



Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities

RESEARCH PAPER

RENATA VIEIRA

FERNANDA OLIVAL

HELENA FREIRE CAMERON

JOAQUIM SANTOS

OFÉLIA SEQUEIRA

IVO SANTOS

**Author affiliations can be found in the back matter of this article*

]u[ubiquity press

ABSTRACT

This work presents an enriched version of the Parish Memories (1758–1761), an essential Portuguese historical source manually transcribed. It is enriched with annotations of named entities of the types PERSON, LOCATION, and ORGANIZATION. The annotation was done automatically for the whole collection where two researchers annotated a portion of it manually for evaluation purposes. In this dataset, we provide the tagged texts, the lists of extracted entities, and frequency counts. The corpus is useful for historians, allowing, for instance, comparative analyses between parishes and regions or to calculate the area of influence of a locality. The paper describes the creation and evaluation of the corpus, discusses its applications and limitations. This first release may be improved by other researchers interested in the historical source itself or in the technology employed in its annotation.

CORRESPONDING AUTHOR:

Renata Vieira

CIDEHUS, University of Évora,
Portugal

renatav@uevora.pt

KEYWORDS:

Digital History; Historical Sources (18th century); Named Entity Recognition; Information Extraction; Named Entity Evaluation; Named Entity Labelled Corpus

TO CITE THIS ARTICLE:

Vieira, R., Olival, F., Cameron, H. F., Santos, J., Sequeira, O., & Santos, I. (2021). Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities. *Journal of Open Humanities Data*, 7: 20, pp. 1–13. DOI: <https://doi.org/10.5334/johd.43>

(1) CONTEXT AND MOTIVATION

The area of Digital Humanities (DH) stands at the intersection of computing and the humanities. Each day, DH involves more collaborative and transdisciplinary research, bringing digital tools and methods for studying the humanities. DH is an essential growing research field for which natural language processing (NLP) has much to offer. NLP may enable researchers to explore large amounts of data and discover new elements and correlations according to different aspects of each research area. This trend is also known for historical research, and it is claimed that computers are already ubiquitous in historical scholarship (Romein et al., 2020).

Named Entity Recognition (NER) is an NLP task able to find proper nouns in a given text and classify them according to various pre-defined categories (Jurafsky & Martin, 2000). In recent years, the advent of neural networks that automatically learn effective features from raw text (Collobert et al., 2011) has allowed NER systems to bypass the manual selection of features. The effectiveness of word and character embedding (continuous real vectors) obtained by pre-trained neural language models (LMs) (Chiu & Nichols, 2016; Collobert et al., 2011; dos Santos & Guimarães, 2015; Lample et al., 2016) has significantly impacted these systems. These embeddings encapsulate semantic, syntactic, and morphological information about the language in a way that can easily be incorporated into a sequence tagger, usually implemented as a recurrent neural network. Such systems are available for easy use in NLP frameworks such as spaCy.¹

More recent state-of-the-art NER systems use contextualised embeddings such as BERT and Flair (Santos et al., 2019a; Souza et al., 2019). This is accomplished by adding to the system the embedding matrix of a pre-trained LM and incorporating the pre-trained LM itself as the initial layers of the system. An example of a system that uses this method for English and German is presented in Akbik et al. (2018). For the Portuguese language, such systems are presented in Souza et al. (2019) and Santos et al. (2019a).

This paper aims to present a corpus of texts dated 1758–1761, the Parish Memories, enriched with the (automatic) annotation of entities of the classes person (PER), location (LOC), and organisation (ORG). The reasons for this choice of classes are both the availability of systems that recognise them and their relevance for historians, as explained later in the paper.

For the automatic annotation, we evaluated two available Portuguese NER systems. One is a commonly used framework, spaCy, which includes Portuguese, among other languages. The second was developed specifically for the Portuguese language and has been well evaluated in many domains (Santos et al., 2019a). The evaluation considered a portion of the corpus, where the automatic extraction was contrasted with human analysis. The manual annotation is presented and discussed regarding the particularities of the corpus, with insights for improving the annotation process, both manual and automatic, in further research. For this paper, however, its main role is to demonstrate the quality of the enriched corpus provided.

The paper is organised as follows: Section 2 presents the original corpus, introduces the NER systems considered for evaluation, and describes the manual annotation; Section 3 discusses the systems' evaluation; Section 4 presents the resulting annotated corpus and the usefulness of obtaining knowledge through the automatic processing of textual historical corpora; Section 5 presents conclusions and future work.

(2) MATERIAL AND METHODS

This section presents the corpus provided, the systems considered for automatic annotation, and the process of manually annotating the entities.

(2.1) THE CORPUS: PORTUGUESE PARISH MEMORIES – ALENTEJO

The Parish Memories-Alentejo (PM-A) corpus is an essential source for obtaining Portugal's description in 1758–1761. Digitised copies of the original documents are available at Arquivo Nacional da Torre do Tombo.² **Figure 1** presents an example of the original manuscript.

