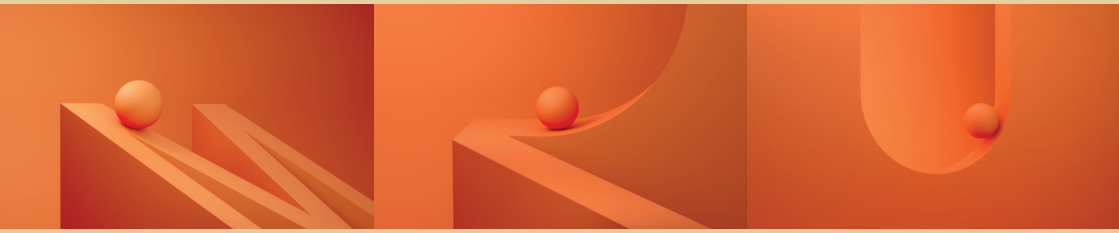
Mykolo Romerio universitetas

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

Presentation Abstracts

Workshop Abstracts

KEYNOTE PRESENTATION ABSTRACTS

► Acquiring Terminological Relations with Neural Models for Multilingual LLOD Resources

Dagmar Gromann,
University of Vienna, Austria, dagmar.gromann@univie.ac.at

Specialized communication strongly benefits from the availability of structured and consistent domain-specific knowledge in LLOD language resources. Manually curating such language resources is cumbersome and time-intensive. Thus, automated approaches for extracting terms, concepts, and their interrelations are required. Recent advances in computational linguistics have enabled the training of highly multilingual neural language models, such as GPT-3 or XLM-R, that can successfully be adapted to various downstream tasks, from sentiment classification and text completion to information extraction. Furthermore, several approaches exist to extract and explore lexico-semantic relations by means of these language models, however, only few focus on curating, representing, and interchanging domain-specific language resources in the LLOD cloud.

In this talk, we will explore recent work on creative methods to acquire LLOD resources by utilizing pre-trained language models with a particular focus on lexico-semantic and terminological relations. Methods to creatively utilize pre-trained multilingual neural language models, such as GPT-3 or XLM-R, to acquire and extract such relations range from sequence classification to question answering. On the other hand, this talk will discuss how existing LLOD resources can contribute to the task of neural relation acquisition. Any such methods require considerations of scalability and adaptability to domains and languages not considered during training. One objective of the LLOD community is to provide highly multilingual fully interoperable and interlinked language resources. Low-resource languages represent a particular challenge in this regard. Thus, we will further discuss how the proposed methodological combination of neural language models and LLOD technologies can support the acquisition, publication, and interoperability low-resource language data and resources.. In addition, their ease of use for and readiness of uptake by different communities, such as language experts, terminologists, domain experts, and the LLOD community, will strongly influence their final success. The question in this regard is how easy it is for, e.g. linguists, to apply these methods to languages and data of their choice and to evaluate the reliability of the results. To this end, some examples of easy to use technologies to acquire relations will be presented.

Keywords: *linguistic linked data, terminological relations, neural language models*

Linked Data as a Cornerstone of Linguistic Data Science

Jorge Gracia,
University of Zaragoza, Spain, jogracia@unizar.es

At present, we are witnessing incredible advancements in language technologies, natural language processing, and related fields, mostly stimulated by the success of deep learning technologies and the increasing availability of huge amounts of textual data on the Web. Algorithms and models that were commonly used a few years ago are rapidly substituted by newer, more effective ones, with little time for researchers and practitioners to adapt to them. In such an evolving and challenging scenario, what is the role of linguistic data science? We understand linguistic data science as a subfield of data science which focuses on the systematic analysis and study of the structure and properties of linguistic data at a large scale, along with methods and techniques to extract new knowledge and insights from it. To that end, it is necessary to provide a formal basis to the analysis, representation, integration, and exploitation of such data at their different levels (syntax, morphology, lexicon, etc.).

In this talk, we will try to answer the previous question, for which we first have to consider the difference between textual data, where current trends put the focus, and linguistic data (with their linguistic features explicitly represented). A reflection on the type of problems that can be solved with the use of the latter type of data is needed. We will also discuss how linked data technologies can be essential to provide a basis for linguistic data science, by enabling an ecosystem of multilingual and semantically interoperable linguistic data at Web scale. We will review recent advancements and open challenges in the field of linguistic linked data, mentioning recent studies carried out in the context of the NexusLinguarum COST Action on that matter. Among other things, they identify the main current challenges that the linguistic linked data community needs to face to maximize the adoption of their technologies. These include the need to break some entry barriers to enable non-experts to adopt the technology seamlessly, the need for more sustainable hosting solutions, as well as developments targeted to increase multilinguality and support for under-resourced languages on the Web. Finally, some examples of how linguistic linked data technologies are currently being exploited for research in linguistics and industrial applications will be provided.

