MOR*Digital*: The Advent of a New Lexicographic Portuguese Project

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

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

NOVAFCSH
FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS
UNIVERSIDADE NOVA DE LISBOA

FCT Fundação para a Ciência e a Tecnologia

Institutions involved

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

Inria

TOSSIO

BELGRADE CENTER FOR DIGITAL HUMANITIES

universidade de aveiro cllc centro de línguas, literaturas e culturas

AS ArcaScience

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

Institutions involved

NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal

Academia das Ciências de Lisboa, Portugal

Istituto Di Linguistica Computazionale 'A. Zampolli', Italy

CLLC, Centro de Línguas, Literaturas e Culturas da Universidade de Aveiro, Portugal

Inria, team ALMAnaCH, France

ROSSIO Infrastructure, Portugal

Arcascience, France

BCDH – Belgrade Center for Digital Humanities, Serbia

Outline

- Introduction
- Theoretical framework
- Historical background
- Morais dictionary
- MORDigital
  - The project
  - Methodology
  - End product
- Concluding remarks

Introduction

- The *Diccionario da Lingua Portugueza* by António de Morais Silva marks the beginning of modern Portuguese lexicography and serves as a model for all subsequent lexicographic production throughout the 19th and 20th centuries.
- The MOR*Digital* project aims to produce high-quality and searchable digital versions of the first three editions (1789; 1813; 1823) of Morais in order to preserve this important European heritage work.
- These digital versions will be converted into structured data and made publicly available with the purpose of guaranteeing the preservation of this legacy resource.
- This project aims to make a substantial contribution to the scientific community and aspires to apply innovative computational methodologies.

Theoretical framework (1)

- Lexicography has undergone a radical change in the past two decades.
- This paradigm shift is directly related to the advancement of digital humanities.
- The perspective underpinning the construction of lexical resources presupposes rethinking the methodologies of the Portuguese lexicographic tradition, perceiving lexicography, terminology, ontologies and computational linguistics as an integral part of the digital humanities, which will imply a paradigm shift in the construction of dictionary resources.
- In this new paradigm, ontologies will play a key role in organising and representing linguistic and metalinguistic knowledge (Carvalho, Costa & Roche, 2018; Almeida, Costa & Roche, 2019), as well as supporting its operationalisation and, therefore, its preservation in the long term.

Theoretical framework (2)

- The European lexicographic scenario is currently quite heterogeneous, in what concerns

  (1) the types of existing lexicographic resources;
  (2) their structural component, which relates to how the data are represented, the adopted models, as well as the respective applied formats.

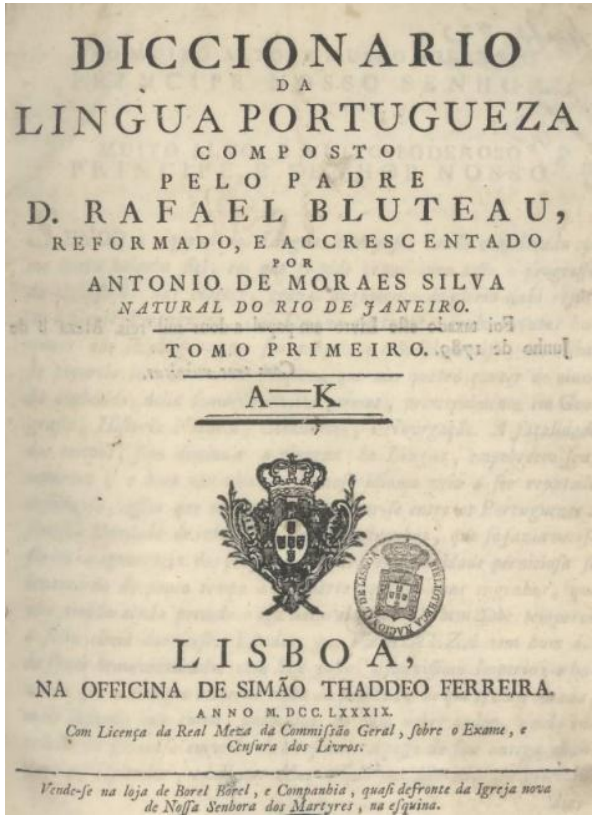
- The diversity of incompatible formats creates severe problems in the digital landscape, making it impossible to interconnect resources and their respective metadata and lexical data → importance of following compatible standards and formats such as LMF (ISO 24613: 2008), TEI Lex-0 (Tasovac and Romary et al., 2018) and Ontolex-Lemon (McCrae et al., 2017).

