

Modelling Usage Information in a Legacy Dictionary: From TEI Lex-0 to Ontolex-Lemon

Bruno Almeida¹, Rute Costa¹, Ana Salgado^{1,2}, Margarida Ramos¹, Laurent Romary³,
Fahad Khan⁴, Sara Carvalho^{1,5}, Mohamed Khemakhem⁶, Raquel Silva¹, Toma Tasovac⁷

¹NOVA CLUNL - Centro de Linguística da Universidade Nova de Lisboa

²ACL - Academia das Ciências de Lisboa

³Inria - Institut national de recherche en sciences et technologies du numérique

⁴ILC-CNR - Istituto di Linguistica Computazionale "Antonio Zampolli"

⁵CLLC - Centro de Línguas, Literaturas e Culturas UA

⁶ArcaScience

⁷BCDH - Belgrade Center for Digital Humanities

mordigital@fcsh.unl.pt

Abstract

This paper describes ongoing work in the modelling of usage information in the context of the MORDigital project. The latter is based on the encoding and publication as linked data of *Dicionario da Lingua Portuguesa*, a Portuguese legacy dictionary authored by António de Morais Silva, whose first edition was published in 1789. In this paper, we will focus on the modelling of domain labels in Ontolex-Lemon, based on a previous encoding of the dictionary's entries in TEI Lex-0. This approach should be reusable for other projects involving the linked data publication of legacy dictionaries.

1 Introduction

The publication of *Dicionario da Lingua Portuguesa* in 1789, authored by António de Morais Silva, marks the beginning of contemporary Portuguese lexicography, following the model set by several modern language dictionaries published in Europe in the 17th and 18th centuries (Silvestre, 2009; Verdelho, 2003). As the first Portuguese monolingual dictionary, it had a fundamental role in the normalisation of this language, and constitutes a reference for studying the evolution of the Portuguese lexicon (Correia, 2009). The first edition of the dictionary had two volumes (Vol. 1, 752 p. and Vol. 2, 541 p.). Morais directly oversaw the 2nd and 3rd editions (published, respectively, in 1813 and 1823). This work was greatly revised and

updated over the years, culminating in the 10th edition, which was published in 12 volumes from 1949 to 1959.

The MORDigital project aims at digitising and publishing in open access the structured data of the first three editions of the dictionary by Morais (Costa et al., 2021). Our methodology involves the reuse of digitised versions of the dictionary, available in the public domain as PDF files with OCR data. The digitised versions will be structured by means of several open standards for encoding and modelling lexical and dictionary data, which will facilitate interoperability with existing systems and datasets. The encoding of the dictionary's editions will be carried out in TEI Lex-0 (Tasovac et al., 2018), a baseline XML encoding for machine-readable dictionaries based on the guidelines of the Text Encoding Initiative (TEI). This encoding will be the basis for a LMF (Lexical Markup Framework) version, which should be facilitated by the present convergence between TEI and the LMF standard (Romary, 2015). Furthermore, the TEI Lex-0 encoding of the Morais dictionary will be transformed to RDF based on Ontolex-Lemon (Cimiano et al., 2016), a model originally developed for enriching ontologies with lexical information, which has become a *de facto* standard for publishing lexical resources as linked data (Cimiano et al., 2020). The recently developed lexicography module, or *lexicog* (Bosque-Gil et al., 2019), facilitates the application of Ontolex to dictionary information.

`tei2ontolex`¹ will be used for converting from TEI to Ontolex. Examples such as the one

¹ <https://github.com/elexis-eu/tei2ontolex>

META'STASE, ou *Metastasis*, f. f. *Med.* de-
 geração de huma doença em outra, especie de
 Crife. § *na Rhet.* figura pela qual o Orador at-
 tribue alguma coisa a outrem, desonerando-fe
 della.

presented in this abstract will be the basis for a

Figure 1: *Metástase* entry

wider coverage of features of this converter.

2 Usage information in the Morais dictionary

Usage information consists of constraints on the
 use of words or senses to certain contexts, or to a
 subset of language users (Landau, 2001; Svensén,
 2009). Dictionaries traditionally include usage in-
 formation in the entries as labels, notes or within
 the definitions themselves.

