# Modelling usage information in a legacy dictionary: from TEI Lex-0 to Ontolex-Lemon

Bruno Almeida, Rute Costa, Ana Salgado, Margarida Ramos, Laurent Romary, Fahad Khan, Sara Carvalho, Mohamed Khemakhem, Raquel Silva, Toma Tasovac

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

**Bruno Almeida**[1], **Rute Costa**[1], **Ana Salgado**[1 2], **Margarida Ramos**[1], **Laurent Romary**[3], **Fahad Khan**[4], **Sara Carvalho**[1 5], **Mohamed Khemakhem**[6], **Raquel Silva**[1], **Toma Tasovac**[7]
[1]NOVA CLUNL - Centro de Linguística da Universidade Nova de Lisboa
[2]ACL - Academia das Ciências de Lisboa
[3]Inria - Institut national de recherche en sciences et technologies du numérique
[4]ILC-CNR - Istituto di Linguistica Computazionale "Antonio Zampolli"
[5]CLLC - Centro de Línguas, Literaturas e Culturas UA
[6]ArcaScience
[7]BCDH - 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 *Diccionario da Lingua Portugueza*, 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.

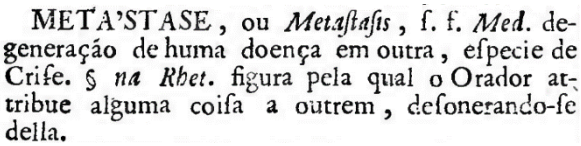
## 1 Introduction

The publication of *Diccionario da Lingua Portugueza* 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*, ſ. f. *Med.* degeneração de huma doença em outra, eſpecie de Criſe. § *na Rhet.* figura pela qual o Orador attribue alguma coiſa a outrem, deſonerando-ſe della.

presented in this abstract will be the basis for a wider coverage of features of this converter.

Figure 1: *Metástase* entry

## 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 information 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 section 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 abbreviations provides valuable insights into the usage information 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).

