Interlinking lexicographic data in the MORDigital project

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

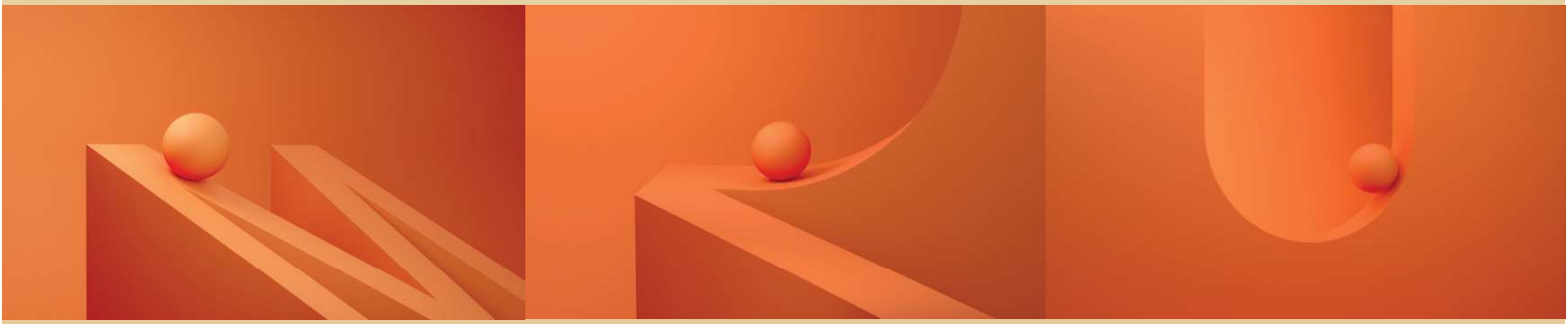
Mykolo Romerio universitetas

www.mruni.eu

LLOD APPROACHES FOR LANGUAGE DATA RESEARCH AND MANAGEMENT

LLODREAM2022

International Scientific Interdisciplinary Conference



LLOD APPROACHES FOR LANGUAGE DATA RESEARCH AND MANAGEMENT

LLODREAM2022

International Scientific Interdisciplinary Conference

Mykolas Romeris University

LLOD APPROACHES FOR LANGUAGE DATA RESEARCH AND MANAGEMENT

LLODREAM2022

International Scientific Interdisciplinary Conference

ABSTRACT BOOK

Supported by the NexusLinguarum COST Action CA18209

in cooperation with The Institute of Croatian Language

September 21-22, 2022

CONFERENCE SCIENTIFIC COMMITTEE

Dr. Florentina Armaselu, University of Luxembourg, Luxembourg

Prof. Dr. Anna Bączkowska, University of Gdansk, Poland

Dr. Ana Balula, University of Aveiro, Portugal

Dr Sara Carvalho, University of Aveiro, Portugal

Prof. Dr. Christian Chiarcos, Goethe University Frankfurt, Germany

Dr. Mariana Damova, Mozaika, Bulgaria

Dr. Milan Dojchinovski, Czech Technical University in Prague, Czech Republic

Dr. Olga Dontcheva-Navratilova, Masaryk University, Czech Republic

Dr. Radovan Garabík, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Slovakia

Dr. Daniela Gifu, Alexandru Ioan Cuza University of Iasi & Romanian Academy – Iasi branch, Romania

Dr. Dagmar Gromann, University of Vienna, Austria

Dr. Nomeda Gudelienė, Mykolas Romeris University, Lithuania

Dr. Gordana Hrzica, University of Zagreb, Hrvatska

Prof. Dr. Violeta Janulevičienė, Mykolas Romeris University, Lithuania

Dr. Mietta Lennes, University of Helsinki, Finland

Dr. Barbara Lewandowska-Tomaszczyk, State University of Applied Sciences in Konin, Poland

Dr. Chaya Liebeskind, Jerusalem College of Technology, Israel

Dr. Viktorija Mažeikienė, Mykolas Romeris University, Lithuania

Dr. Barbara McGillivray, University of Cambridge and The Alan Turing Institute, United Kingdom