Figure 1 shows an entry with domain labels
 (Silva, 1789, vol. 2, p. 79). This entry for *metástase*
 ('metastasis') has two senses, separated by the sec-
 tion sign (§). Each of these senses is associated
 with different subject fields, namely medicine
 (*Med.*) and rhetoric (*na Rhet.*).

The analysis of the dictionary's list of abbrevia-
 tions provides valuable insights into the usage in-
 formation that was more relevant to the late 18th
 century lexicographer. Our analysis resulted in the
 following typology of labels (the corresponding
 types of usage in TEI Lex-0, following Salgado et
 al. (2019), are shown in parentheses:

- *Diatechnical information* (domain). These labels indicate that the lexical unit belongs to the specialised language of a subject field (e.g., *Med.*, for medical terms).
- *Diatextual information* (textType). These labels identify the text or discourse types in which the lexical units are used (e.g., *Poet.*, for poetic words).
- *Diaevaluative information* (attitude). These labels associate a lexical unit with a specific attitude on the speaker's part (e.g., *t. Chulo*, for ironic or malicious usages).
- *Diastratic information* (socioCultural). These labels associate a lexical unit with a particular social group (e.g., *Vulg.*, for words associated with the common people).

- *Diaphasic information* (socioCultural). These labels associate a lexical unit with a register (e.g., *Fam.*, for words used in an informal register).
- *Diatopic information* (geographic). These labels associate a lexical unit with a regional variety of a language (e.g., *Asiat.*, for words used in the former Portuguese colonies in India).
- *Diachronic information* (temporal). These labels associate a lexical unit with a period in the history of language (e.g., *Ant.*, for dated words).
- *Diainegrative information* (hint). These labels indicate that a lexical unit is a loanword (e.g., *Lat.*, for Latin words integrated in Portuguese).
- *Diafrequential information* (frequency). These labels indicate the frequency of occurrence of a lexical unit (e.g., *P. us.*, for rarely used words).

3 Encoding in TEI Lex-0

Figure 2 shows the encoding in TEI Lex-0 of the above-mentioned entry.

```
<entry xmlns="http://www.tei-c.org/ns/1.0"
type="monolexicalUnit" xml:lang="pt"
xml:id="MORAIS_1.metastase">
  <form type="lemma">
    <orth>METÁSTASE</orth>
    <pc>, ou</pc>
    <form type="variant">
      <orth>Metastasis</orth>
    </form>
  </pc></pc>
  <gramGrp>
    <gram type="pos"
norm="NOUN">s.</gram>
    <gram type="gen">f.</gram>
  </gramGrp>
  <sense xml:id="MORAIS_1.metastase_1">
    <usg type="domain">Med.</usg>
    <def>degeneração de huma doença em
outra, espécie de Crise</def>
  </sense>
  <sense xml:id="MORAIS_1.metastase_2">
    <pc>na</pc>
    <usg type="domain">Rhet.</usg>
    <def>figura pela qual o Orador attribue
alguma coisa a outrem , desonerando-se
della.</def>
  </sense>
</entry>
```

