Linguistic and orthographical classic Portuguese variants

Challenges for NLP

Helena Freire Cameron[1], Maria Filomena Gonçalves[2], Paulo Quaresma[3]

[1] Instituto Politécnico de Portalegre - Escola Superior de Tecnologia e Gestão, Campus Politécnico, 10, 7300-555 Portalegre, Portugal
helenac@ipportalegre.pt

[2] Universidade de Évora, ECS/Department of Linguistic and Literatures, CIDEHUS-UÉ/FCT, Largo dos Colegiais, 7002-554 Évora, Portugal
mfg@uevora.pt

[3] Universidade de Évora, ECS/Department of Computer Science, Laboratory of Informatics, Systems and Parallelism, R. Romão Ramalho, 59, 7000 Évora, Portugal
pq@uevora.pt

Abstract. In recent times, it was made a great investment in transfer from physical ancient Portuguese texts to digital support. This support transfer allows not only the access to the texts, bringing them to the public in general, but also the possibility of texts to be readable and processed by machines. NLP tools are addressed, mainly, to contemporary Portuguese and the application of NLP to classic texts has several difficulties. The elaboration of big lexical corpora of forms previous to modern Portuguese is an opportunity for multidisciplinary field of studies allowing the enlargement of linguistic studies and also the possibility of obtaining, by NLP, validated corpora, collections and ontologies, that can be input in NLP tools for ancient Portuguese texts. In this work we will present, briefly, the problem of lexical variation of forms in processing classic Portuguese texts, the challenges that emerge from them and future perspectives of work.

Keywords: classic Portuguese texts, NLP, linguistic variation.

1 Introduction

The aim of this paper is to demonstrate the problem of lexical variation in classic period of Portuguese in Natural Language Processing (NLP) of ancient Portuguese texts. Therefore, the main purpose of this work is to present a preliminary essay of a major work in systematization of lexical variants, orthographical, typographical, his-

torical or semantic, and proposal of equivalence/grouping of forms. The typology of lexical variation will enable a more accurate processing of these texts, as there are no known NLP tools able to process any classic Portuguese texts fully automatically.

2 Some challenges to NLP in classic Portuguese texts

Lately, it has been done a great effort to transfer textual documents booked in Patrimonial Libraries to a digital support, bringing them from restricted access bookshelves and provide them to public in general. With this investment and the progress in Natural Language studies, the access to the "inside" of these texts is a very interesting field of study for Digital Humanities.

Digital Humanities (DH) is an area of confluence of Computer Sciences and Language Sciences (Schreibman, Siemens, & Unsworth, 2004). Nowadays, DH constitute a growing and challenging field of knowledge concerning antique books (Gonçalves & Banza, 2013). Also, automatic reading of texts has been developed by computational tools for Natural Language Processing for the Portuguese (Quaresma, 2013), allowing automatic access to the text. However, the application of these tools to ancient texts brings new challenges to automatic text reading or text processing (Finatto, Quaresma, & Gonçalves, 2018). Automatic character recognition, in most of known tools for the Portuguese, is still not properly supported by updated lexicon that is contemporary to the forms pretended to be recognized in texts. On the other hand, the lexical forms in ancient texts have specific characteristics that must be annotated so, when automatic NLP procedures occur, they can be properly analysed and processed.

Some tools, the VARD2 and the TICCL, have been adapted and used to Portuguese in ancient texts, ported to the Portuguese, discussed by (Reynaert, at allii, 2012).

Recently, the CARDS-FLY project (Marquilhas & Hendrickx, 2014) has collected and transcribed historical Portuguese personal letters from the 16th to the 19th century in a digital support, constituting a *corpus* of 2000 letters, the CARD corpus, and, in sequence, enlarged with the FLY *corpus* of 20th century personal papers. All the texts have been transcribed by hand and the project already presents results of automatic spelling normalization in this *corpus*.

The classic linguistic period of Portuguese is very rich in variation. It has remarkable lexical renovation, with the entry of many new terms in Portuguese language (Verdelho, 1987). Also, classic lexical Portuguese forms have great variation, in linguistic level, coexisting archaic and renovated forms (Teyssier, 1997) and, in graphical level, with many orthographical and typographical variants (Gonçalves M. F., 2003). In the classic period of Portuguese, some linguistic phenomena stabilize, and some lexical forms are close to contemporary forms; others still maintain archaic forms, coexisting with renovated forms, sometimes, even in a same text document.

Concerning Portuguese orthography of classic period, as there was not still a standardization, the variation is very expressive, for example, in nasal diphthongs, in

the vocalic or consonantal use of <j> and <v>, in pseudo-etymological spelling or in the use of double consonants, sometimes in a very "creative" registration of Portuguese language.

Also, the printing press process at that time produced a great variety of different spelling, as printers, limited by the availability of typography types or the lack of linguistic criteria, use freely allographs, increasing the variation of Portuguese forms.

The description and registration of the lexical variants of this period of the Portuguese language is fundamental for tasks of pre-processing and post-processing of texts in this period.

3 An example of lexical variation in classic Portuguese in dictionaries (classic period)

The lexical variation in a text is noticed by a human reader but, in an automatic processing, the variants are treated as autonomous forms when, in fact, they are related, linked by the history of language. The example of the forms *giolho*, *geolho*, *joelho*, and *juelho*, all used during the classic period of Portuguese, present, clearly, a good example of this linguistic variation. The search of each of these forms in *Corpus Lexicográfico do Português* produces interesting results, showing the lexicographical and textual registration of forms across time.

The form *giolho* is used by Jerónimo Cardoso (1569-70), Bento Pereira (1697), Bluteau (1712-1728) and Madureira Feijó (1734) and others.

However, *geolho* is only registered by Bento Pereira, in the Latin-Portuguese dictionary *Prosodia* (1697). The lexical form *joelho* is not registered in this *Corpus* by Jerónimo Cardoso and, in Bento Pereira, it is only register in *Prosodia*. This form is register in all the subsequent dictionaries:

	Jerónimo Cardoso	**Bento Pereira**		**Bluteau**	**Fonseca**
	Dictionarium	*Tesouro*	*Prosodia*	*Vocabulario*	*Parvum lexicon*
giolho	5	1	--	2	--
geolho	--	--	5	--	--
joelho	--	1	23	120	15

Table 1. Occurrences of the forms *giolho*, *geolho* and *joelho* in *Corpus Lexicográfico do Português*, in (Cameron, 2012, p. 210)

Bluteau (1712- 1728) and Folqman (1755) dictionaries also register the form *juelho* and they are the only ones in this *Corpus* that register this variation. The forms

giolho, *geolho* and *juelho* are no longer active in contemporary language and they don't have actual lexicographical registration.

4 Linguistic variation in historical online corpora

The search in historical corpora is made with queries for each word separately. In *Corpus do Português* the forms *giolho* and *joelho* are bothe registered. The variants *geolho* and *juelho*, registered in dictionaries, are not mentioned in this *corpus*.

Concerning the *corpus Tesouro do léxico Patrimonial Galego e Português*, although it is not an historical *corpus* for the Portuguese, gives information about dialectal variants. The query for one of the forms gives back also all the equivalent forms and their location. The variants *giolho*, *joelhe*, *joelho* are found in this corpus.

5 Some considerations and future perspectives

The NLP of lexical variants in Classic Portuguese must assume, in pre-processing and/or post-processing of texts, among other steps, an exhaustive lexical description of variants made from dictionaries and a verification of other possible variants by use included in texts of the classic Portuguese. The registration of lexical forms in dictionaries is much more enlarged than the occurrences available in texts in actual online *corpora*. Probably, this is due to the fact that some *corpora* choose representative texts, and, in consequence, they may not have enough volume of different lexical forms. The probability of existence of a particular word in a *corpus* may depend much more of the amount and variety of texts that constitute the *corpus* than the probable existence of that lexical variant, itself. For that, only the combination of these two steps can contribute for a validation of lexical forms from the use in texts.

