An annotation proposal based on TEI Schema for Portuguese Corpora editions: A solution for e-Dictor XML annotation problem

Aline Silva Costa*[1][0000-0003-1434-3242], Cristiane Namiuti[2] [0000-0002-1451-8391], Maria Clara Paixão de Sousa[3][0000-0002-8422-417X], Bruno Silvério Costa[1][0000-0003-0233-9262] and Jorge Viana Santos[2][0000-0002-8548-4379]

[1] Federal Institute of Education, Science and Technology of Bahia, Vitória da Conquista, Brazil
[2] State University of Southwest Bahia, Vitória da Conquista, Brazil
[3]University of São Paulo, São Paulo, Brazil
alinecosta@ifba.edu.br,
cristianenamiuti@uesb.edu.br,mariaclara@usp.br,
viana.jorge.viana@uesb.edu.br, brunosilverio@ifba.edu.br

Abstract. Annotated multilayer data approaches are becoming widespread in historical corpora. In this context, the Tycho Brahe Corpus of Historical Portuguese was the pioneer in this approach and several other initiatives emerged based on its annotation, generated by e-Dictor software. During the research conducted with corpora based on this annotation scheme, some gaps were identified to cater to the requirements for manuscript sources and conformity increase using annotation standards. These requirements include the TEI (Text Encoding Initiative) Guidelines, which provide a standard format for encoding of texts, to achieve more interoperability, while at the same time solving problems of missing adequacy in the current encoding, among others. This work presents an annotation scheme proposal for the philological editions (or editorial interventions) and morphological analyses of Portuguese historical corpus encoded by XML (eXchange Markup Language). The presented encoding scheme fulfills the reliability requirement for this kind of corpora while achieving more adequacy, conformity, and, consequently, more interoperability.

Keywords: TEI, Historical Corpora, Multilayer Corpora Annotation.

1 Introduction

Multilayer corpora are an emerging methodology within the linguistics corpus, pushing linguistic research to new frontiers. These corpora bring together multiple independent analyzes of the same linguistic phenomena by favoring the interplay of these concurrent analyzes. Multilayer approaches are also spreading in historical corpora, allowing to merge the manuscript structure around the text with representations of morphological analyzes and syntactic treebanks [1]. In the context of the historical corpora of the Portuguese language, the Tycho Brahe Corpus - TBC [2] was the pioneer in this approach, bringing together the philological editions encoding with morphological and syntactic analyzes. The TBC is a historical corpus composed of Portuguese texts from the 14th to the 19th centuries, which were ported to digital support. The current editorial interventions annotation, designed by [3], is expressed using XML. It defines XML tags for each variant point in the text, favoring to keep and recover the original forms and their edited versions. This encoding is applied by e-Dictor software [4], according to editorial interventions made by the user in its graphic interface. The scheme generated by the e-Dictor must fulfill the requirement conformity of the original text to the historical sources, postulated by [5] and corroborated by [6] in the context of historical manuscript sources.

Other Portuguese corpora project initiatives emerged in Brazil based on e-Dictor annotation, and nowadays at least seven large projects are using it [7]. Among these projects is the DOViC Corpus (‘*Corpus de Documentos Oitocentistas de Vitoria da Conquista*’, Vitoria da Conquista Nineteenth Century Documents Corpus, South-west region of Bahia, Brazil) [8]. New requirements arose for the e-Dictor XML annotation scheme as the research progressed. For manuscript documents of DOViC, it was necessary to encode several data for this kind of document, which was not designed by the e-Dictor annotation scheme. As the Portuguese language is among the low-resources languages [9], it is important to adopt standard formats or models under well-documented and accepted norms, such as the TEI Guidelines [10]. The adoption of less idiosyncratic models or representations of data contributes to the dissemination and expansion of research with the Portuguese language, as this favors the exchange of data and its conversion for use with new annotation tools, increasing the probability that other researchers can use the corpus and even extend the data. The e-Dictor scheme also had problems in the adequacy of word representation when segmentation or join edits occurred.

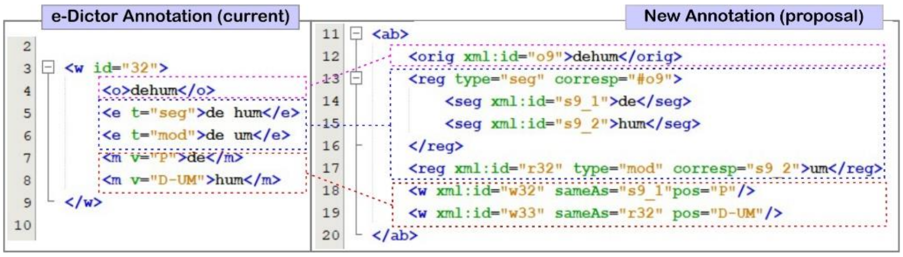
Aiming to achieve more sharing, merging, and comparison of Portuguese language resources and resolve the found problems, a proposal of a new annotation scheme for the historical corpora of the Portuguese language encoded in XML is being developed. This action is part of ongoing doctoral dissertation research within the scope of the Postgraduate Program in Linguistics (PPGLIN) offered by *Universidade Estadual do Sudoeste da Bahia* – UESB (Southwest of Bahia State University). The scheme under development targets greater adequacy, conformity, and, consequently, more interoperability. It must be TEI-conformant and meet the requirement of reliability to original texts. A new version of e-Dictor, as a web-based application, will be developed, incorporating the new annotation proposed. The result can be used to any written corpus of the Portuguese language, either manuscript, oral transcript, or written.

