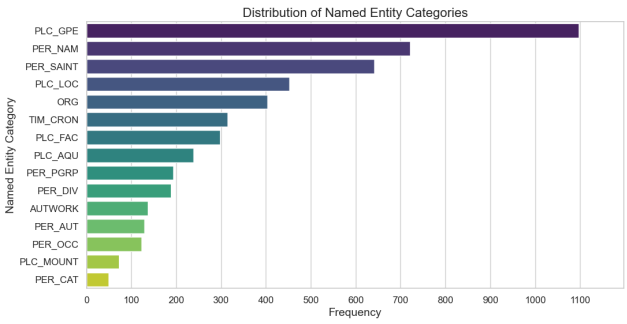

Figure 1: General distribution of NE categories

Understanding the general distribution of named entity categories across historical texts is essential to contextualising the research. Table 2 highlights the predominance of PLACE and PERSON categories compared to the others, totalling more than 80% of the named entities noted in these texts.

[2] https://github.com/DanielReeyes/inception-entity-parser

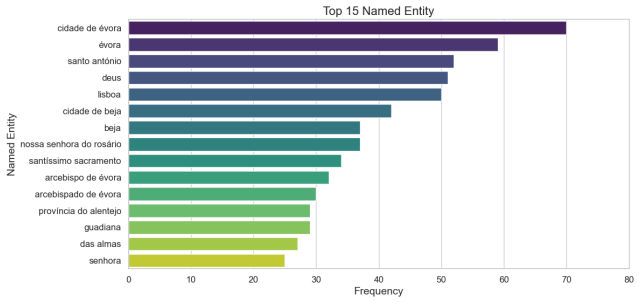

Figure 2: Top 15 NEs among all municipalities

The distribution is unbalanced, where the major categories represented in the corpus are related to geopolitical entities, person names, and saints. Persons are referenced only by category, and mountains are the least represented.

Figure 1 provides an insightful representation of the global distribution of named entities categorised into subcategories. By examining this chart, one can gain a panoramic view of the main categories present in historical texts and identify areas of focus and concentration of information. This visual representation makes comprehending the distribution of NEs and their subcategories easier, allowing for a more comprehensive analysis of the text data.

Based on the distribution of macro categories, the subcategories that appear the most are PLC_GPE (referring to geopolitical entities), PER_NAM (referring to personal names), and PER_SAINT (referring to saint names), with 1097, 721, and 642 annotations, respectively.

To gather more specific information across all parishes, Figure 2 shows the 15 most referenced named entities across all texts from all analysed municipalities. It summarises the most significant entities found in the historical texts of all parishes, offering a comprehensive perspective on the most recurrent and important elements in analysing the named entities. Specifically, Évora, Lisbon, and Beja were the most frequently mentioned named entities in the GPE category among all entities. They were followed by named entities in the saints and holiness categories, such as Saint Anthony, God, and Our Lady of the Rosary, a widely spread devotion in Portugal after the Counter-Reformation. Lisbon also received significant mentions. In the first group, the most surprising inclusion is Beja, considering that at that time, Beja was not the episcopal capital. The question posed in the inquiry was: "How far is the parish from the episcopal capital city, and how far is it from Lisbon, the capital of

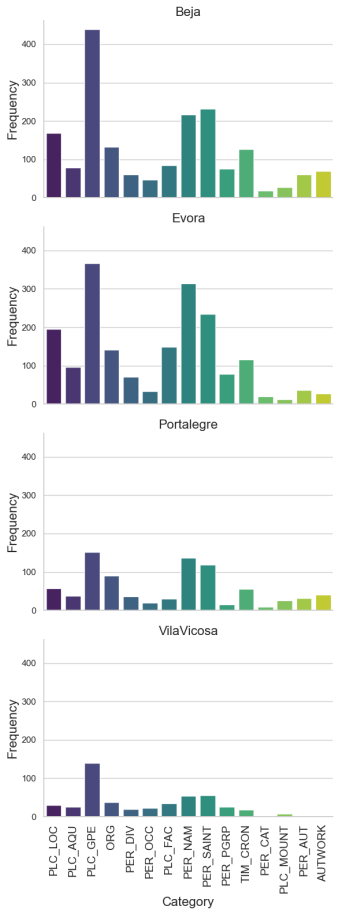

Figure 3: Distribution of NE subcategories by municipality

the Kingdom?" As a result, Lisbon was consistently cited in each parish.

3.2 Distribution of named entities by municipality

To analyse thematic variations between parishes, we analysed the distribution of named entity categories. Figure 3 displays the frequency of categories in each municipality.

The graph analysis reveals that they all follow the same pattern of subcategory distributions as the global context. All have a prevalence of PLC_GPE categories and subcategories, even when analysed scenario by scenario. The other two subcategories highlighted in the general analysis, PER_NAM and PER_SAINT, are also present when we analyse the data by the municipality.