Dr. Amália Mendes, University of Lisbon, Portugal

Prof. Dr. Odeta Merfeldaitė, Mykolas Romeris University, Lithuania

Dr. Jelena Mitrović, University of Passau, Germany

Prof. Dr. Liudmila Mockienė, Mykolas Romeris University, Lithuania

Dr. Hugo Gonçalo Oliveira, University of Coimbra, Portugal

Dr. Petya Osenova, Institute of Information and Communication Technologies, Bulgaria

Dr. Ana Ostroški Anić, Institute of Croatian Language and Linguistics, Croatia

Prof. Dr. Sigita Rackevičienė, Mykolas Romeris University, Lithuania

Dr. Jorge Gracia del Río, University of Zaragoza, Spain

Prof. Dr. Marko Robnik-Šikonja, University of Ljubljana, Slovenia

Dr. Eglė Selevičienė, Mykolas Romeris University, Lithuania

Prof. Dr. Linas Selmistraitis, Mykolas Romeris University, Lithuania

Dr. Gilles Sérasset, University Grenoble Alpes, France

Dr. Purificação Silvano, University of Porto, Portugal

Dr. Renato Rocha Souza, Austrian Academy of Sciences, Austria

Prof. Dr. Nadežda Stojković, University of Niš, Serbia

Prof. Dr. Giedrė Valūnaitė Oleškevičienė, Mykolas Romeris University, Lithuania

Prof. Dr. Vojtech Svatek, University of Economics, Czech Republic

Dr. Kristina Štrkalj Despot, Institute of Croatian Language and Linguistics, Croatia

Dr. Lora Tamošiūnienė, Mykolas Romeris University, Lithuania

Dr. Dimitar Trajanov, Faculty of Computer Science and Engineering, North Macedonia

Dr. Ciprian-Octavian Truica, Uppsala University, Sweden

Dr. Andrius Utka, Vytautas Magnus University, Lithuania

Dr. Vilhelmina Vaičiūnienė, Mykolas Romeris University, Lithuania

Dr. Sandra Vieira Vasconcelos, University of Aveiro, Portugal

Dr. Deniz Zeyrek, Middle East Technical University, Turkey

Dr. Slavko Žitnik, UL FRI, Slovenia

ORGANIZING COMMITTEE

Coordinators:

Dr. Radovan Garabík, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Slovakia

Prof. Dr. Giedrė Valūnaitė Oleškevičienė, Mykolas Romeris University, Lithuania

Members:

Dr. Viktorija Mažeikienė, Mykolas Romeris University, Lithuania

Prof. Dr. Liumila Mockienė, Mykolas Romeris University, Lithuania

Avigėja Novikovienė, Mykolas Romeris University, Lithuania

Dr. Ana Ostroški Anić, Institute of Croatian Language and Linguistics, Croatia

Prof. Dr. Sigita Rackevičienė, Mykolas Romeris University, Lithuania

Dr. Eglė Selevičienė, Mykolas Romeris University, Lithuania

Prof. Dr. Linas Selmistraitis, Mykolas Romeris University, Lithuania

Dr. Kristina Štrkalj Despot, Institute of Croatian Language and Linguistics, Croatia

Dr. Lora Tamošiūnienė, Mykolas Romeris University, Lithuania

Olga Usinskiene, Mykolas Romeris University, Lithuania

Dr. Vilhelmina Vaičiūnienė, Mykolas Romeris University, Lithuania

KEYNOTE SPEAKERS

Dr. Dagmar Gromann, University of Vienna, Austria

Dr. Jorge Gracia del Río, University of Zaragoza, Spain

ISBN 978-609-488-041-4

CONTENTS

Keynote Presentation Abstracts

Dagmar Gromann. ACQUIRING TERMINOLOGICAL RELATIONS WITH NEURAL MODELS FOR MULTILINGUAL LLOD RESOURCES / 10

Jorge Gracia. LINKED DATA AS A CORNERSTONE OF LINGUISTIC DATA SCIENCE / 11

Presentation Abstracts

Alfonso Rascón Cabellero. ELECTRONIC LEXICOGRAPHY: BETWEEN INFORMATION OVERLOAD AND USER-FRIENDLINESS / 12

Anas Fahad Khan, Rute Costa, Sara Carvalho, Laurent Romary, Bruno Almeida, Margarida Ramos, Mohamed Khemakhem, Raquel Silva, Toma Tasovac. INTERLINKING LEXICOGRAPHIC DATA IN THE MORDIGITAL PROJECT / 14