This paper focuses on the philological editions encoding layer annotation, cutting through the proposal annotation scheme under development, presenting specifically an example of segmentation edition followed by modernization of the word spelling. The objective of this work is to present and shortly discuss a proposal of a TEI-conformant scheme for the editorial interventions, meeting the requirements required by research with historical corpora. The corpus used in this research extract is DOViC, using an excerpt from the manuscript document ‘*Carta de Liberdade do Cabrinha Bernardo*’ (Manumission Letter from Slave Bernardo), which is part of this corpus.

2 TEI-conformant proposal for philological editions

The e-Dictor XML encoding uses the <w> element to represent a word, the <o> element to represent the original form of transcript text, and <e> element for any changes made by the editor. The editing type is encoded in the “type” attribute, which can have the values “seg” for segmentation, “jun” for junction, “mod” for modernization, and “gra” for spelling edits assigned to it. Figure 1 presents an excerpt of the document annotated in the current scheme on the left side, where the passage “*dehum*” (“of a”) is composed of two words, which were originally written together (line 4 of Figure 1). The editor segmented the passage, whose encoding is represented in line 5 of Figure 1. Then, the word “*hum*” (“a” or “an” in English), written as “*um*” nowadays, was modernized, which is represented in line 6 of Figure 1. The <m> elements encode the morphological analysis, assigning the POS category to the “v” attribute value. The element content should match the word analyzed by the POS tagger.

We consider that the e-Dictor annotation in this situation is semantically misleading, as it keeps two words “*de*” and “*hum*” in a single <w> element. The attribution of the “D-UM” tag (equivalent to indefinite determinant) was realized by the POS tagger from the string “de hum”, <e> element content of line 6, but the content that received this tag POS corresponds to the unmodernized version “hum”, as encoded in line 8. The annotation also does not explicit that the modernization corresponds only to the word “hum”, rather than the whole string “de um” such as modernization of “de hum”, as the encoding indicates in lines 5 and 6.



```
2
3   <w id="32">
4       <o>dehum</o>
5       <e t="seg">de hum</e>
6       <e t="mod">de um</e>
7       <m v="P">de</m>
8       <m v="D-UM">hum</m>
9   </w>
10
```

```
11  <ab>
12      <orig xml:id="o9">dehum</orig>
13      <reg type="seg" corresp="#o9">
14          <seg xml:id="s9_1">de</seg>
15          <seg xml:id="s9_2">hum</seg>
16      </reg>
17      <reg xml:id="r32" type="mod" corresp="s9_2">um</reg>
18      <w xml:id="w32" sameAs="s9_1"pos="P"/>
19      <w xml:id="w33" sameAs="r32" pos="D-UM"/>
20  </ab>
```

Fig. 1. Comparison between the e-Dictor annotation and the new annotation proposed schemes

Figure 1 presents the same excerpt from the previous example encoded with the new proposal on the right side. For the new annotation scheme, we propose that each “token” coming from the transcription step is encoded in the <orig> element, recommended by TEI as an indication that the reading follows the original. In the edition phase, each <orig> element will be wrapped in an <ab> element (defined by TEI as “anonymous block”). All changes made by the editor to the <orig> content will be recorded in sequential <reg> (regularization) elements within the parent <ab> block. The TEI Guidelines define that <reg> element may be used for any kind of regularization, including normalization, standardization, and modernization [10]. The edition type will be encoded in the “type” attribute and the “corresp” attribute links the edition realized to the element being normalized. The tokens generated by the segmentation are encoded in <seg> elements (defined to segment units) as descendants of the element representing the edition that generated them. Each word will be encoded by the <w> element and the POS category will be encoded by the “pos” attribute. This corresponds to morphological layer annotation and the “pos” value attribution will be realized after tagger POS runs.

We encode <w> with the “same as” attribute, which points to an element whose content is the same as the current element. It is useful to represent the fact that one element of a text is identical to others. Alternatively, the <w> elements (lines 18 and 19 of Figure 1) could be removed from the <ab> block and encoded by the *stand-off* method, into a single block that brings together all the words with the POS category annotation elsewhere in the same file or even in another file. The proposal of a new encoding scheme represents adequately the words in two <w> elements and explicitly marks the linking between the edition and the element that corresponds to it. In the modernization of “hum”, encoded in line 17 of Figure 1, the attribute “corresp” value references the segment whose “xml:id” is “s9_2”, which encodes what content is being modernized.