Keywords: *linguistic data science, linked data, linguistic linked data*

PRESENTATION ABSTRACTS

▶ Electronic Lexicography: Between Information Overload and User-Friendliness

Alfonso Rascón Cabellero,
Vilnius University
Lithuania
alfonso.rascon@flf.vu.lt

Purpose: This study aims to provide guidelines for the user-friendly display of lexicographical information in electronic dictionary entries. Particular attention is paid to the example as it illustrates the generally metalinguistic, abstract information provided by other components (definition or equivalent, morphological indications, syntactic indications, indications of lexical combination, etc.), and conveys complementary information.

Design/methodology/approach: A comprehensive analysis of the examples of the letter L in the print Lithuanian-Spanish/French/Italian dictionaries was carried out according to their form (condensed and sentences) and functions: to illustrate the meaning, to show its use in a specific syntactic environment, its combination with other words (collocations), and the relationship of the use of the lemmatized word with the real context of the situation (pragmatic aspects). On the other hand, bilingual and monolingual electronic dictionaries were used as a reference in order to establish how to distribute and dose information between the components so that it is exhaustive, but in an effective format and design.

Findings: This analysis suggests that more systematic, predictable, and generalizable information on word usage can be presented through components such as semantization, cotext indication (collocators), syntactic construction, and condensed examples, while more idiosyncratic and unpredictable uses should preferably be shown with sentence examples, but also with explanatory glosses. A condensed type of example that is especially suitable for electronic dictionaries is proposed: the translated cotext (collocator), which not only serves as a discriminator of equivalents, but also shows the basic combinatorics of the lemmatized word in both languages.

Research limitations/implications: Electronic media, which do not have the space limitations that have always restricted printed dictionaries, can present much more information, but can also lead to information overload that causes cognitive fatigue for users.

Practical implications: Lexicographers, together with the computer specialists who design electronic dictionaries, must strive to present information in a user-friendly way, without compromising on quality.

Originality/Value: This study is based on research on the bilingual lexicographical example and focuses directly on the development of bilingual electronic dictionaries taking into account the electronic format, which facilitates the search for specific information.

Keywords: *Dictionary example, dictionary entry components, electronic lexicography, language and speech.*

Research type: Research paper.

► 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

▶ A LexO-Server Use Case: Languages and Cultures of Ancient Italy

Andrea Bellandi, Fahad Khan, Monica Monachini, Valeria Quochi
Institute for Computational Linguistics "A. Zampolli", Italy
{name.surname}@ilc.cnr.it

Purpose: This paper presents a set of REST services called a LexO-server (https://github.com/andreabellandi/LexO-) for the management of lexical resources modeled as the *OntoLex-Lemon* model. This comes as a software backend, providing data access and manipulation to frontend developers, and will be exemplified through the Languages and Cultures of Ancient Italy CLARIN-IT related project. This is a use case where the creation and edition of an integrated system of LRs for ancient fragmentary languages will be shown in compliance with current digital humanities and Linked Open Data principles. Other relevant use cases will be mentioned for demonstrating the versatility of these services and how they can be easily integrated within more complex systems and/or interact with other independent back-ends.

Design/methodology/approach: A LexO-server is a system based on Service-Oriented Architecture (SOA), an architectural style where services are provided to the other components by application components, through a communication protocol (REST in our case) over a network. Each service is a discrete unit of functionality that can be accessed remotely and acted upon and updated independently, such as by retrieving a lexical entry or adding a lexical sense to a lexical entry. According to the aim of the *OntoLex-Lemon* model, that is the enrichment of conceptual ontologies with linguistic information, a LexO-server provides services at both the lexical and conceptual levels. A LexO-server natively provides services for the management of SKOS ontologies only, but it makes possible lexical entries refer to external existent OWL ontologies.

Findings: A LexO-server is a free and open source backend that relies on the GraphDB semantic repository. It is implemented as a set of Representational State Transfer (REST) services based on the HTTP protocol, and exchanges data in JSON format. Services conform to OpenAPI, a specification for machine-readable interface files to describe, produce, consume, and display REST services.

Research limitations/implications: On the one hand, a LexO-server allows the target of *OntoLex-Lemon* users to be broadened, while on the other it opens up the possibility for the construction of applications oriented at different tasks, such as editing, linking, dictionary making, and linguistic annotation.

Practical implications: Service orientation architectures allow a strong frontend-backend separation of application concerns to be maintained in a way that makes

most services potentially reusable in different contexts. This allows developers to build different end-user applications on the same backend.

Originality/Value: Some similar initiatives have been developed. Among them, K Dictionaries by the University of Cambridge, and REST services provided by the University of Oxford for accessing their dictionaries. Another experience worthy of note is a REST interface developed in the context of the ELEXIS infrastructural project for accessing collections of monolingual and multilingual resources with a broad range of usages. Differently from these, the LexO-server also provides editing services, in order to serve a wider set of possible tasks for the development of dedicated applications by third parties.

Keywords: *LLOD, OntoLex-Lemon, digital historical linguistics, lexicon web services.*

Research type: Technical paper.

► From Knobhead to Sex Goddess: Swearwords in English Subtitles, Their Functions and Representation as Linguistic Linked Data

Anna Bączkowska,
English Language and Theoretical Linguistics Division, Institute of English and American Studies, University of Gdańsk, Poland, anna.baczkowska@ug.edu.pl
Dagmar Gromann,
Centre for Translation Studies, University of Vienna, Austria, dagmar.gromann@univie.ac.at

Purpose: Swearwords represent an important social vehicle for human communication that, beyond mere insults, are conventionally used for expressing solidarity, bonding, and banter, among other functions. Representing an empirically validated typology of such functions and annotated swearwords in subtitles as Linguistic Linked Open Data (LLOD) provides a rich linguistic research platform for foul language. As a first case study, we classify swearwords in the English subtitles of Bridget Jones's Diary into twelve unique functions and represent them with OntoLex-Lemon (Cimiano et al., 2016).

Design/methodology/approach: Extending existing lists of swearwords (e.g., as proposed by Dewaele, 2016) with, e.g., Wiktionary lists. This approach automatically detects swearwords in English subtitles by means of a profanity checker. Two linguistic experts annotated each extracted example with functions from the proposed typology, including pain, flirting/teasing, banter, and giving advice. The annotated primary corpus data and their metadata and functions are represented as LLOD resources with the OntoLex-Lemon model and OntoLex-Frac module for corpus data, extolling its usefulness with some sample exploration of the LLOD swearwords.

Findings: The initial calculation of the kappa value provided moderate inter-annotator agreement of 0.429, upon which we refined the definitions of the functions and discussed examples to generate a gold standard. The function typology thereby created is easily reusable as an LLOD resource, for which we provide sample uses on the proposed dataset.

Research limitations/implications: This initial case study is limited to subtitles of a single, yet rich, movie, which naturally limits the variety of swearwords explored in this first approach. Nevertheless, it provides an excellent first testbed for porting foul language and its diverse functions to the LLOD.

Practical implications: Representing functionally classified swearwords as LLOD allows, for instance, for exploring relations between swearwords, frequencies of

functions, and the attitudes of movie characters with simple queries.

Originality/Value: To the best of our knowledge, this is the first proposal to represent swearwords and their functions as LLOD, which provides an interesting use case for existing modeling approaches for the LLOD community, and an incentive for linguists investigating offensive language to benefit from the interoperability and ease of reuse of LLOD resources.

Keywords: *swearwords, functions, subtitles, Linguistic Linked Open Data*

Research type: Research paper

Bibliography

Cimiano, P., McCrae, J. P., & Buitelaar, P. (2016). *Lexicon Model for Ontologies: Community Report*. W3C Community Group.

Dewaele, J. M. (2016). Thirty shades of offensiveness: L1 and LX English users' understanding, perception and self-reported use of negative emotion-laden words. *Journal of Pragmatics*, *94*, 112–127.

