Integrating Terminological and Ontological Principles into a Lexicographic Resource

Rute Costa[1], Ana Salgado [1, 2], Margarida Ramos [1], Fahad Khan[3], Sara Carvalho[1,4], Toma Tasovac [5], Bruno Almeida[1,6], Mohamed, Khemakhem [7], Laurent Romary [8], Raquel Silva [1]

[1] *CLUNL – Centro de Linguística da Universidade Nova de Lisboa, Lisboa, Portugal*
[2] *Academia das Ciências de Lisboa, Lisboa, Portugal*
[3] *CNR – Istituto di Linguistica Computazionale "Antonio Zampollo" Pisa, Italy*
[4] *CLLC – Cetnro de Línguas. Literaturas e Culturas, Aveiro. Portugal*
[5] *BCDH – Belgrade Center for Digital Humanities, Belgrade. Serbia*
[6] *ROSSIO - ROSSIO Infrastructure - Social Sciences, Arts and Humanities, Lisboa, Portugal*
[7] *ArcaScience. Paris, France*
[8] *ALMAnaCH – Automatic Language Modelling and ANAlysis & Compuatational Humanities, INRIA, Paris, France*

Abstract

In this paper we will present the research that is taking place at the NOVA CLUNL[1] where an international team is working on a financed project MORDigital[2]. MORDigital's goal is to encode the selected editions of *Diccinario de Lingua Portugue*za by António de Morais Silva (MOR), first published in 1789.

Keywords

dictionary, lexicography, digital humanities, standards

1. Introduction

MORDigital's ultimate goals are, on the one hand, to promote accessibility to cultural heritage while fostering reusability and, on the other hand, to contribute towards a more significant presence of lexicographic digital content in Portuguese through open tools and standards. MOR represents a significant legacy, since it marks the beginning of Portuguese dictionaries, having served as a model for all subsequent lexicographic production. The team follows a new paradigm in lexicography, which results from the convergence between lexicography, terminology, computational linguistics, and ontologies as an integral part of digital humanities and linked (open) data. In the Portuguese context, this research fills a gap concerning searchable online retrodigitised dictionaries, built on current

[1] https://clunl.fcsh.unl.pt/grupos_clunl/lexicologia-lexicografia-terminologia/
[2] https://www.fct.pt/apoios/projectos/consulta/vglobal_projecto?idProjecto=164850&idElemConcurso=14818

1st International Conference on "Multilingual digital terminology today. Design, representation formats and management systems", June 16 – 17, Padova, Italy

EMAIL: rute.costa@fcsh.unl.pt (A. 1); anacastrosalgado@gmail.com (A. 2); fahad.khan@ilc.cnr.it (A. 3) ; mvramos@fcsh.unl.pt (A.4) ; sara.carvalho@ua.pt (A.5) ; ttasovac@humanistika.org (A.6) ; brunoalmeida@fcsh.unl.pt (A.7); mohamed.khemakhem@inria.fr (A.8) laurent.romary@inria.fr (A.9) ; (raq.silva@fcsh.unl.pt A.10)

ORCID: 0000-0002-3452-7228 (A. 1); 0000-0002-6670-3564 (A. 2); 0000-0002-1551-7438 (A. 3) ; 0000-0001-7209-3806 (A.4) ; 0000-0002-7501-5405 (A.5); 0000-0002-3919-993X (A.6); 0000-0002-5777-5574 (A.7); 0000-0003-3529-2990 (A.8); 0000-0002-0756-0508 (A.9); 0000-0002-0505-4863 (A.10)

standards and methodologies which promote data sharing and harmonisation, namely TEI Lex-0[4] and Ontolex-Lemon[5]. The team will further ensure the connection to other existing systems and lexical resources, particularly in the Portuguese-speaking world.