Figure 2: *Metástase* entry in TEI Lex-0

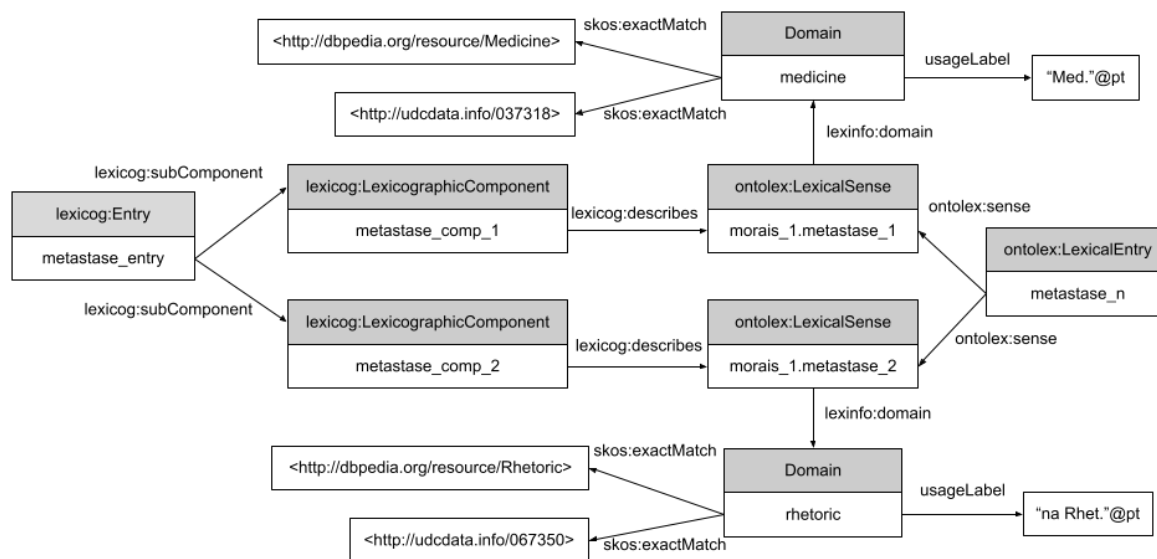


Figure 3: Usage information for the *metastase* entry in Ontolex.

134 In TEI Lex-0, usage information is encoded
 135 through the `<usg>` typed element (with the value
 136 `domain`, in this case). The domain label is linked
 137 to a hierarchy of domains, which will be defined
 138 and modelled through an ontology, as carried out in
 139 a related project (Costa et al., 2020). This approach
 140 should be reusable in other projects going forward
 141 (Costa et al., 2021).

142 4 Modelling in Ontolex-Lemon

143 Ontolex-Lemon already provides most of the neces-
 144 sary elements for modelling information associat-
 145 ed with lexical entries. The lexicographic mod-
 146 ule, `lexicog`, provides additional elements for
 147 dictionary data. This allows, e.g., to distinguish be-
 148 tween *lexical entries* (which must belong to the
 149 same part of speech, such as noun or adjective) and
 150 *dictionaries entries* (which often aggregate differ-
 151 ent parts of speech). The LexInfo ontology² pro-
 152 vides data categories for Ontolex, including several
 153 usage sub-properties aligned with TEI Lex-0 (e.g.
 154 `domain`, `socioCultural`). Finally, SKOS (Sim-
 155 ple Knowledge Organization System) (Miles &
 156 Bechhofer, 2009) allows to organise the subject
 157 fields corresponding to the domain labels, and align
 158 them with external resources and knowledge or-
 159 ganisation systems (KOS).

² <https://lexinfo.net/>

³ The representation of grammatical and semantic infor-
 mation was omitted to simplify the diagram.

160 Figure 3 shows the *metastase* entry modelled in
 161 Ontolex, along with elements from the above-men-
 162 tioned ontologies.³ Structural elements are mod-
 163 elled through `lexicog`, allowing to distinguish be-
 164 tween the dictionary entry and the *metastase* noun.
 165 The noun’s senses are constrained to different do-
 166 mains (`medicine` and `rhetoric`), which are
 167 aligned with DBPedia and the Universal Decimal
 168 Classification for interoperability as linked data.⁴

169 Acknowledgments

170 This paper is supported by (1) the MORDigital –
 171 Digitalização do Dicionário da Língua Portuguesa
 172 de António de Morais Silva [PTDC/LLT-
 173 LIN/6841/2020] project financed by the Portu-
 174 guese National Funding through the FCT – Funda-
 175 ção para a Ciência e Tecnologia (2) Portuguese Na-
 176 tional Funding through the FCT – Fundação para a
 177 Ciência e Tecnologia as part of the project Centro
 178 de Linguística da Universidade NOVA de Lisboa –
 179 UID/LIN/03213/2020 and (3) the European Un-
 180 ion’s Horizon 2020 research and innovation pro-
 181 gramme under grant agreement No 731015
 182 (ELEXIS) (European Lexicographic Infrastruc-
 183 ture).

184 References

185 Bosque-Gil, J., Gracia, J., McCrae, J. P., Cimiano, P.,
 186 Stolk, S., Khan, F., Depuydt, K., Does, J., Frontini,
 187 F., & Kernerman, I. (2019). *The OntoLex Lemon*

⁴ Alternative approaches for modelling domain labels will
 be discussed in the full paper.