Identifying the main named entities in each municipality is crucial to gaining a more specific view of local particularities. Figure 4 representing the top 15 NEs by municipality provides a better understanding of the local context. This visualisa-

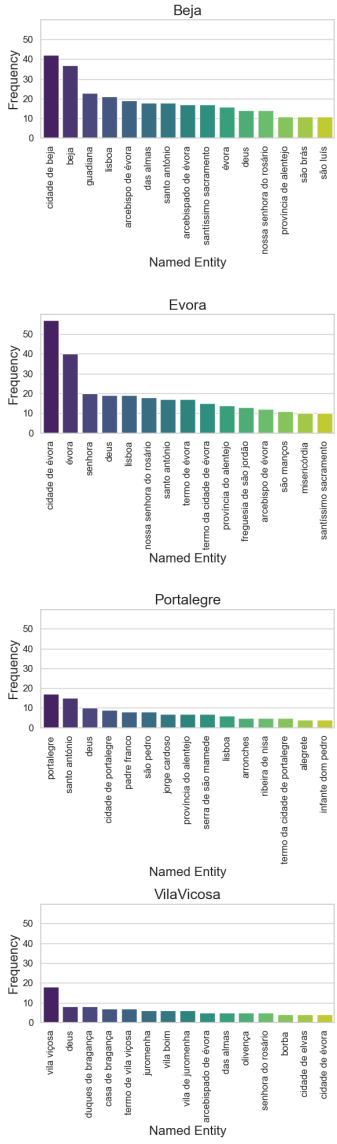

Figure 4: Top 15 NEs by municipality

tion highlights the most prominent entities in each location, contributing to a deeper understanding of local history. For instance, we can see the importance of Évora across municipalities, we see different saints mentioned across regions, Saint Anthony ("Santo António") in Beja and Évora, Saint Peter ("São Pedro") in Portalegre. These two male saints represent two of the three most venerated in Portugal. Saint Anthony, who is believed to have been born in Lisbon, has been widely invoked since the 17th century to locate lost objects, while Saint Peter serves as the guardian of the keys to heaven, the patron saint of the Church, and the papacy (Farmer, 1997). The latter also symbolises the reaffirmation of the triumphant Catholic Church after the Counter-Reformation. The Bragança family is highly mentioned in Vila Viçosa. It represented a key element of the town's identity due to its direct ties to royalty and its contributions to the local community through the sponsorship and establishment of convents, chapels, and other communal facilities. It was consistently referenced positively, and in just one of the parishes, there was mention of the obligation for the population to pay taxes when constructing mills and other devices in the local aquifers (Olival et al., 2023b).

4 Conclusion

In this work, we presented a study on the collection of *Parish Memories*, which describes aspects of Portugal from the 18^{th}century. In this work, Named Entities were analysed regarding their distribution in the parishes of the Alentejo region. The predominance of GPE can reinforce the idea that this survey was launched in 1758 to resume the project of a Geographical Dictionary of Portugal, initiated before the earthquake of 1755 and interrupted by this catastrophe (Olival et al., 2023a).

Exploring this data helps achieve valuable insights from historical registers about the parishes in Portugal at that time, helping to gain a richer and more contextualised understanding of local history. By studying these Named Entities through manual annotation of historical texts, we can create more robust and reliable datasets and compare between parishes. The annotation enables us to conduct experiments to develop and test methodologies such as Artificial Intelligence models for extracting named entities, making it possible to automate this type of task (Santos et al., 2024).

Acknowledgements

This work has received financial support from the Portuguese Science Foundation FCT in the context of the projects CEECIND/01997/2017 and UIDB/00057/2020 - https://doi.org/10.54499/UIDB/00057/2020.

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153.

Helena Freire Cameron, Fernanda Olival, Renata Vieira, and Joaquim Francisco Santos Neto. 2022. Named entity annotation of an 18th century transcribed corpus: problems, challenges. In *Proceedings of the*

Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Virtual Event, Fortaleza, Brazil, 21st March, 2022, volume 3128 of *CEUR Workshop Proceedings*, pages 18–25. CEUR-WS.org.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

David Farmer. 1997. *The Oxford dictionary of saints, 4th ed.* Oxford University Press, Oxford, UK.

Sara Grilo, Márcia Bolrinha, João Silva, Rui Vaz, and António Branco. 2020. The BDCamões collection of Portuguese literary documents: a research resource for digital humanities and language technology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 849–854, Marseille, France. European Language Resources Association.

Fernanda Olival, Helena Freire Cameron, Fátima Farrica, and Renata Vieira. 2023a. As memórias paroquiais (1758) do atual concelho de vila viçosa. *Callipole: revista de Cultura*, (29):85–128.

Fernanda Olival, Helena Freire Cameron, and Renata Vieira. 2023b. As memórias paroquiais: Do manuscrito ao digital. *Atas da Jornada de Humanidades Digitais do CIDEHUS*.

Joaquim Santos, Renata Vieira, Fernanda Olival, Helena Cameron, and Fatima Farrica. 2024. Named entity recognition specialised for portuguese 18th century history research. In *Proceedings of International Conference on the Computational Processing of Portuguese (PROPOR 2024)*.

Renata Vieira, Fernanda Olival, Helena Cameron, Joaquim Santos, Ofélia Sequeira, and Ivo Santos. 2021. Enriching the 1758 portuguese parish memories (alentejo) with named entities. *Journal of Open Humanities Data*, 7:20.

Leonardo Zilio, Maria Jose Bocorny Finatto, and Renata Vieira. Named entity recognition applied to portuguese texts from the 18th century. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022)*, volume 3128.