Historical background

- *Diccionario da Lingua Portugueza* by António de Morais Silva was elaborated during the Age of Enlightenment.
- The eighteenth century brought a renewal in several fields of knowledge, namely those concerning the description of living languages, at a time when Latin was still the language of instruction.
- The publication of the Morais dictionary in 1789 inaugurated modern Portuguese lexicography.

Morais dictionary

- Morais does not claim to be the author, assigning this condition to Bluteau, author of the *Vocabulario Portuguez and Latino.*
- Morais recognises in the '*Prólogo ao Leitor*' [Prologue to the Reader] that the additions he brought to the dictionary are quite relevant.
- Morais represents the first modern work to systematise the lexicon of the Portuguese language, a model and example for all the ones that followed.
- The first edition was first published in two volumes: first, from the letters A to K, in a total of 752 pages, and then, from the letters L to Z, with 541 pages.
- The following two editions (1813; 1823) are considered new dictionaries, due to both their enrichment and the updating.


DICCIONARIO
DA
LINGUA PORTUGUEZA
COMPOSTO
PELO PADRE
D. RAFAEL BLUTEAU,
REFORMADO, E ACCRESCENTADO
POR
ANTONIO DE MORAES SILVA
NATURAL DO RIO DE JANEIRO.
TOMO PRIMEIRO.
A—K
LISBOA,
NA OFFICINA DE SIMÃO THADDEO FERREIRA.
ANNO M. DCC. LXXXIX.
Com Licença da Real Meza da Commissão Geral, sobre o Exame, e Censura dos Livros.
Vende-se na loja de Borel Borel, e Companhia, quasi defronte da Igreja nova de Nossa Senhora dos Martyres, na esquina.

Frontispiece of Morais (1789), first volume

MORDigital – The project (1)

Aims of the project:

- to encode the selected editions of *Diccionario de Lingua Portugueza* by António de Morais Silva MORDigital – Digitalização do *Diccionario da Lingua Portugueza* de António de Morais Silva;
- to promote accessibility to cultural heritage while fostering reusability and contributing towards a greater presence of lexicographic digital content in Portuguese through open tools and standards;
- to connect data and metadata within the same lexicographic resource and between different resources, through the Web of Data;

MORDigital – The project (2)

Aims of the project:

- to concentrate on the linguistic and lexicographic knowledge that permeates the entire project and contributes to the necessary systematisation of data and metadata;
- to add value of bringing a historical resource into the LLOD cloud in a language (Portuguese) that is still underrepresented;
- to put forward a methodology that can be replicated in other legacy paper dictionaries, using tools that allow the automatic extraction of lexicographic content, as well as the modernisation of the spelling in an automated way.

MORDigital – Methodology (1)

- To analyse all components that comprise the dictionary's macro- and microstructure;
- To identify, organise and describe the different levels of linguistic knowledge to apply the aforementioned standards systematically;
- To develop methodologies that can be replicated for other applications and test the alignment of the different encodings of Morais;
- To participate in reviewing the corresponding standards as members of the standard bodies and scientific forums;
- To propose best practices for harmonising the encoding of lexicographic resources;
- To make Morais available via an open-access platform.

MORDigital – Methodology (2)

- Our methodology is based on 5 central axes:
    - high-quality retrodigitisation of Morais and automatic structuring of the lexical content for the creation of a computer-readable resource;
    - lexicographically-oriented language description;
    - Morais encoding, using the TEI Lex-0 specifications mapped to the LMF standard and their respective serialisations, as well as to OntoLex-Lemon;
    - creation of an ontology for alignment purposes;
    - conception of a platform for Morais, enriched with both lexicographic and ontological modules.