The new annotation scheme will be implemented in a new version of e-Dictor (version 2.0) that will be developed within the scope of this work. e-Dictor 2.0 will replace the annotator of edits in the current format (e-Dictor 1.0), which has a desktop GUI (Graphic User Interface). The new version will be a web-based application and like e-Dictor 1.0, it will have the POS tagger embedded in its code, calling it after editorial interventions made by the user. The link between edits and segments (<seg> elements) shown in the annotation proposal will be generated according to the user's actions in the software's GUI. This proposal is also suitable for join edits, which generate a word from two or more <orig> elements. Changing the annotation does not impose any overload on e-Dictor software.

The complete result from ongoing research aims to develop a syntactic annotation scheme aligned to other layers plus the annotation of metadata. The resulting scheme can be used to any written corpus of the Portuguese language, either manuscript, oral transcript, or written. Although focused on historical corpora, the scheme is also intended for contemporary corpora. The Carolina Corpus (‘*Corpus Aberto para Linguística e Inteligência Artificial’*, Open Corpus for Linguistics and Artificial Intelligence), a large open corpus of Brazilian Portuguese texts, which was released in March 2022, has already adopted the metadata annotation scheme developed in this work [11].

3 Final considerations

The proposed scheme for morphological analysis and editorial interventions for Portuguese historical corpora, presented in this paper, solves the inadequate representation of words and editorial interventions found in e-Dictor encoding XML. By changing the current to TEI-conformant schema, it achieves a less idiosyncratic scheme, based on a widely accepted standard for annotation of digital texts. Thus, greater interoperability is achieved and there will be a greater possibility that other researchers will use the corpora that use the e-Dictor software, annotated in this format. The complete research aims to present to the research community a multilayered annotation schema for Portuguese historical corpora, with TEI-conformant syntactic annotation aligned to the other layers, joining appropriately developed software for its adoption, thus contributing to the expansion of research with the Portuguese Language and to moving it out of the low-resources languages scenario.

Acknowledgment

We thank FAPESB and CNPq as this work is linked to thematic projects funded by FAPESB (APP0007/2016 and APP0014/2016) and CNPq (436209/2018-7); the Graduate Program in Linguistics (PPGLIN); to the Corpus Linguistics Research Laboratory (LAPELINC); to the State University of Southwest Bahia – UESB; and advisors Prof. Dr. Jorge Viana Santos and Prof. Dr. Cristiane Namiuti.

References

1. Zeldes, A.: Multilayer Corpus Studies. 1st eds. Routledge, New York (2019).
2. Galves, C., Andrade, A. L. D., Faria, P.: Tycho Brahe parsed corpus of historical Portuguese. Unicamp, Campinas (2017).
3. Paixão de Sousa, M. C.: Sistema de Edições Eletrônicas do Corpus Tycho Brahe. Unicamp, Campinas (2007). Homepage https://www.tycho.iel.unicamp.br/corpus/manual/prep/manual_frameset.html, last accessed 2021/10/23.
4. Paixão de Sousa, M. C., Kepler, F., Faria, P.: eDictor: novas perspectivas na codificação e edição de corpora de textos históricos. In: Shepherd, T., Sardinha, T.B., Pinto, M. V. Caminhos da Linguística de Corpus. Mercado de Letras, Campinas (2010).
5. Paixão de Sousa, M. C.: Memória do Texto. In: Revista Texto Digital, n. 2. Universidade Federal de Santa Catarina, Santa Catarina (2006). Homepage http://www.textodigital.ufsc.br/num02/paixao.htm, last accessed 2019/02/23.
6. Paixão de Sousa, M. C.: e-Dictor Homepage. Universidade de São Paulo (USP), São Paulo (2022). Homepage https://edictor.net/projetos-envolvidos/, last accessed 2022/03/08.
7. Santos, J. V., Namiuti, C.: O objeto livro: a complexidade da forma e o digital. In: X Congresso Internacional da ABRALIN. Universidade Federal Fluminense, Niterói (2017).
8. Santos, J. V., Namiuti, C.: DOViC - Documentos Oitocentistas de Vitória da Conquista. Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista (2016).
9. Center for Artificial Intelligence. Research in the C4AI. Universidade Estadual de São Paulo (USP), São Paulo (2021). HomePage: http://c4ai.inova.usp.br/research/#NLP2, last accessed 2022/03/02.
10. TEI Consortium: TEI P5: Guidelines for Electronic Text. 4.2.1 version (2021).
11. Center for Artificial Intelligence. Carolina (Corpus Aberto para Linguística e Inteligência Artificial). Universidade Estadual de São Paulo (USP), São Paulo (2022). Homepage http://sites.usp.br/corpuscarolina, last accessed 2019/02/23.