Andrea Bellandi, Fahad Khan, Monica Monachini, Valeria Quochi. A LEXO-SERVER USE CASE: LANGUAGES AND CULTURES OF ANCIENT ITALY / 16

Anna Bączkowska, Dagmar Gromann. FROM KNOBHEAD TO SEX GODDESS: SWEARWORDS IN ENGLISH SUBTITLES, THEIR FUNCTIONS AND REPRESENTATION AS LINGUISTIC LINKED DATA / 18

Anna Bączkowska, Barbara Lewandowska-Tomaszczyk, Slavko Žitnik, Giedre Valunaite-Oleskeviciene, Chaya Liebeskind, Marcin Trojszczak. IMPLICIT OFFENSIVE LANGUAGE TAXONOMY AND ITS APPLICATION FOR AUTOMATIC EXTRACTION AND ONTOLOGY / 20

Barbara Lewandowska-Tomaszczyk, Slavko Žitnik, Liebeskind, Chaya, Giedre Valunaite-Oleskevicienė, Anna Bączkowska, Paul A. Wilson, Marcin Trojszczak, Ivana Brač, Lobel Filipić, Ana Ostroški Anić, Olga Dontcheva-Navratilova, Agnieszka Borowiak, Kristina Despot, Jelena Mitrović. ANNOTATION SCHEME AND EVALUATION: THE CASE OF OFFENSIVE LANGUAGE / 23

Christian Chiarcos, Maxim Ionov, Katerina Gkirtzou, Anas Fahad Khan, Penny Labropoulou, Marco Passarotti, Matteo Pellegrini. ONTOLEX-MORPH: MORPHOLOGY FOR THE WEB OF DATA /26

Christian Chiarcos, Purificação Silvano, Mariana Damova, Giedrė Valūnaitė-Oleškevicienė, Chaya Liebeskind, Dimitar Trajanov, Ciprian-Octavian Truica, Elena-Simona Apostol,Anna Baczkowska. AN OWL ONTOLOGY FOR ISO-BASED DISCOURSE MARKER ANNOTATION / 28

Danguolė Straižytė, Paul Gregor Droessiger. WORD-FORMATION PATTERNS OF NOMINA LOCI (PLACE NAMES) IN GERMAN, ENGLISH, AND LITHUANIAN: A CASE STUDY OF GRIMMS' FAIRY TALES / 31

Daria Koloda. GENERAL CHARACTERISTICS OF LANGUAGE INTERFERENCE IN SOCIAL MEDIA IN THE CONTEXT OF RUSSIA'S AGGRESSION AGAINST UKRAINE / 33

David Sanchez, Thomas Louf, Jose J. Ramasco. MULTILINGUAL SOCIETIES FROM TWITTER DATA: EMPIRICAL ANALYSIS AND THEORETICAL MODELLING / 35

Eglė Selevičienė. THE USE OF FREE WEB-BASED MIND MAPPING TOOLS IN THE STUDIES OF ESP: A REVIEW OF RESEARCH / 37

Eugénie Bonner-Bestchastnova, Antoinette Bestchastnova. MOTIVATION, CORRECTION, FOREIGN LANGUAGE ACQUISITION THROUGH THE USE OF PSYCHOLINGUISTIC TESTS / 39

Florentina Armaselu, Elena-Simona Apostol, Christian Chiarcos, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truică, Giedrė Valūnaitė-Oleškevičienė. TRACING SEMANTIC CHANGE WITH MULTILINGUAL LLOD AND DIACHRONIC WORD EMBEDDINGS / 41

Giedrė Valūnaitė-Oleškevičienė, Gražina Čiuladienė, Lora Tamošiūnienė, Liudmila Mockienė. PRACTICES OF ONLINE LANGUAGE TEACHING AND LEARNING: A SURVEY IN LITHUANIA / 43

Giedrė Valūnaitė-Oleškevičienė, Lora Tamošiūnienė, Gražina Čiuladienė. DIAL4U: DIGITAL PEDAGOGY TO DEVELOP AUTONOMY, MEDIATE AND CERTIFY LIFEWIDE AND LIFELONG LANGUAGE LEARNING FOR (EUROPEAN) UNIVERSITIES / 45

Giedrė Valūnaitė-Oleškevičienė, Chaya Liebeskind. PROCESSING MULTI-WORD DISCOURSE MARKERS IN TRANSLATION: ENGLISH TO HEBREW AND LITHUANIAN / 47