- *Diaintegrative 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>
</entry>
```

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

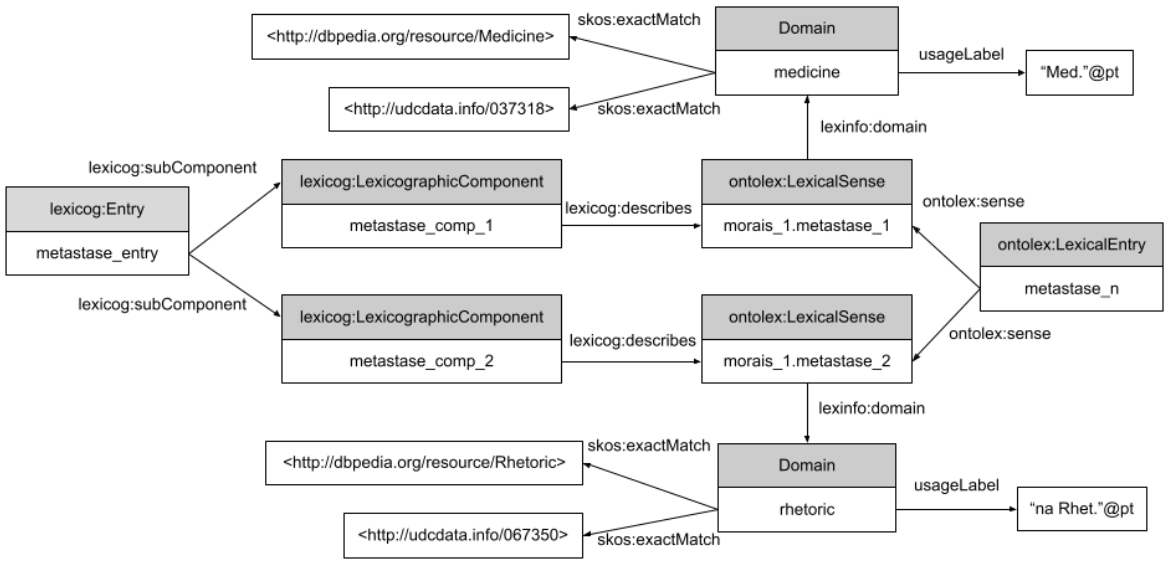



Figure 3: Usage information for the *metástase* entry in Ontolex.

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

## 4 Modelling in Ontolex-Lemon

Ontolex-Lemon already provides most of the necessary elements for modelling information associated with lexical entries. The lexicographic module, `lexicog`, provides additional elements for dictionary data. This allows, e.g., to distinguish between *lexical entries* (which must belong to the same part of speech, such as noun or adjective) and *dictionaries entries* (which often aggregate different parts of speech). The LexInfo ontology[2] provides data categories for Ontolex, including several usage sub-properties aligned with TEI Lex-0 (e.g. `domain`, `socioCultural`). Finally, SKOS (Simple Knowledge Organization System) (Miles & Bechhofer, 2009) allows to organise the subject fields corresponding to the domain labels, and align them with external resources and knowledge organisation systems (KOS).

Figure 3 shows the *metástase* entry modelled in Ontolex, along with elements from the above-mentioned ontologies.[3] Structural elements are modelled through `lexicog`, allowing to distinguish between the dictionary entry and the *metástase* noun. The noun's senses are constrained to different domains (`medicine` and `rhetoric`), which are aligned with DBPedia and the Universal Decimal Classification for interoperability as linked data.[4]

## Acknowledgments

This paper is supported by (1) the MORDigital – Digitalização do Diccionario da Lingua Portugueza de António de Morais Silva [PTDC/LLT-LIN/6841/2020] project financed by the Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia (2) Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020 and (3) the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS) (European Lexicographic Infrastructure).

## References

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

[2] https://lexinfo.net/

[3] The representation of grammatical and semantic information was omitted to simplify the diagram.

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

*Lexicography Module: Final community report*. W3C Ontology-Lexicon Community Group. https://www.w3.org/2016/05/ontolex/

Cimiano, P., Chiarcos, C., McCrae, J. P., & Gracia, J. (2020). *Linguistic Linked Data: Representation, Generation and Applications*. Springer.

Cimiano, P., McCrae, J. P., & Buitelaar, P. (2016). *Lexicon Model for Ontologies: Community Report*. W3C Ontology-Lexicon Community Group. https://www.w3.org/2016/05/ontolex/

Correia, M. (2009). *Os dicionários portugueses*. Caminho.

Costa, R., Carvalho, S., Salgado, A., Simões, A., & Tasovac, T. (2020). Ontologie des marques de domaines appliquée aux dictionnaires de langue générale. *Langue(s) & Parole*, *5*, 201–230.

Costa, R., Salgado, A., Khan, A., Carvalho, S., Romary, L., Almeida, B., Ramos, M., Khemakhem, M., Tasovac, T., & Silva, R. (2021). MORDigital: The advent of a new lexicographical Portuguese project. In I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, & C. Tiberius (Eds.), *Electronic lexicography in the 21st century: Post-editing lexicography. Proceedings of eLex 2021* (pp. 312–324). Lexical Computing.

Landau, S. (2001). *Dictionaries: The art and craft of lexicography* (2nd ed). Cambridge University Press.

Miles, A., & Bechhofer, S. (2009, August 18). *SKOS Simple Knowledge Organization System Reference*. http://www.w3.org/TR/skos-reference

Romary, L. (2015). TEI and LMF crosswalks. *Journal for Language Technology and Computational Linguistics*, *30*(1), 47–70.

Salgado, A., Costa, R., & Tasovac, T. (2019). Improving the Consistency of Usage Labelling in Dictionaries with TEI Lex-0. *Lexicography*, *6*, 133–156. https://doi.org/10.1007/s40607-019-00061-x

Silva, A. M. (1789). *Diccionario da lingua portugueza composto pelo padre D. Rafael Bluteau, reformado, e accrescentado por Antonio de Moraes Silva, natural do Rio de Janeiro* (Vol. 1–2). Officina de Simão Thaddeo Ferreira.

Silvestre, J. P. (2009). *Bluteau e as origens da lexicografia moderna*. INCM.

Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge University Press.

Tasovac, T., Romary, L., Banski, P., Bowers, J., Does, J., Depuydt, K., Erjavec, T., Geyken, A., Herold, A., Hildenbrandt, V., Khemakhem, M., Lehečka, B., Petrović, S., Salgado, A., & Witt, A. (2018). *TEI Lex-0: A baseline encoding for lexicographic data. Version 0.9.0*. DARIAH Working Group on Lexical Resources. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

Verdelho, T. (2003). O dicionário de Morais Silva e o início da lexicografia moderna. *História Da Língua e História Da Gramática: Actas Do Encontro*, 473–490.