Digital Humanities and Portuguese Processing: a research pathway [⋆]

Renata Vieira[1][0000−0003−2449−5477], Ana P. Banza[1][0000−0003−4467−9521], Ana S. Ribeiro[1][0000−0002−1822−5908], Cassia Trojahn[2][0000−0003−2840−005X], Fernanda Olival[1][0000−0003−4762−3451], Helena Cameron[3][0000−0001−7719−6894], Hermínia Vilar[1][0000−0003−3300−8335], Ivo Santos[1][0000−0001−5152−6027], Joaquim Santos[1][0000−0002−0581−4092], Maria F. Gonçalves[1][0000−0001−8262−6514], Maria J. Finatto[4][0000−0002−6022−8408], and Paulo Quaresma[5][0000−0002−5086−059X]

[1] CIDEHUS,Universidade de Évora
[2] IRIT, Université de Toulouse
[3] CIDEHUS, Instituto Politécnico de Portalegre
[4] PPGLA, Universidade Federal do Rio Grande do Sul
[5] Department of Computer Science, Universidade de Évora

Abstract. This paper reflects on the whole path of work in digital humanities, on the light of the projects related to text processing under development at CIDEHUS. These projects deal with a rich heritage related to the Portuguese culture, history and language. This paper reflects on the many challenges to be faced and how NLP techniques may broaden the capabilities of organising and sharing knowledge related to these resources.

Keywords: Natural language processing · Portuguese language · Digital Humanities.

1 Introduction

This paper presents and discusses the challenges faced by projects under development at CIDEHUS (Center for History, Culture and Societies), in the area of Digital Humanities, particularly those related to text processing. The main point in common in the interdisciplinary projects we discuss here is having Portuguese texts as their primary knowledge source, mostly pre-contemporary ones. The differences are mainly the historical periods of the sources, whether manuscripts or printed material, and their stage of digitisation, which varies from digital images, PDF texts and digitised texts. The paper presents an overview of the collections that are of interest. Then, we discuss the common starting points to deal with not yet digitised material, and then discuss the first organisation

[⋆] Copyright © 2022 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). Supported by by Portuguese Foundation for Science and Technology (FCT) under the projects CEECIND/01997/2017, UIDB/00057/2020.

requirements once they are digitised or transcribed. Next, we present the more advanced processing undertaken in some of the collections under study, as well as discuss current research goals and how NLP is required or involved. At last, we present our concluding remarks.

2 DH at CIDEHUS: dealing with sources from the 15th century to date

The Centre for History, Culture and Societies (CIDEHUS) is an interdisciplinary research centre of the University of Évora. It develops research in the intersection of heritage, language, history, social sciences and more. CIDEHUS has been conducting research in the area of Digital Humanities, dealing with texts, 3d models, georeference, maps, music, tourism experiences etc. In this paper we focus on those research in the area of DH which are more closely related to text collections. These collections under analysis cover a broad range period, as listed below:

- From the 14ht to the 16th century, "Cortes Portuguesas" is a source under analysis by Hermínia Vilar.
- From the 16th and 17th centuries: (i) the letters collected in "Livro das Monções" are studied in a project led by Ana Sofia Ribeiro [13]; (ii) António Vieira´s unfinished work ("História do Futuro") is studied by Ana Paula Banza [2].
- From the 18th-century, we have: (i) the Parish Memories, with related projects led by Fernanda Olival [15] [11]; (ii) the medical books and treatises of Curvo Semedo, studied in a project led by Maria Finatto [4] [12] and (iii) the first Portuguese nursery handbook "Postilla Religiosa, e Arte de Enfermeiros" (1741) studied by Filomena Gonçalves [5].
- Addressing contemporary scientific literature, we have the research conducted by Ivo Santos based on archaeology reports [16]

The heterogeneity of these sources poses a series of challenges, from digitisation up to natural language processing and the organisation of the acquired knowledge in semantically structured databases. In the next, section we will discuss these challenges.

3 Initial steps

DH projects may benefit from current AI and NLP techniques for better access to the sources, including digitisation, the addition of metadata, information extraction, knowledge representation techniques, annotated corpora, creation of data sets and linked data. NLP and AI methods, however, must be adapted to different needs, and better user interfaces are required for the developed methods to be used out of the context of programming frameworks. We start by presenting the initial steps for DH projects, which are full of challenges themselves.

3.1 Starting points: digitisation

OCR technology is often applied for sources available in PDF files. In our current projects, we have the Monsoon letter ("O livro das Monções"), the books and treatises of Curvo Semedo, and some of the Archaeology reports at this stage. OCR output quality varies a lot depending on the format and quality of the input. It is a basic but not solved problem [19]. It is often the case that OCR outputs must pass through extra processing or manual correction to make the source fit for the next steps.

Besides the challenges posed by OCR quality, there are projects that need digitisation from original manuscripts, which is yet a different problem. Transcription tools such as Transkribus [7] help digitisation in these cases.

3.2 Processing digitised sources

Once a source, or corpus of study, is already digitised, the organisation of its metadata is of great relevance. The digital material must identify itself well, that is, it should inform to which collection it belongs, what is the place of each file, and so on. The metadata can also describe the document structure, when required. It is also important to separate metadata from the actual data. For transcribed material, for instance, one should identify headers, page numbering or comments in an organised way. This is essential for further processing. Metadata is also important to connect a source with other sources, and connect them to the linked open data [10].

Beyond metadata organisation, there is the problem of normalisation. In Digital Humanities it is often the case that the sources, being from a distant time, present grammatical variations, both spelling and morpho-syntactic. Understanding these differences and being able to translate or associate ancient writings to the current standards is an essential step towards other processing levels [3].

Another way of mitigating the writing variants is to create language models that include substantial corpora from other time periods, with their naturally occurring variants or to add a final training phase (tuning) to adapt the model to the variants [1].

From this point, with digitised, normalised texts or with suitable language models, we can make better use of current NLP tools that are already developed, such as automatic translation, named entity recognition, event extraction, correference resolution, question answering, and many others. However, such developments require not only current NLP tools, but also closer interaction with scholars, for the final adaptations for their needs and suitable interfaces.

4 Next (more advanced) steps

The DH group at CIDEHUS has advanced in processing some of its collections, using AI and natural language processing tools to create knowledge bases that may be useful for other researchers. Further ahead we describe some of the developed resources.

4.1 Resources under development

The Parish Memories One example is the work done with the Parish Memories. It is a rich collection, very well studied in Portugal. The digitised version of microfilms from the originals are available at the Portuguese National Archive (Arquivo Nacional da Torre do Tombo). There are also many printed books reproducing parts of the material. A first digital version was made freely available through CIDEHUS Digital Portal[6]. The application of NLP techniques is now possible due to past projects that worked on the manual transcription of the original manuscripts. From this collection, a named entity dataset was automatically built using previously developed systems for named entity recognition [17], based on machine learning techniques and language models. The initial entity categories considered were the usual, person, location, and organisation. The named entities extracted from the Parish Memories constituted a dataset that was made available to the community. The digitised texts are provided with their respective lists of named entities [20]. As it was based on a completely automated process, a curation phase for this data is still foreseen in the future developments of this source.

The Archaeology Corpus A Portuguese corpus is being built in the domain of archaeology. The main sources of information considered are reports of archaeological works, academic theses and specialised bibliography. Among these sources, the “Portal do Arqueólogo” (Archaeological Data Management Tool) stands out for housing structured information. In this portal, the information is distributed in three main groups: sites, works and projects. A project can be made up of several archaeological works, and a work can refer to one or more sites. As of June 2021, this corpus had 36275 records of archaeological sites and 39947 works. The goal is to extract and organise information using NLP methods. Some analysis made on this corpus allows identifying periods in time with more intensive archaeological work and many other aspects of archaeology in Portugal. This initial corpus and the analyses made upon it is described in more detail in [16].

4.2 Further research goals

The various projects of the group have different processing needs. For the work being developed with the writings of Padre António Vieira, História do Futuro [2], for instance, the assessment of semantic similarity [9] is quite significant. Although semantic similarity is a well-developed problem, finding a suitable, applicable tool and creating helpful interfaces for the source analysis are some of the challenges.

Event Extraction is another interesting NLP task, with resources developed for Portuguese [14]. The challenge, in this case, is to adapt these tools to the

[6] http://www.cidehusdigital.uevora.pt

language of the period, an effort under exploration in the project related to the Monsoon letters [13].

Ontology linking is one of the goals of the Curvo Semedo project [12], where health terminology can be found and mapped to existing ontologies [18]. These mappings are a way of enriching the resource and establishing what areas of medicine, anatomy and which kind of diseases and medications were known from that time, according to that source. Currently, the vocabulary - terminologies and related expressions - used within Semedo's manuals are contrasted with those used along with the nursery manual printed in 1741 [5]. This contrast may be useful to identify different perspectives, concepts and behaviours of the various health care professionals and practitioners at that time.

Regarding the source of "Cortes Portuguesas", one of the purposes is to identify and analyse concepts occurring in the speeches of the Courts and the argumentation used in the royal legislation. The quest is to grasp the origins of an ethics of behaviour of royal officers, to the establishment of control and accountability systems and the control of corruption in low-middle-age societies. On top of the other mentioned task, argumentation mining [8] might be the kind of processing that may accelerate the analysis made by the scholar in this kind of project.

In fact, all these different tasks are complementary, and one ambitious goal would be a unified environment combining them to study these sources and others. On the other hand, there are many other possibilities, as dealing with textual information towards meaning is an endless effort.

4.3 FAIR Data and Ontology Development

Naturally, with the evolution of the studies mentioned here, data will be produced, as was the case for the named entities in the Parish Memories [20]. The efforts of sharing data has a long way to go in terms of standardisation. It is very important to make data compliant to the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) [21]. Also, an essential step towards improving FAIRness of data is using vocabularies and ontologies for data and metadata representation [6]. While diverse vocabularies are available for metadata representation (as DCAT, PROV-O, etc.), extending existing domain ontologies or developing new ones to better fit the specificities of each corpus or source is required.

5 Concluding Remarks

CIDEHUS has a rich portfolio of projects with a great potential to be explored through the advancements of the area of Digital Humanities. The ideal is that all collections in this portfolio follow the same digitisation standards, such as those proposed by TEI (the Text Encoding Initiative) and the FAIR principles of Open Data. AI techniques are needed for treating the problems in all phases, since those initial steps related to OCR quality, manuscripts transcription, the addition of metadata, and normalisation as well as the processing phases required for translation, information retrieval and extraction, creation of knowledge basis, and their association to ontologies. It is very crucial that a human centred AI perspective must be taken into account to provide suitable user interfaces for accessing the sources and the extracted data. All these projects and collections, even with different goals, may gain by coexisting in the same environment and thus sharing the use of the same powerful tools to deal with texts and their encoded knowledge.

References

1. Arevalo, E.M., Fonteyn, L.: Macberth: Development and evaluation of a historically pre-trained language model for english (1450-1950). In: ICON Workshop on Natural Language Processing for Digital Humanities (2021)
2. Banza, A.P.: A edição digital da história do futuro, de antónio vieira: arquivo e ferramentas. In: Actas da Jornada de Humanidades Digitais do CIDEHUS (to appear) (2022)
3. Cameron, H.F., Gonçalves, M.F., Quaresma, P.: Linguistic and orthographical classic portuguese variants challenges for NLP. In: Proceedings of the 14th International Conference on the Computational Processing of Portuguese. pp. 43–48 (2020)
4. Finatto, M.J.B., Quaresma, P., Gonçalves, M.F.: Portuguese corpora of the 18th century: old medicine texts for teaching and research activities. In: Proceedings of the conference on Language Technologies Digital Humanities (2018)
5. Gonçalves, M.F.: A arte de enfermeiros (1741): aspetos do léxico relativo a doenças e remédios no século xviii. Revista Panace@ **21**(52) (2020)
6. Guizzardi, G.: Ontology, Ontologies and the "I" of FAIR. Data Int. **2**(1-2), 181–191 (2020)
7. Kahle, P., Colutto, S., Hackl, G., Mühlberger, G.: Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 4, pp. 19–24. IEEE (2017)
8. Lawrence, J., Reed, C.: Argument mining: a survey. Computational Linguistics **45**(4), 765–818 (2020)
9. Lopez-Gazpio, I., Maritxalar, M., Gonzalez-Agirre, A., Rigau, G., Uria, L., Agirre, E.: Interpretable semantic textual similarity: Finding and explaining differences between sentences. Knowledge-Based Systems **119**, 186–199 (2017)
10. Nair, S.S., Jeeven, V.: A brief overview of metadata formats. DESIDOC Journal of Library & Information Technology **24**(4) (2004)
11. Olival, F., Cameron, H., Vieira, R.: As memórias paroquiais: do manuscrito ao digital. In: Actas da Jornada de Humanidades Digitais do CIDEHUS (to appear) (2022)
12. Quaresma, P., Finatto, M.J.: Information extraction from historical texts: a case study. In: Workshop on Digital Humanities and Natural Language Processing, collocated with PROPOR) (2020)
13. Ribeiro, A.S.: O projecto monsoon: perspectivas digitais da Índia portuguesa. In: Actas da Jornada de Humanidades Digitais do CIDEHUS (to appear) (2022)

14. Sacramento, A.d.S.B., Souza, M.: Joint event extraction with contextualized word embeddings for the portuguese language. In: Brazilian Conference on Intelligent Systems. pp. 496–510. Springer (2021)
15. Santos, I., Olival, F., Sequeira, O.: Excavating the data pit: the portuguese parish memories (1758) as a gold standard. In: Workshop on Digital Humanities and Natural Language Processing, collocated with PROPOR) (2020)
16. Santos, I., Vieira, R.: Semantic information extraction in archaeology: Challenges in the construction of a portuguese corpus of megalithism. In: 15th International Conference on Metadata and Semantics Research, Springer Communications in Computer and Information Science Series, Vol. 1537. (2021)
17. Santos, J., Consoli, B., dos Santos, C., Terra, J., Collonini, S., Vieira, R.: Assessing the impact of contextual embeddings for portuguese named entity recognition. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS). pp. 437–442. IEEE (2019)
18. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A.: Disease ontology: a backbone for disease semantic integration. Nucleic acids research **40**(D1), D940–D946 (2012)
19. van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the impact of ocr quality on downstream nlp tasks. In: ICAART 2020 - Proceedings of the 12th International Conference on Agents and Artificial Intelligence (2020)
20. Vieira, R., Olival, F., Cameron, H., Santos, J., Sequeira, O., Santos, I.: Enriching the 1758 portuguese parish memories (alentejo) with named entities. Journal of Open Humanities Data **7**, 20 (2021)
21. Wilkinson, M., Dumontier, M., Aalbersberg, e.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)