MORDigital – Methodology (3)

- transforming the native encoding format into a TEI/XML compliant one (the encoding will be based on TEI standards according to the TEI Lex-0 specification) and LMF metamodels into advanced techniques for semi-structured text acquisition → The result will be a model of a historical dictionary whose entries are structured in a standard format, namely TEI Lex-0;
- the creation of an ontology of all the previously identified and systematised labels (e.g. domain, register, grammar, among others). This will be implemented by resorting to Protégé, a free, open-source ontology editor. The ontology will be represented in OWL.
- alignment of the dictionary versions, which will be carried out in stages: i) alignment of the entries; ii) alignment of the senses; iii) alignment of other lexicographic content.
- build a platform that integrates all Morais versions while also mapping the different heterogeneous annotation models, in order to provide access to high-quality digital lexicographic content enhanced by ontologies.

MORDigital – Methodology (4)

- We aim for our lexical resources to maintain the original spelling. However, making a resource available to the public today, and considering the prevalence of search engines, requires the modernisation of the spellings, especially at the lemma level.
- The original spelling of the lemma will have to be aligned with more current spellings. To this end, the original forms will be noted as a lemma, but we will first match them with the most current spellings and simultaneously work on their encoding in the XML annotation file.

MORDigital – End product

- We expect to have encoded a vital heritage dictionary, compliant with the most advanced standards for scholarly digital editions and made available via an open license.
- The versions will be accessible and searchable through an advanced interface, which will enable the selective querying of text by lemma and type of lexicographic content.
- The project will have significantly contributed towards the analysis and annotation of dictionaries through computer-assisted processes.

Concluding remarks (1)

- This project will represent a substantial contribution to the scientific community, aiming to create innovative and data-driven computational methods for text digitisation and encoding, based on a comprehensive analysis of lexicographic articles and their respective components.
- Tests on automatic text capture will refine processes and techniques, advancing the state of the art regarding semantic annotation of semi-structured documents. A rigorous linguistic treatment will make it possible to organise and structure the lexicographic components, and to elicit lexical relationships between various elements.
- The linking mechanisms of the resulting structured dictionary to other resources will constitute a prototype that can be replicated in other works, namely in the Portuguese-speaking world.

Concluding remarks (2)

- MOR*Digital* will be a user-friendly, open-access web interface, equipped with a robust research system that will not only facilitate the search on a more traditional lexicographic perspective but will also allow undertaking research on various types of structured lexicographic and terminological information (Costa et al., 2020).

- Combining semasiological and onomasiological approaches applied to the three editions of Morais will be possible via the inclusion of ontologies (e.g. diasystematic marking, namely domain labels, registers and part of speech categories).

- This method will make a new type of dictionary emerge which will contribute to creating a digital linguistic resource that is central to digital humanities. End-users will be predominantly scholars dealing with language and historical issues.

Acknowledgments

- MOR*Digital – 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

- 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

- European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS) (European Lexicographic Infrastructure)

Obrigada

Danke schön

Grazie

Merci

Thank you

Хвала вам

شكرا لك

Rute Costa: rute.costa@fcsh.unl.pt

Ana Salgado: anasalgado@campus.fcsh.unl.pt

Anas Fahad Khan: fahad.khan@ilc.cnr.it

Sara Carvalho: sara.carvalho@ua.pt

Laurent Romary: laurent.romary@inria.fr

Bruno Almeida: brunoalmeida@fcsh.unl.pt

Margarida Ramos: mvramos@fcsh.unl.pt

Mohamed Khemakhem: medkhemakhemfsegs@gmail.com

Raquel Silva: raq.asilva@gmail.com

Tomas Tasovac ttasovac@humanistika.org

CLUNL
CENTRO DE LINGUÍSTICA DA UNIVERSIDADE NOVA DE LISBOA

NOVAFCSH
FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS
UNIVERSIDADE NOVA DE LISBOA

FCT Fundação para a Ciência e a Tecnologia