1 <https://spacy.io>.

2 <https://digitarq.arquivos.pt/details?id=4238720>.

[...] naquellas **Ilhas | LOC** especialmente na de **São Thiago | LOC** estão muitas das nobres famílias desta **Villa | LOC** como **Pedro Martins Chamquino | PER** que fes morgado e capela na **Matris de Monforte | ORG** [...]

Figure 1 Digital copies of the original Parish Memories, in the Portuguese National Archive of Torre do Tombo.

The collection corresponds to a survey sent to those responsible for all country dioceses at the beginning of 1758, and then distributed to the parish priests. The Portuguese crown sent the survey to better know the country's real situation less than three years after the big earthquake of 1755. At that time, it was usual for the sovereign states and some academies to conduct surveys to improve their knowledge about certain territories or topics. Portugal is no exception, and as such, the coverage and great detail of the Parish Memories make it one of the most important surveys of its kind (Chorão, 1987). It contains detailed descriptions of almost each parish of which there were about 4,232 of these religious units in Portugal, in 1798 (Santos, 1995). The parish was the smallest organisational unit in the country at that time.

The survey was organised into three major parts: land, mountain, and river. It included 27 questions about the land, 13 about the mountains, and 20 about the rivers. The parish priests only responded to what suited their territory, and their texts were handwritten. Regarding the land section, the questions included historical, administrative, jurisdictional (ecclesiastical and secular), geographical, and economic topics. It also asked about the impact of the 1755 earthquake in each locality and requested people to describe the reconstruction's status. Questions about the mountains were asked to obtain descriptions of orography, fountains, medicinal herbs, mines, lagoons, villages, and monasteries. Concerning the river section, the detail was also significant: size and intensity of flow, navigability, the direction of the stream, fish and fishery-related activities, bridges, mills, cultivation of the banks, and many other subjects were asked about. Each of the three parts ended with an invitation to describe what was specific and relevant about each place and was not analysed in the previous topics (Santos et al., 2020a). Some parish priests accepted this last invitation, and others had nothing special to point out.

Transcribed digital versions of these reports are available through CIDEHUS website.³ These transcribed versions constitute a sub-corpus with texts from 366 parishes of today's Alentejo region, Southern Portugal's largest administrative area. The PM-A represents 90% of the total of Alentejo's parishes in 1758.

In these originally handwritten texts, the orthography variations are remarkable, as there was no standardisation at that time, in the sense that there was no uniform spelling that everyone followed (Mateus & Cardeira, 2007). This classical Portuguese language period is, in fact, characterised by significant variation in orthography (Gonçalves, 2003). Nasal diphthongs have many spellings, and the use of double consonants and pseudo-etymological spelling were frequent. The sibilants register is also an excellent example of variation, among others (Kemmler, 2001).

³ <http://www.cidehusdigital.uevora.pt/portugal1758/memorias>.

In the PM-A corpus, these phenomena are pervasive and illustrate the orthographic variation: ‘tãobêm/tambêm’ [also], ‘administrassam/administração’ [administration], ‘officiais/oficiais’ [oficial], ‘parcho/paroco’ [priest], ‘oitosentos/oitocentos’ [800], and ‘freguesia/freguezia’ [parish]. The variation is sometimes difficult to predict, as we can find many spelling variants of a single word, like in ‘noticcia/noticia/notiçia/notticia’ [news]. In this period of the history of the Portuguese language, there was a remarkable renewal in the lexicon; many new words entered the written language (Verdelho, 1987), enlarging the possibilities of variation in words.

Due to the large number of texts (one for each of the 366 parishes), the original documents were transcribed by a group of people, including undergraduate students and research fellows, over a period of 12 years (2008–2020). It also comprises transcriptions previously published by other authors outside our research group. The transcriptions were collected as MS Word documents. Although not having a significant impact on the historical research objectives, these different origins reflect distinct, and sometimes contrasting, transcription criteria. The resulting variation is a challenge for the automatic processing of classical Portuguese as most existing tools operate in the contemporary Portuguese language (Cameron et al., 2020). In addition, uppercase and lowercase letters were used randomly in the original documents, with some transcribers keeping the randomness and others interpreting and updating it to current Portuguese. This constitutes a particular challenge for NER, since uppercase is usually a strong indicator. Also, the syntax and punctuation marks do not precisely follow the contemporary usage and conventions, which constitutes another challenge.

(2.1.1) Annotated Entities in the Parish Memories

Named entities recognised by most NER systems are generally of the types PER, LOC, and ORG. These entities are also essential topics for historians as they are for NLP in general. They structure a necessary part of the historian’s inquiry, answering basic questions: Who? Where? Which institution? In the past, everything happened in a specific place, was developed or executed almost always by someone, even if it is not known by whom, and organisations/institutions often framed the actions. **Figure 2** shows examples of each category occurring in the PM-A.