Before the great variance of Portuguese classic forms, the need of having historical vocabularies made with strong linguistic criteria is essential so vocabularies and ontologies that will be used in NLP of ancient texts may respond to the demands of processing and parsing.

Also, facing this range of variation of historic forms during time, when lexical variants appear in some texts in certain dates but are no longer used in other texts from posterior dates, the automatic recognition of variant lexical forms in a text may also validate the authenticity and probable date of documents.

We pretend to collect a lexical classic Portuguese *corpus*, from texts of that period, obtained automatically or semi automatically, using Optical Character Recognition (OCR). It will be processed and annotated, and it will allow a validated study of lexical variants in context, not only orthographical. This *corpus* will also help to a better optimization in character capture, supported by the results obtained.

To enlarge our lexicon in pre-processing tasks, we collected a list of 46 000 words from the linguistic corpus of the Latin-Portuguese-Latin dictionary *Prosodia*, from its 7th edition of 1697, one of the largest *corpora* from bilingual dictionaries in classic Portuguese. We transcribed the complete text to a Word document ans, in format .txt,

we processed it with AntConc© tool, using conventional information to separate Latin from Portuguese. This corpus, that was not lemmatized, is representative of Classic Portuguese (Cameron, 2012) and it will support, in first stage, the automatic capture of texts with OCR.

Although the application of NLP to classic Portuguese texts may need an initial effort in order to have validated vocabularies and ontologies, in future, this investment may provide very interesting results allowing automatic access to the inside of texts and all that may result of that.

References

1. Álvarez, R. (. (2009-). *Tesouro do léxico patrimonial galego e português*. (Santiago de Compostela: Instituto da Língua Galega) Obtido em 21 de dezembro de 2019, de http://ilg.usc.es/Tesouro
2. Anthony, L. (2014). AntConc (Version 3.4.4w) [Windows)]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software
3. Banza, A. P., & Gonçalves, M. (2018). *Roteiro de História da Língua Portuguesa.* (U. C.-H. Heritage, Ed.) Évora: Universidade de Évora.
4. Cameron, H. F. ([no prelo]). Desafios da aplicação da análise de *corpus* em dicionários antigos latim-português. *Atas do V Seminário de I&DT: novos desafios ... novas agendas de investigação*.
5. Cameron, H. F. (2012). *A Prosodia de Bento Pereira - Contributos para o estudo lexicográfico e filológico.* Aveiro: Universidade de Aveiro.
6. Cameron, H. F. (2018). *O conjunto lexicográfico Prosodia (1634-1750), de Bento Pereira, S.J.* (Vol. 7). Évora: Coleções do CIDEHUS, Coleção Fontes & Inventários - série geral, disponível em:https://books.openedition.org/cidehus/3321.
7. Cardeira, E. (2006). *O essencial sobre a História do Português.* Alfragide: Editorial Caminho.
8. Cardeira, E., & Mateus, M. (2008). *Norma e Variação.* Alfragide: Editorial Caminho.
9. Castro, I. (2006). *Introdução à História do Português* (2ª edição revista e muito ampliada ed.). Lisboa: Colibri.
10. *Corpus Lexicográfico do Português*. (s.d.). Obtained from http://clp.dlc.ua.pt/inicio.aspx
11. Davies, M., & Ferreira, M. (2006-). Corpus do Português: 45 million words, 1300s-1900s: Obtained from http://www.corpusdoportugues.org/hist-gen/
12. Finatto, M. J., Quaresma, P., & Gonçalves, M. (2018). Portuguese corpora of the 18th century: old medicine texts for teaching and research. In D. Fiser, & A. Pancur (Edits.), *Proceedings of the Conference on Language Technologies & Digital Humanities* (pp. 114-120). Ljubljana: University of Ljubljana, Slovenia, disponível em:http://hdl.handle.net/10174/23606.
13. Gonçalves, M. F. (2003). *As ideias ortográficas em Portugal - de Madureira Feijó a Gonçalves Viana (1734-1911).* Lisboa: Fundação Calouste Gulbenkian.
14. Gonçalves, M. F., & Banza, A. P. (Edits.). (2013). *Património Textual e Humanidades Digitais: da antiga à nova Filologia.* Évora: Publicações do CIDEHUS, disponível em: https://books.openedition.org/cidehus/1073.

15. Gonçalves, T., Silva, C., Quaresma, P., & Vieira, R. (2006). Analysing part-of-speech for Portuguese text classification. *CICLing-06, 7th International Conference on Intelligence Text Processing and Computational Linguistics* (pp. 551-561). Berlin, Heidelberg: Springer-Verlag, disponível em: https://link.springer.com/chapter/10.1007%2F11671299_57.
16. Guerreiro, D., & Borbinha, J. (Jan-Jun de 2014). Humanidades digitais: novos desafios e oportunidades. *Cadernos BAD, nº1*, pp. 63-78.
17. Han, J., & Kamber, M. (2012). *Data mining: concepts and Techniques* (3rd edition ed.). Waltham, MA: Morgan Kaufman.
18. Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining.* Boston: MIT Press.
19. Kemmler, R. (Outubro de 2001). Para uma História da Ortografia Portuguesa: o texto metaortográfico e a sua periodização do século XVI até à reforma ortográfica de 1911. *Lusorama 47-48*, pp. 128-319.
20. Marquilhas, R. & Hendricx, I. (2014). Manuscripts And Machines: The automatic replacement of spelling variants in a Portuguese Historical Corpus. *International Journal of Humanities and Arts Computing.* Edinburgh: Edinburgh University Press, 8.1, 65–80.
21. Quaresma, P. (2013). Análise linguística de documentos da Biblioteca Pública de Évora: uma abordagem informática. In M. F. Gonçalves, & A. Banza (Edits.), *Património Textual e Humanidades Digitais: da antiga à nova Filologia.* Évora: Publicações do CIDEHUS, disponível em: https://books.openedition.org/cidehus/1091.
22. Reynaert, M., Hendrickx, I. & Marquilhas, R. Historical spelling normalization. A comparison of two statistical methods: TICCL and VARD2. Mambrini, F., Passarotti, M. & Sporleder, C. (edits) (2012) *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2), Lisbon: Ediçoes Colibri, 87-98.*
23. Schreibman, S., Siemens, R., & Unsworth, J. (Edits.). (2004). *A Companion to Digital Humanities.* Oxford: Blackwell.
24. Sousa, M. C. (2013). A Filologia Digital em língua Portuguesa: alguns caminhos. In M. F. Gonçalves, & A. Banza (Edits.), *Património Textual e Humanidades Digitais.* Évora: Publicações do CIDEHUS, disponível em: https://books.openedition.org/cidehus/1089.
25. *Tesouro do léxico patrimonial Português e Galego* (s.d.) Obtained from http://ilg.usc.es/Tesouro/
26. Teyssier, P. (1997). *História da Língua Portuguesa.* Lisboa: Sá da Costa.
27. Verdelho, T. (vol. XXIII de 1987). Latinização na história da Língua Portuguesa - o testemunho dos dicionários. *Arquivos do Centro Cultural Português (volume de homenagem a Paul Teyssier)*, pp. 157-187.
28. Verdelho, T. (1998). Terminologias na língua Portuguesa. Perspectiva diacrónica. In J. (. Brumme, *La història dels llenguatges iberoromànics d'especialitat (segles XVII-XIX): solucions per al present* (pp. 98-131). Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.
29. Vieira, R., Mendes, A., Quaresma, P., Fonseca, E., Collovini, S., & Antunes, S. (2018). Corref-PT: A Semi-automatic Annotated Portuguese Coreference Corpus. In *Computación y Sistemas* (Vols. 22, nº4, pp. 1259-1267). disponível em: http://dspace.uevora.pt/rdpc/handle/10174/24429.