- 188 *Lexicography Module: Final community report.* 239
 189 W3C Ontology-Lexicon Community Group. 240
 190 <https://www.w3.org/2016/05/ontolex/> 241
- 191 Cimiano, P., Chiarcos, C., McCrae, J. P., & Gracia, J. 242
 192 (2020). *Linguistic Linked Data: Representation,* 243
 193 *Generation and Applications.* Springer. 244
- 194 Cimiano, P., McCrae, J. P., & Buitelaar, P. (2016). *Lex-* 245
 195 *icon Model for Ontologies: Community Report.* 246
 196 W3C Ontology-Lexicon Community Group. 247
 197 <https://www.w3.org/2016/05/ontolex/>
- 198 Correia, M. (2009). *Os dicionários portugueses.* Cami-
 199 nho.
- 200 Costa, R., Carvalho, S., Salgado, A., Simões, A., &
 201 Tasovac, T. (2020). Ontologie des marques de do-
 202 maines appliquée aux dictionnaires de langue gé-
 203 nérale. *Langue(s) & Parole*, 5, 201–230.
- 204 Costa, R., Salgado, A., Khan, A., Carvalho, S., Ro-
 205 mary, L., Almeida, B., Ramos, M., Khemakhem, M.,
 206 Tasovac, T., & Silva, R. (2021). MORDigital: The
 207 advent of a new lexicographical Portuguese project.
 208 In I. Kosem, M. Cukr, M. Jakubiček, J. Kallas, S.
 209 Krek, & C. Tiberius (Eds.), *Electronic lexicography*
 210 *in the 21st century: Post-editing lexicography. Pro-*
 211 *ceedings of eLex 2021* (pp. 312–324). Lexical Com-
 212 puting.
- 213 Landau, S. (2001). *Dictionaries: The art and craft of*
 214 *lexicography* (2nd ed). Cambridge University Press.
- 215 Miles, A., & Bechhofer, S. (2009, August 18). *SKOS*
 216 *Simple Knowledge Organization System Reference.*
 217 <http://www.w3.org/TR/skos-reference>
- 218 Romary, L. (2015). TEI and LMF crosswalks. *Journal*
 219 *for Language Technology and Computational Lin-*
 220 *guistics*, 30(1), 47–70.
- 221 Salgado, A., Costa, R., & Tasovac, T. (2019). Improv-
 222 ing the Consistency of Usage Labelling in Diction-
 223 aries with TEI Lex-0. *Lexicography*, 6, 133–156.
 224 <https://doi.org/10.1007/s40607-019-00061-x>
- 225 Silva, A. M. (1789). *Diccionario da lingua portugueza*
 226 *composto pelo padre D. Rafael Bluteau, reformado,*
 227 *e accrescentado por Antonio de Moraes Silva, natu-*
 228 *ral do Rio de Janeiro* (Vol. 1–2). Officina de Simão
 229 Thaddeo Ferreira.
- 230 Silvestre, J. P. (2009). *Bluteau e as origens da lexico-*
 231 *grafia moderna.* INCM.
- 232 Svensén, B. (2009). *A handbook of lexicography: The*
 233 *theory and practice of dictionary-making.* Cam-
 234 bridge University Press.
- 235 Tasovac, T., Romary, L., Banski, P., Bowers, J., Does,
 236 J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A.,
 237 Hildenbrandt, V., Khemakhem, M., Lehečka, B., Pe-
 238 trović, S., Salgado, A., & Witt, A. (2018). *TEI Lex-*
 239 *0: A baseline encoding for lexicographic data. Ver-*
 240 *sion 0.9.0.* DARIAH Working Group on Lexical Re-
 241 sources. [https://dariah-eric.github.io/lexicalre-](https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html)
 242 [sources/pages/TEILex0/TEILex0.html](https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html)
- 243 Verdelho, T. (2003). O dicionário de Morais Silva e o
 244 início da lexicografia moderna. *História Da Língua*
 245 *e História Da Gramática: Actas Do Encontro*, 473–
 246 490.