Gordana Hržica, Sara Košutar, Dario Karl, Matea Kramarić. SELECTION, IMPLEMENTATION AND TESTING OF LANGUAGE SAMPLE ANALYSIS MEASURES FOR THE WEB-BASED APPLICATION MULTIDIS / 49

Kris Heylen, Ilan Kernerman, Carole Tiberius. LINKING LEXICOGRAPHIC RESOURCES TO LANGUAGE PROFICIENCYLEVEL APPLICATIONS / 51

Livija Puodžiūnaitė, Giedrė Valūnaitė-Oleškevičienė. LITHUANIAN TRANSLATION OF THE DISCOURSE MARKER AND IN SOCIAL MEDIA TEXTS / 53

Marco Passarotti, Francesco Mambrini. ISSUES IN BUILDING THE LILA KNOWLEDGE BASE OF INTEROPERABLE LINGUISTIC RESOURCES FOR LATIN / 55

Maryna Bielova. VERBO-VISUAL PUN IN MEMETIC WARFARE AGAINST RUSSIA'S AGGRESSION IN UKRAINE / 57

Olena Kholodniak. THEORETICAL BASICS OF STYLIZED COLLOQUIAL SPEECH FUNCTIONING IN WORKS OF FICTION / 58

Olga Usinskiene, Sigita Rackevičienė. ENGLISH AND LITHUANIAN TERMS IN THE PARALLEL MIGRATION CORPUS / 60

Penny Labropoulou, David Lindemann, Christiane Klaes, Katerina Gkirtzou. THE LEXMETA METADATA MODEL FOR LEXICAL RESOURCES: THEORETICAL AND IMPLEMENTATION ISSUES / 62

Sigita Rackevičienė, Andrius Utka, Liudmila Mockienė, Agnė Bielinskienė. DEVELOPING A CYBERSECURITY TERMBASE / 64

Thierry Declerck. TOWARDS THE INTEGRATION OF SIGN LANGUAGES DATA IN THE LINGUISTIC LINKED OPEN DATA CLOUD / 66

Vilhelmina Vaičiūnienė, Akvilė Šimėnienė. LITHUANIAN POETRY TRANSLATION TO SPANISH: A REVIEW OF TWO DECADES (2000–2021) / 68

Vitalija Jankauskaitė-Jokūbaitienė. THE EFFECTIVENESS OF VIDEO CREATION IN THE ESL CLASSROOM IN LITHUANIA: A CASE STUDY / 70

Yuliia Shpak, Ganna Krapivnyk. STYLISTIC MEANS OF VERBALIZING IMAGES OF THE RUSSIAN-UKRAINIAN WAR / 72

Workshop Abstracts

Emma Angela Montechiari, Stanko Stankov, Kostadin Mishev, Mariana Damova. MACHINE LEARNING METHODS FOR DISCOURSE MARKER DETECTION IN ITALIAN / 74

Hugo Gonçalo Oliveira. EVALUATING SYNONYM AND ANTONYM ACQUISITION FROM A PORTUGUESE MASKED LANGUAGE MODEL / 81

Lucía Pitarch, Lacramioara Dranca, Jorge Bernad, and Jorge Gracia. LEXICO-SEMANTIC RELATION CLASSIFICATION WITH MULTILINGUAL FINETUNING / 86

Martina Kramarić. EXTRACTING AND LINKING MORPHOLOGICAL DATA FROM THE PRE-STANDARD CROATIAN GRAMMARS USING TEI /89

Radovan Garabík, Denis Mitana. ACCURACY OF SLOVAK LANGUAGE LEMMATIZATION AND MSD TAGGING – MORPHODITA AND SPACY / 93

► 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

The publication is based upon work from the NexusLinguarum COST Action CA18209, supported by COST (European Cooperation by Science and Technology). COST is a Pan-European intergovernmental framework. Its mission is to enable breakthrough scientific and technological development leading to new concepts and products and thereby contribute to strengthening Europe's research and innovation capacities.

Edited by MB Kopis
Layout by Jovita Jankauskienė

Published by:
Mykolas Romeris University
Ateities st. 20, LT-08303 Vilnius, Lithuania
www.mruni.eu


Nexus
Linguarum


cost
EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY