▶ Interlinking Lexicographic Data in the MORDigital Project

Anas Fahad Khan,
Istituto di Linguistica Computazionale "Antonio Zampolli", Italy, fahad.khan@ilc.cnr.it
Ana Salgado,
Centro de Linguística da Universidade Nova de Lisboa & Academia das Ciências de Lisboa, Portugal, ana.salgado@fcsh.unl.pt
Rute Costa,
Centro de Linguística da Universidade Nova de Lisboa, Portugal, rute.costa@fcsh.unl.pt
Sara Carvalho,
Centro de Linguística da Universidade Nova de Lisboa & CLLC – Centro de Línguas, Literaturas e Culturas, Portugal, sara.carvalho@ua.pt
Laurent Romary,
Automatic Language Modelling and ANAlysis & Computational Humanities Inria de Paris, France, laurent.romary@inria.fr
Bruno Almeida,
Centro de Linguística da Universidade Nova de Lisboa, Portugal, brunoalmeida@fcsh.unl.pt
Margarida Ramos,
Centro de Linguística da Universidade Nova de Lisboa, Portugal, mvramos@fcsh.unl.pt
Mohamed Khemakhem,
ArcaScience, France, medkhemakhemfsegs@gmail.com
Raquel Silva,
Centro de Linguística da Universidade Nova de Lisboa, Portugal, raq.silva@fcsh.unl.pt
Toma Tasovac,
BCDH – Belgrade Center for Digital Humanities, Serbia, tasovac@humanistika.org

Purpose: To introduce MORDigital as an innovative Portuguese national project that incorporates the latest results in computational lexicography, the digital humanities, and linguistic linked data. In particular, we will show how it brings together work in the development of TEI Lex-0 and OntoLex-Lemon, as well as recent innovations on the conversion of retrodigitized dictionaries into computational lexical resources (using in this case the GROBID-dictionaries tool).

Design/methodology/approach: The aim of the project is to convert three editions (1789; 1813; 1823) of the important legacy Portuguese-language lexicographic re-

source, the Diccionario da Lingua Portugueza by António de Morais Silva (hereinafter – Morais), into a computer-readable resource. The lexical content of the high-quality OCR of the Morais will be automatically structured (using the GROBID-dictionaries tool) into TEI Lex-0, and this will then be converted to a TEI encoding according to the LMF standards. This will be subsequently converted to RDF using the OntoLex model using an XSLT stylesheet, allowing us to make the dictionary available both using a dedicated platform and via a SPARQL endpoint, and permitting users to download versions of the dictionaries in RDF and TEI-XML. The RDF versions of each edition of the dictionary will be added to the LOD cloud, thus adding a historically significant Portuguese language lexical resource to the cloud.

Findings: We will describe the pipeline used for the production of the first edition of the Morais, as well as the specific challenges of modelling lexicographic articles in both TEI-Lex0 and OntoLex and the more general implications this has both for the creation of lexical resources in the Portuguese language and for the digitization of historical (and historically important) dictionaries. At the end of the project, we will propose technical guidelines to help lexicographers and digital humanists. This document will be openly available on the dedicated platform.

Research limitations/implications: As mentioned above, our work should be useful for anyone working on converting historical dictionaries into digital lexical resources using TEI-Lex0, LMF, and OntoLex. We will also look at some of the limitations in these models and currently existing tools when working with historical retrodigitized dictionaries.

Practical implications: The pipeline used in this project, as well as our more general practical observations of working with historical dictionaries, should be useful for anyone working on similar tasks.

Originality/Value: This project is fairly innovative in its modelling of a retrodigitized dictionary in two of the latest digital lexicographic standards; this will enable us to make this important lexicographic work as accessible as possible. Furthermore, we also intend to apply terminological methods to lexicographic work by combining semasiological and onomasiological approaches, thereby providing added value via the use of ontologies, something that is currently missing in general language dictionaries. These results will be evaluated at different levels, namely regarding: the quality of the OCR systems; the ontology (the quality of the modelling); and the platform (based on end-user satisfaction).

Keywords: *Dictionary, lexicography, Portuguese, Linked Open Data, TEI*

Research type: Case study