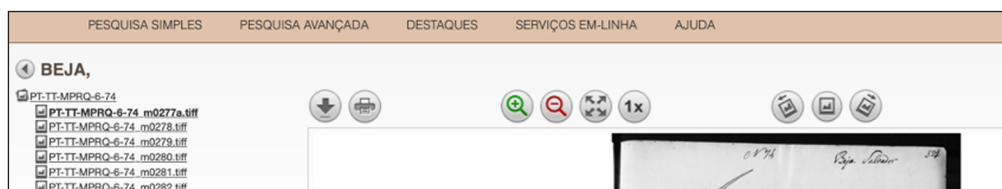


Figure 2 Examples of named entities in the PM-A.

(2.2) THE NER SYSTEMS

There are many software options regarding the NER task (Schmitt et al., 2019). We evaluated two NER systems: the Portuguese NER model available at the spaCy framework and a Portuguese NER model, based on Flair contextual embeddings assessed in several domains (Santos et al., 2019a). The former was developed to consider industrial needs, and the latter is an academic product that has shown the best results in several evaluations. The intention was to choose the NER system with the best performance for these classical Portuguese texts. Both spaCy and Flair frameworks provide their own tokenisation.

(2.2.1) spaCy

The spaCy framework is ready to use for several natural language tasks. It features Convolutional Neural Network (CNN) models for NER, POS tagging, among other tasks. Pre-built statistical neural network models to perform these tasks are available for several languages, including contemporary Portuguese. It contains three Portuguese models based on word embeddings. We have used the most significant model available at the time.⁴ The Portuguese multi-task CNN sequence tagging model was trained on the Portuguese WikiNER (Nothman et al., 2013)

⁴ pt_core_news_lg-2.3.0.

(2.2.2) BCF

The second system used was BCF⁶ (BiLSTM-CRF+FlairBBP), which is composed of a neural network previously used for NER in English and German (Akbik et al., 2018) and a language model called FlairBBP. This model was developed based on a raw text corpus of 4.9 billion words from contemporary Portuguese texts (Santos et al., 2019a). BCF was previously evaluated in several domains, including geoscience (Consoli et al., 2020), law (Santos et al., 2019b), and health (Santos et al., 2020b), and has been the one with the best performance across domains. We trained the BCF system in the First HAREM corpus (Santos & Cardoso, 2007), a Portuguese corpus manually annotated to develop and evaluate Portuguese NER systems. BCF is publicly available.

(2.3) MANUAL ANNOTATION

We performed manual annotation in a part of the PM-A corpus. We chose one of the large texts, aiming for a good number of examples for each class. Our choice of text was considered adequate, since the smallest number of examples in a class is 368 (as presented in [Table 1](#)). We followed the guidelines proposed by the HAREM project (Santos & Cardoso, 2007). We considered only three main categories, as previously explained. We made no distinctions between different sub-types of mentions. For instance, in HAREM, the PER category was also marked for sub-types regarding its kind as being an individual or an occupation, as in Carlos I or Sua Majestade (The King), respectively. For performing the manual annotation, two annotators, historians, used the INCEPTION open-source tool (Klie et al., 2018).⁷

CATEGORY	NUMBER
PER	474
LOC	514
ORG	368
Total	1356

Table 1 Number of manually annotated items per class.⁹

[Table 1](#) presents the number of manually annotated instances for each class. The two annotators agreement was measured based on Kappa statistics, usually employed for measuring annotation quality (McHugh, 2012), using the script provided by Python's sklearn (Pedregosa et al., 2011) library.⁸ The resulting Kappa was 0.71. Although higher agreement has been achieved for this type of annotation task, for this work we considered it satisfactory since our main goal was not to deliver a gold standard, but instead, to provide the automatic annotation of the collection. Another reason for accepting this agreement level is that the range 0.61–0.80 is, in general, considered as substantial agreement.

(3) EVALUATION, RESULTS, AND DISCUSSION

This section discusses the performance of the automatic annotation, based on the usual metrics for evaluating the NER systems: recall, precision, F-measure. For obtaining the metrics, we exported the manually annotated text in IOB tagging scheme, CoNLL-2002 (Sang & Erik, 2002). We used the original CoNLL-2002 Perl script.¹⁰ We present the results below. Regarding the number of annotated instances by the two systems ([Table 2](#) and [Table 3](#)), we can see that the class ORG is considerably smaller than the one in the manual annotation. One of the

⁵ <https://dumps.wikimedia.org/>.

⁶ <https://github.com/jneto04/ner-pt>.

⁷ <https://inception-project.github.io/>.

⁸ https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html.

⁹ For all tables, PER = person, LOC = location, ORG = organisation.

¹⁰ <https://www.clips.uantwerpen.be/conll2002/ner/bin/conllevl.txt>.

CATEGORY	NUMBER
PER	446
LOC	905
ORG	58
Total	1409

CATEGORY	NUMBER
PER	375
LOC	624
ORG	35
Total	1034

Table 2 Total annotated by spaCy.

Table 3 Total annotated by BCF.

systems (BCF) is more conservative than the other, resulting in a smaller number of instances in the three classes.

Tables 4 and **5** present the evaluation measures. Considering that the systems were trained on contemporary texts and have not been adapted for the 18th century language, and neither had new examples that could reflect the entity annotation differences, the results point to a fairly good performance. In general, considering all classes, we can see that the BCF system performed better in the identification and classification of named entities, being more conservative it achieved a better precision, with general F1 of 38% (spaCy) and 45% (BCF). Note that these systems resulted in a general F1 of 60.74% (spaCy) and F1 of 80.58% (BCF) when evaluated in a test set of contemporary Portuguese, the Mini-Harem golden corpus,¹¹ considering these same three categories.

CATEGORY	PRE	REC	F1
General	36.98%	39.98%	38.42%
LOC	33.59%	61.41%	43.43%
ORG	22.41%	3.57%	6.16%
PER	45.74%	45.95%	45.84%

Table 4 spaCy results.

CATEGORY	PRE	REC	F1
General	52.84%	40.34%	45.75%
LOC	46.81%	57.00%	51.40%
ORG	64.86%	6.52%	11.85%
PER	61.70%	48.43%	54.27%

Table 5 BCF results.

These differences in the results point to the impact of differences in the language regarding current vs 18th century language. Since uppercase is less consistent in these texts and there is a lot of orthographic variations, it affects both the nature of entities and context which is also used for the identification of the entities, for instance for hospital (ORG) we have *ospital*, *ispital*, *espital*.

As expected, we confirmed the better performance of BCF, which might be due to the use of language models that provide high-quality representations based on context (Akbik et al., 2018; Santos et al., 2019a).

For the category LOC, BCF presents an F1 of 51%, for PER, 54%, with precision higher than recall for PER and recall higher for LOC. There is a remarkable fall in the recall for the ORG category in both cases, meaning that the system did not find many of the ORG cases.

¹¹ https://www.linguateca.pt/aval_conjunta/HAREM.

In the example below (from Monforte – Nossa Senhora da Graça, p. 1195), the context points to a kind of relation (paying money) that indicates that the referent is an ORG and not LOC.

A igreja tem de fabrica 12 mil reis e o Reverendo Prior tem obrigação de dar para a <LOC> Real Capela de Villa Vicoza </LOC>, vinte e sinco mil reis.
The amount of money the church has for works is twelve-thousand reis, and the Reverend Prior must give the <LOC> Royal Chapel of Villa Vicoza </LOC> twenty-six thousand reis.

Despite the difficulties, based on this analysis, we chose BCF to annotate the corpus. The corpus enriched with the automatic identification of three categories of named entities is described next.

(4) RESULTING DATASET AND ITS APPLICATIONS

This section presents a description of the enriched dataset provided with examples and gives some notes on entity extraction applications in history research.

(4.1) DESCRIPTION OF THE DATASET

The dataset, Parish Memories with named entities,¹² consists of 366 transcribed texts from the original handwritten collection, where each text contains the description of a parish. It amounts to approximately 650,000 word tokens and 35,000 word types. These texts are annotated with entities PER, LOC, ORG. The dataset also contains the list of the extracted entities for each text and a global list with all entities from the collection. All these lists have frequency counts of terms in each category, number of mentions in each text. The transcribed texts with no annotations are also included. We performed the annotation automatically, and the agreement of the system with the manual annotation was around 0.65, whereas the agreement between human annotators was 0.71. The manually annotated text used in the evaluation and the corresponding list of entities are provided in the dataset.

In total, the collection has 13,600 person mentions, 23,511 location mentions, and 1,321 organisation mentions. The total number of automatically identified mentions is 38,432. The annotation is given in two IOB formats: (i) CoNLL, which lists each word/token with both positive and negative tags, and (ii) the in-text tags in which positive tags follow the respective words. In both formats (B-CAT) identifies the first word of an entity belonging to category CAT and (I-CAT) identifies the continuation of an entity, as exemplified in *Table 6*. The dataset provides the manually annotated texts, used for evaluation purposes, in the CoNLL format. A better visualisation interface of the annotated texts can be obtained with the INCEpTION tool, as shown in *Figure 3*.

Table 6 Annotation formats.

CoNLL:	In text tags:
edificada 0	
na 0	
Estremadura B-LOC	
, 0	
e 0	... edificada na Estremadura <B-LOC> , e
vizinhancas 0	vizinhancas de Castella <B-LOC> , em distancia
de 0	de huma legoa , na Provincia <B-LOC> do <I-LOC>
Castella B-LOC	Alenteio <I-LOC>
, 0	
em 0	
distancia 0	
de 0	
huma 0	
legoa 0	
, 0	
na 0	
Provincia B-LOC	
do I-LOC	
Alenteio I-LOC	

¹² <https://doi.org/10.5281/zenodo.4946479>.

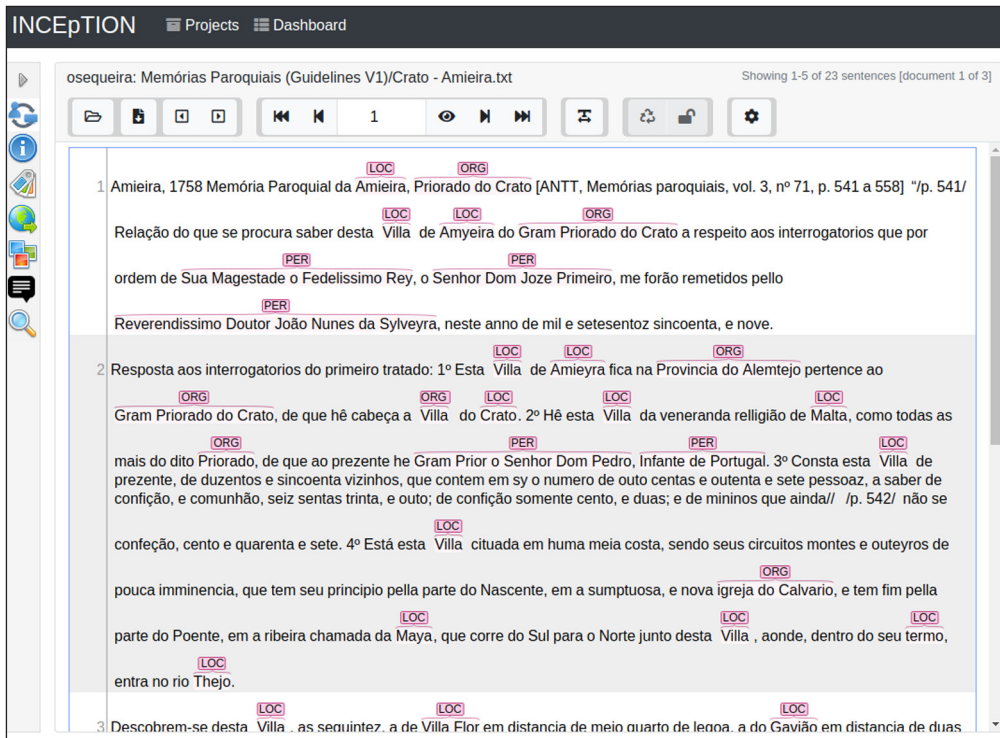


Figure 3 Annotated Entities in the INCEpTION Tool.

Table 7 shows examples of automatically extracted entities for each category, while **Table 8** illustrates some of the most frequent expressions referring to entities. For counting the global frequencies, we normalised words to lowercase and removed diacritics. However, there are other spelling variant issues to deal with for further statistical analysis (e.g., *matris/matriz*).

CATEGORY:PER	CATEGORY:LOC	CATEGORY:ORG
Cezar	Alandroal	Boa Nova
Christo	Alentejo	Câmara Municipal
Deos	Aviz	Companhia de Jesus
Divus Augustus Cesar	Elvas	Conselho
Eliza	Espanha	Coroa
el-Rey D. Deniz	Estremadura	Fontalia
Francisco de Freytas	Lisboa	Igreja Parochial
Lourenço Affonso	Nascente	Matris
Mouro	Norte	Matriz
Rey de Portugal	Occidente	Ordem da Comarca
Salvator Mundi	Poente	Senado da Camara
Sancta Maria	Porta da Trayção	Torre dos Coelheyros
Senhor Rey D. Deniz	Reyno	Villa

Table 7 Examples of extracted entities.

In terms of using this type of resulting information, we can say that these frequencies confirm the impact of the two most prominent characters in Ancien Régime society: God and the King. Regarding LOC entities, the domination of “villa” is very expressive and the main towns of the South are also present as well as Lisbon, since the distance from each village to the Court is mentioned in every text. Addressing the organisations, “villa” is also dominant. The disambiguation among these two uses of the term is indeed a challenging factor for the annotations.

CATEGORY: PER	#	CATEGORY: LOC	#	CATEGORY: ORG	#
deos	163	villa	1592	villa	150
sua magestade	128	evora	552	igreja matris	41
vossa excelencia	87	norte	409	igreja parochial	29
parochio	77	sul	375	camara	28
santo antonio	71	lisboa	360	parochia	26
deus	66	poente	355	conselho	25
santos	62	nascente	300	corte	23
el rey	60	beja	255	coroa	20
sao pedro	58	elvas	214	villa de monforte	19
juis	56	reyno	192	igreja matriz	16

Table 8 Some of the most frequent entities in each category.

(4.2) APPLICATIONS

Available in an open-access format, this corpus provides both the list of extracted entities and the entities annotated in-text, along with its original textual context. Keeping context is essential for historians to decode the precise meaning of a word and to assess the relevance of historical data, and this approach allows us to preserve it.

Entities, once identified, can be linked to other information, and texts can be reused by other researchers and projects. When we tag historical sources, enabling better document retrieval and information extraction, we improve the capacity of analysis. It is possible to find correlations between entities faster than manually or even to find new ones that would not be found otherwise. In this way, when we extract named entities from this kind of corpora, we can cross this information with other datasets from the same period, working with a larger number of sources. For instance, the PER category can be linked to prosopographical data, making it easier to identify biographical aspects, like birth or burial places, among others. This type of approach allows us to extract data to compare and carry out different analyses.

Regarding the connection with geographic information systems, although we are not aware of the exact borders of all the *freguesias* in the corpus, we know the exact location of the most important villages and the locations of some points of interest referred to in the texts (e.g., archaeological ruins). By mapping this data with artificial grids (lattices) (Birch et al., 2007) for representing the regions geographically, we can compare the parishes with each other to verify the institutions' homogeneity in the various geographical areas. We can also find the most valued institutions of that time through the frequency of citations in certain regions, clarifying the profile: ecclesiastic, secular, or mixed. Inspired by central-place theory (Christaller & Baskin, 1966; Romão, 2019) that measures the influence area of a geographical point, we can also verify the area of influence of a parish, organisation, or person by measuring the distance over where it is mentioned in other Parish Memories. These are only a few of the many possibilities of research that can be developed through this corpus.

We acknowledge that the data, being entirely automatically produced, is far from perfect, especially for the class ORG, which is rather incomplete. The idea is to provide some input, even if there is missing information; the findings brought about by the extraction may serve as partial evidence when seen through the lens of the historians. For instance, even for the case of ORG, which is low in recall, by looking at the data we can observe the prevalence of churches, councils, counties, brotherhoods, hospitals, and parishes. The material made available can be further improved by other research groups interested in either the material themselves or in the technical challenges for improving it. This first release serves as silver data to train other NER systems that may be more apt for texts from the classical period.

(5) CONCLUDING REMARKS AND FUTURE WORK

This paper presents a research resource for Digital History, a transcribed subset of originally handwritten texts from the 18th century, describing the geography and economy of Alentejo in

Portugal, which is now enriched with the annotation of named entities in three categories: PER, LOC, and ORG. The annotation was made automatically for the whole corpus, whereas a part of it was manually annotated for evaluation purposes. The dataset provides both automatic and manual annotations.

There are similar studies made for other languages, including Hubková et al. (2020) who present a study for named entities in a Czech historical corpus. In our case, it is a Portuguese historical corpus, which can be useful not only for historians, but also for architects, demographers, territory administrators, and planners.

As future work, we plan to develop further manual annotations that consider new categories, such as time, occupations, and social categories of people, which are crucial distinctions for historical research purposes. The last of which intends to adjust these labels to the society of the time and its legal system, strongly marked by the inequality of each person or group before the law. Based on the difficulties of distinguishing ORG and LOC, we will consider the GPE class (geopolitical entities for municipalities, countries, and others), first introduced in the ACE¹³ project, mainly meant to overcome (or ignore) the metonymy problem between ORG and LOC.

We aim to develop new and specific guidelines based on the lessons learned with this first manual annotation. We will improve the manual annotation process with discussions between annotators and will add a curation phase as we had a reasonable difference in the number of annotated entities among the human annotators. To improve the annotation quality, it is essential to create new guidelines adequate to the classical period of Portuguese language. With such improved annotation, we can tune the system with the new annotated categories. A new LM that includes 18th century texts with a detailed spelling variants description is also relevant to building a more robust system. We believe that future more robust NER system can be equally helpful to other 18th century handwriting sources and perhaps to earlier times.

There are problems with variations in spelling to be tackled in order to achieve better statistics and indexation. One of the solutions can be to lemmatise variants or to annotate variation, indexing all variants to the correspondent lemma. Examples of variants are: *igreja matris/igreja matriz, parochia/parrochia*. This pre-processing task is very challenging (Baron & Rayson, 2008; Dereza, 2018). There are some tools for classic languages, like CLTK: the classical language toolkit,¹⁴ but concerning the classic period of the Portuguese language, existing tools¹⁵ still need to be trained for this period.

From here we can link the Parish Memories to other sources, such as the book *Corografia Portuguesa* (Costa, 1706–1712), which contains data for the same region about the foundations of the cities and convents, bishops' catalogues, comments on illustrious men and genealogies of noble families, topographic and nature descriptions, and other observations.

Another possibility for future study is to identify other entities and concepts particular to specific domain studies, such as more fine-grained location categorisation with a focus on mountains and rivers, or mapping the descriptive data of the destruction caused by the Lisbon earthquake in 1755, for an Earthquake Damage Assessment. Studies from the corpus can serve to answer questions such as: Do the descriptions in the corpus reflect current scientific data? Can they be corroborated with archaeological data?

The INCEpTION annotation tool (Klie et al., 2018) includes semantic annotation (e.g., concept linking, fact linking, knowledge base population, semantic frame annotation) which are relevant features for historians. In this stage of the study, we only used it for annotating the three categories mentioned above; other features of the tool are relevant for the future development just described.

SUPPLEMENTARY FILES

Supplementary File 1: Parish Memories with Named Entities Dataset. ParishMemorieswithNEs_V2.zip. DOI: <https://doi.org/10.5281/zenodo.4946479>

13 <https://www ldc.upenn.edu/collaborations/past-projects/ace>.

14 <https://github.com/cltk/cltk>.

15 <http://lxcenter.di.fc.ul.pt/services/pt/LXServicesLemmatizerPT.html>.

1) README

Parish Memories with Named Entities (ParishMemorieswithNEs) contains digitized, transcribed texts from 1758 Portuguese parochial surveys. This collection refers to the surveys from the Alentejo region. Here they are provided with annotation of named entities (person, organization, location). The annotation was done automatically, it is therefore incomplete and not totally precise, but potentially useful and can be further improved by other research groups, either interested in the material themselves or in the technical challenges for improving it. All the issues are discussed in more detail in the paper.

2) NE-Totals

- GlobalTotal: all classified elements with frequencies for each category
- TextTotals: number of classified elements and number of occurrences for each text

3) Texts+Entities(AutomaticAnnotation): 366 automatically annotated with NEs

- PM-1: for each parish memory PM
 - * PM1.txt (original texts)
 - * ptTagged-PM1(NE tagged texts)
 - * CoNLL-PM1(NE CoNLL format)
 - * Named_Enties-PM1 (list of named entities)
- PM-2: same as above for PM2
- PM-366: same as above for PM366

obs: empty lines in CoNLL files correspond to original document breaks

4) Texts+Entities(ManualAnnotation): manually annotated texts with NEs

- ManualAnnot_CoNLL-PM.txt
- ManualAnnot_Named_Entities-PM.txt
- PM_Manual_SourceText.txt

FUNDING INFORMATION

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

AUTHOR CONTRIBUTIONS

Renata Vieira: conceptualisation, supervision, methodology, and writing original draft, review and editing.

Fernanda Olival: conceptualisation, data curation, supervision, writing original draft, review and editing.

Helena Freire Cameron: conceptualisation, supervision, writing original draft, review and editing.

Joaquim Santos: software and investigation.

Ofélia Sequeira: data curation and investigation.


Ivo Santos: resources, data curation and writing original draft.


COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Renata Vieira  orcid.org/0000-0003-2449-5477
CIDEHUS, University of Évora, Portugal

Fernanda Olival  orcid.org/0000-0003-4762-3451
CIDEHUS, University of Évora, Portugal; Department of History – University of Évora, Portugal

Helena Freire Cameron  orcid.org/0000-0001-7719-6994
CIDEHUS, University of Évora, Portugal; VALORIZA-Polytechnics of Portalegre, Portugal

REFERENCES

- Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fé, USA.
- Baron, A., & Rayson, P. (2008). VARD2: A tool for dealing with spelling variation in historical corpora. *Postgraduate Conference in Corpus Linguistics*, Birmingham, U.K.
- Birch, C., Oom S., & Beecham, J. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling*, 206(3–4), 347–359. DOI: <https://doi.org/10.1016/j.ecolmodel.2007.03.041>
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. DOI: https://doi.org/10.1162/tacl_a_00051
- Cameron, H. F., Gonçalves, M. F., & Quresma, P. (2020). Linguistic and orthographical classic Portuguese variants challenges for NLP. *Proceedings of the 14th International Conference on the Computational Processing of Portuguese*, Évora, Portugal.
- Chiu, J. P., & Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4, 357–370. DOI: https://doi.org/10.1162/tacl_a_00104
- Chorão, M. J. M. B. (1987). Inquéritos promovidos pela Coroa no século XVIII. *Revista de História Económica e Social*, 1^a série, 21, 93–130.
- Christaller, W., & Baskin, C. W. (1966). *Central places in southern Germany*. Engelwood Cliffs, NJ: Prentice-Hall.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12, 2493–2537. Aug.
- Consoli, B. S., Santos, J., Gomes, D., Cordeiro, F., Vieira, R., & Moreira, V. (2020). Embeddings for named entity recognition in geoscience Portuguese literature. *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France.
- Costa, A. C. (1706–1712). *Corografia Portuguesa 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, Tomo primeyro[-terceyro]*, vol. 1-2-3. Lisboa: na officina de Valentim da Costa Deslandes.
- Dereza, O. (2018). Lemmatization for ancient languages: Rules or neural networks? In D. Ustalov, A. Filchenkov, L. Pivovarova & J. Žižka (Eds.), *Artificial Intelligence and Natural Language. AINL 2018. Communications in Computer and Information Science*, vol 930. Springer. DOI: https://doi.org/10.1007/978-3-030-01204-5_4
- dos Santos, C. N., & Guimarães, V. (2015). Boosting named entity recognition with neural character embeddings. *Proceedings of the 5th Named entity workshop*, Beijing, China. DOI: <https://doi.org/10.18653/v1/W15-3904>
- Gonçalves, M. F. (2003). *As ideias ortográficas em Portugal: de Madureira Feijó a Gonçalves Viana (1734–1911)*. Lisboa: Fundação Calouste Gulbenkian.
- Hubková, H., Kral, P., & Pettersson, E. (2020). Czech historical named entity corpus v 1.0. *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille: France.
- Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed. draft)*. Stanford University.
- Kemmler, R. (2001). Para uma história da ortografia portuguesa: o texto metaortográfico e a sua periodização do século xvi até a reforma ortográfica de 1911. *Lusorama. Zeitschrift für Lusitanistik. Revista de Estudos sobre os Países de Língua Portuguesa*, 47–48, 128–319.
- Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., & Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, Santa Fé, USA.
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). *Neural architectures for named entity recognition*. arXiv preprint arXiv:1603.01360.
- Mateus, M. H. M., & Cardeira, E. (2007). *Norma e variação*. Alfragide: Editorial Caminho.
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica*, 22(3), 276–282. DOI: <https://doi.org/10.11613/BM.2012.031>

- Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J. R.** (2013). Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194, 151–175. DOI: <https://doi.org/10.1016/j.artint.2012.03.006>
- Ortiz Suárez, P. J., Romary, L., & Sagot, B.** (2020). A monolingual approach to contextualised word embeddings for mid-resource languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. DOI: <https://doi.org/10.18653/v1/2020.acl-main.156>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É.** (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85), 2825–2830.
- Romão, R. M. A.** (2019). *Os lugares centrais em Portugal: a área de influência de Coimbra*. Masters Thesis, Instituto Superior de Economia e Gestão.
- Romein, C. A., Kemman, M., Birkholz, J. M., Baker, J., DeGrujter, M., Meroño-Peñuela, A., Ries, T., Ros, R., & Scagliola, S.** (2020). State of the field: digital history. *History*, 105(365), 291–312. DOI: <https://doi.org/10.1111/1468-229X.12969>
- Sang, T. K., & Erik, F.** (2002). Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. *Proceedings of CoNLL-2002, Conference on Natural Language Learning*, Taipei, Taiwan.
- Santos, J. A.** (1995). *As freguesias: História e actualidade*. Oeiras: Celta.
- Santos, D., & Cardoso, N.** (2007). *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Retrieved from https://www.linguateca.pt/aval_conjunta/LivroHAREM/Livro-SantosCardoso2007.pdf
- Santos, J., Consoli, B., dos Santos, C., Terra, J., Collonini, S., & Vieira, R.** (2019a). Assessing the impact of contextual embeddings for Portuguese named entity recognition. *Proceedings of the 8th Brazilian Conference on Intelligent Systems*, Salvador, Brasil.
- Santos, J., dos Santos, H. D. P., & Vieira, R.** (2020b). Fall detection in clinical notes using language models and token classifier. *Proceedings of the 33rd International Symposium on Computer-Based Medical Systems, CBMS 2020*, Rochester, USA.
- Santos, I., Olival, F., & Sequeira, O.** (2020a). Excavating the data pit: the Portuguese parish memories (1758) as a gold standard. *DHandNLP@PROPOR, Workshop on Digital Humanities and Natural Language Processing*, Évora, Portugal.
- Santos, J., Terra, J., Consoli, B. S., & Vieira, R.** (2019b). Multidomain contextual embeddings for named entity recognition. *Proceedings of the 35th Conference of the Spanish society for natural language processing*, Bilbao, Spain.
- Schmitt, X., Kubler, S., Robert, J., Papadakis, M., & Le-Traon, Y.** (2019). A replicable comparison study of NER software: Stanford NLP, NLTK, Open NLP, Spacy, Gate. *Sixth International Conference on Social Networks Analysis, Management and Security*, Granada, Spain.
- Souza, F., Nogueira R., & Lotufo, R.** (2019). *Portuguese named entity recognition using BERT-CRF*. arXiv preprint arXiv:1909.10649.
- Verdelho, T.** (1987). Latinização na história da Língua Portuguesa – o testemunho dos dicionários. *Arquivos do Centro Cultural Português* (volume de homenagem a Paul Teyssier), XXIII, 157–187.

TO CITE THIS ARTICLE:

Vieira, R., Olival, F., Cameron, H. F., Santos, J., Sequeira, O., & Santos, I. (2021). Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities. *Journal of Open Humanities Data*, 7: 20, pp. 1–13. DOI: <https://doi.org/10.5334/johd.43>

Published: 09 September 2021

COPYRIGHT:

© 2021 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.