For this paper, after posing the theoretical background (terminology and lexicography) that /underpins our methodology, we will present 4 interrelated tasks:

1. Structuration of MOR's digitised versions using GROBID-Dictionaries[6], a specific software for the parsing, extraction and structuring of information extracted from dictionary text. In our case, the tool will be used to parse the constituent parts of each dictionary entry, which involves the preparation of a native encoding format that is compliant with the XML/TEI metamodel.
2. Presentation of a systematic analysis of the Mathematical Sciences and Medical Sciences domains, their related domain labels [6], [1] and other mechanisms, such as the use of formulae present in the definition which identifies the specialised field of knowledge. We will propose a hierarchical organisation that constitutes the foundation of domain ontologies.
3. Representation of the model in OWL resorting to Protégé[7], a free, open-source ontology editor. This means each class or individual in the ontology will be assigned a URI (Universal Resource Identifier), used to reference the label present in each of the lexicographic entries in accordance – whenever possible – with the TEI schemas.
4. Conversion of the TEI Lex-0 output of Task 4 into linked data using the RDF-based model Ontolex-Lemon; the conversion will be based on work already carried out in the scope of previous initiatives in rendering the two models more interoperable. The Ontolex-Lemon model has recently been extended by a lexicography module – lexicog[8] –, which facilitates interoperability in modelling dictionaries as linked data.

At the end of the paper, we will discuss the results, highlighting the challenges that we faced.

2. Acknowledgements

This paper is supported by 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.

3. References

[1] R. Costa, S. Carvalho, A. Salgado, A. Simões, T. Tasovac (2020). Ontologie des marques de domaines appliquée aux dictionnaires de langue générale, in [éditeur : Xavier Blanco] La lexicographie en tant que méthodologie de recherche en linguistique Revue de Philologie Française et Romane – Langue(s) & Parole, n. 5 . Mons: Edition du CIPA. pp. 201–230. ISSN papier 2466-7757, ISSN numérique 2684-6691.

[2] R. Costa, A. Salgado, B. Almeida (2021). SKOS as a key element for linking lexicography to digital humanities. Information Organization in Digital Humanities: A Global Perspective. Coll. Digital Research in the Arts and Humanities. [Editors: Koraljka Golub / Ying-Hsang Liu], Routledge, pp. 178–204. ISBN 97803675516.

[3] R. Costa, A. Salgado, F. Khan, S. Carvalho, L. Romary, B. Almeida, M. Khemakhem. M. Ramos, R. Silva, T. Tasovac (2021). MORDigital: the advent of a new lexicographical Portuguese project. Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference., Lexical Computing CZ s.r.o., Brno, Czech Republic, pp. 321–324. ISSN 2533-5626.

[4] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html
[5] https://www.w3.org/community/ontolex/
[6] https://github.com/MedKhem/grobid-dictionaries
[7] https://protege.stanford.edu/
[8] https://www.w3.org/2019/09/lexicog/

[4] F. Kahn, A. Salgado (2021). Modelling Lexicographic Resources Using CIDOC CRM, FRBRoo and Ontolex Lemon. In: A. Bikakis et al., eds., SWODCH 2021 – Semantic Web and Ontology Design for Cultural Heritage 2021. Proceedings of the International Joint Workshop on Semantic Web and Ontology Design for Cultural Heritage co-located with the Bolzano Summer of Knowledge 2021 (BOSK 2021). Bozen-Bolzano: CEUR-WS, pp. 1–12. ISSN 1613-0073.

[5] F. Khan, L. Romary, A. Salgado, J. Bowers, M. Khemakhem, T. Tasovac (2020). Modelling Etymology in LMF/TEI: The 'Grande Dicionário Houaiss da Língua Portuguesa' Dictionary as a Use Case. In: N. Calzolari et al., eds., LREC 2020 Conference Proceedings. Paris: ELRA, pp. 3172–3180. ISBN 979-10-95546-34-4.

[6] A. Salgado, R. Costa, (2019). Marcas temáticas en los diccionarios académicos ibéricos: estudio comparativo. RILEX: Revista sobre investigación léxicos, 2(2), pp. 37–63. e-ISSN 2605-3136.

[7] A. Salgado, R. Costa, T. Tasovac (2019). Improving the consistency of usage labelling in dictionaries with TEI Lex-0. Lexicography: Journal of ASIALEX. e-ISSN 2197-4306.

[8] A. Salgado, R. Costa, T. Tasovac, A. Simões, Alberto (2019). TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In: I. Kosem et al., eds., Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., pp. 417–433. ISSN 2533-5626.