► Implicit Offensive Language Taxonomy and Its Application for Automatic Extraction and Ontology

Anna Bączkowska,
University of Gdansk, Poland, University of Luxembourg, Luxembourg, e-mail: anna.baczkowska@ug.edu.pl
Barbara Lewandowska-Tomaszczyk,
State University of Applied Sciences in Konin, Poland, e-mail: barbara.lewandowska.tomaszczyk@gmail.com
Slavko Žitnik,
University of Ljubljana, Slovenia, e-mail: Slavko.Zitnik@fri.uni-lj.si
Giedre Valūnaitė-Oleškevičienė,
Mykolas Romeris University, Lithuania, e-mail: gentrygiedre@gmail.com
Chaya Liebeskind,
Jerusalem College of Technology, Israel, e-mail: liebchaya@gmail.com
Marcin Trojszczak,
University of Bialystok, University of Applied Sciences in Konin, e-mail: marcintrk@gmail.com

Purpose: In this current study, we intend to explore varying forms of implicit (mostly figurative) offensiveness (e.g., irony, metaphor, hyperbole, etc.) in order to propose a linguistic taxonomy of implicit offensiveness (and how it permeates explicit forms), and an ontology of offensive terms readily applicable to fine-tuned, pre-trained language models (word and phrase embedding). Offensive language has recently attracted great attention from computational scientists (e.g., Zampieri et al., 2019) and linguists alike (e.g., Haugh & Sinkeviciute, 2019). While in NLP scholars focus on ways of automatic extraction of what is generally and most often referred to as toxic language, in linguistics the concept of hate speech is frequently explored. Implicit offensive language, however, as opposed to explicit offence, has received little scholarly attention which so far has focused solely on single and unrelated concepts/terms. This paper aims at proposing an overarching model where varying subtypes of implicitness used in the context of offensive language are conceptually linked (Bączkowska et al., 2022).

Design/methodology/approach: The linguistic model of implicit offensive language results from a thorough literature review on implicit language seen from three perspectives: Gricean, post-Gricean, and neo-Gricean. Resulting from the analysis of existing typologies and definitions, a new model embedded mostly in a neo-Gricean approach to implicitness has been proposed. Offensiveness, on the other hand, is anchored in current approaches to offensive as well as impolite language (Culpeper, 2011, 2021; Haugh & Sinkeviciute, 2019). This taxonomy is further validated by computational methods (word and phrase embedding) aiming at finding an algorithm to cluster sim-

ilar concepts/terms and show existing dependencies among select word clusters. This study is conducted in line with the focus and approach adopted within the framework of COST ACTION CA 18209, European Network for Web-centred Linguistic Data Science (NexusLinguarum).

Findings: The validation of the linguistic model of implicit offensive language conducted by means of computational approaches to language analysis based on neural networks generally supports the linguistic taxonomy proposed in the preliminary model (Bączkowska et al., 2022). Linguistically, the research proves that the implicit model of implicitness and the explicit forms of offensiveness we proposed in our earlier model (Lewandowska-Tomaszczyk et al., 2021; Bączkowska, 2021; Lewandowska-Tomaszczyk et al., 2022; Žytnik et al., in press) are intertwined.

Research limitations/implications: The research results are dependent to a large extent on the choice and size of datasets used for the word and phrase embedding as well as the methods used for embedding (FastText, word2vec, Glove, ELMo, BERT), which are taken into account in our analysis. The implications of the study are easily transferable to offensive language annotation practice and to Linguistic Linked Data.

Practical implications: The taxonomy proposed here can be readily applied to other languages as well as to real linguistic data in order to implement automatic detection of offensive language in discourse, in particular online discourse (Twitter, Facebook, etc.). The taxonomy has already been used for linguistic data annotation with the aid of a semantic annotation tool INCEpTION (https://github.com/inception-project/inception) as part of Cost Action WG 4.1.1. Incivility in Media and Social Media.

Originality/Value: Even though the topic of offensiveness has received some attention both in the realm of linguistics and computer science, the terms ascribed to offensiveness are not well-defined and the relations among them (such as abusive, bullying, profane, obscene, insulting, etc.) rarely go beyond ad-hoc typologies and "non-systematic lexicography" (Goddard, 2018, p. 498). Our study marshals the terms that refer to various forms of offensiveness, shows relations held among them, and validates the proposed taxonomy by resorting to computational methods of language analysis. The study is thus original in the choice of methods used and the depth and breadth of concepts/terms involved in building the model of implicit offensiveness.

Keywords: *implicitness, offensiveness, embeddings, NLP.*

Research type: Research paper.

References

Bach, K. (1994). Conversational impliciture. *Mind and Language*, *9*(2), 124–162.

Bączkowska, A. (2021). "You're too thick to change the station" – Impoliteness, insults and responses to insults on Twitter. *Topics in Linguistics*, *22*(2), 62–84.

Bączkowska, A., Lewandowska-Tomaszczyk, B., Valunaite-Oleškevičiene, G., Žitnik, B., Liebeskind, C. (2022). *Jerusalem workshop: A taxonomy of implicit language*.

Culpeper, J. (2011). *Impoliteness: Using Language to Cause Offence*. Cambridge: CUP.

Culpeper, J. (2021). Impoliteness and hate speech: Compare and contrast. *Journal of Pragmatics*, *179*, 4–11.

Goddard, K. (2018). "Joking, kidding, teasing": Slippery categories for cross-cultural comparison but key words for understanding Anglo conversational humor. *Intercultural Pragmatics*, *15*(4), 487–514.

Haugh, M., & Sinkeviciute, V. (2019). Offence and conflict talk. In: M. Evans, L. Jeffries and J. O'Driscoll (Eds.), *The Routledge handbook of language in conflict* (pp. 196–214). Abingdon, Oxon, UK: Routledge.

Lewandowska-Tomaszczyk, B. et al. (2021). LOD-connected offensive language ontology and tag set enrichment. In S. Carvalho & R. Rocha Souza (Eds.), *LDK Workshops and Tutorials 2021* (Vol. 3064).

Lewandowska-Tomaszczyk, B., Liebeskind, C., Žitnik, B., Bączkowska, A., Valunaite-Oleškevičiene, G., & M. Trojszczak (2022). *An offensive language taxonomy and a webcorpus discourse analysis for automatic offensive language identification*. Presentation at 3rd International Conference: Approaches to Digital Discourse Analysis (ADDA 3). St Petersburg, Florida May 13–15, 2022.

Žitnik, B., Lewandowska-Tomaszczyk, B., Bączkowska, A., Liebeskind, C., Valunaite-Oleškevičiene, G., & Mitrovic, J. (in press). *Detecting Offensive Language: A new approach to offensive language data preparation*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 1415–1420).

► Annotation Scheme and Evaluation: The Case of OFFENSIVE Language

Barbara Lewandowska-Tomaszczyk,
State University of Applied Sciences in Konin, Poland
Slavko Žitnik,
University of Ljubljana, Slovenia
Chaya Liebeskind,
Jerusalem Institute of Technology, Israel
Giedre Valūnaitė-Oleškevičienė,
Mykolas Romeris University, Vilnius, Lithuania
Anna Bączkowska,
University of Gdansk, Poland
Paul A. Wilson,
University of Lodz, Poland
Marcin Trojszczak,
State University of Applied Sciences in Konin, University of Bialystok, Poland,
Ivana Brač,
Lobel Filipić,
Ana Ostroški Anić,
Institute of Croatian Language and Linguistics, Zagreb, Croatia
Olga Dontcheva-Navratilova,
Masaryk University, Brno, Czech Republic
Agnieszka Borowiak,
State University of Applied Sciences in Konin, Poland
Kristina Despot,
Institute of Croatian Language and Linguistics, Croatia
Jelena Mitrović,
University of Passau, Germany; Institute for AI R&D of Serbia

Purpose: Offensive discourse refers to the presence of explicit or implicit verbal attacks towards individuals or groups and has been extensively analyzed in linguistics (e.g., Culpeper, 2005; Haugh & Sinkeviciute, 2019) and in NLP (e.g., OffensEval (Zampieri et al., 2020), HASOC (Mandl et al., 2019)), under the names of *hate speech*, *abusive language*, *offensive language*, etc. The paper focuses on the presentation and discussion of aspects of the linguistic annotation of OFFENSIVE LANGUAGE, including creation, annotation practice, curation, and evaluation of an OFFENSIVE LANGUAGE annotation taxonomy scheme first proposed in Lewandowska-Tomaszczyk et al. (2021) and Žitnik et al. (in press). An extended offensive language ontology in terms of 17 categories, structured in terms of 4 hierarchical levels, has been shown to represent the encoding of the defined offensive language schema, trained in terms of non-contextual

word embeddings – i.e., Word2Vec and Fast Text – and eventually juxtaposed to the data acquired by using pairwise training and testing analysis for existing categories in the HateBERT model.

Approach: The system is used for annotation practice in WG 4.1.1. *Incivility in media and social media* in the context of COST Action CA 18209 *European network for Web-centred linguistic data science* (Nexus Linguarum) with the INCEpTION tool (https://github.com/inception-project/inception) – a semantic *annotation* platform offering assistance with annotation. The current authors are the taxonomy proposers, annotators, and curators of the annotation practice. We identify and discuss corresponding offensive category *levels* (types of offence target, etc.) and *aspects* (offensive language property clusters) as well as categories of *expressiveness* (*explicit – implicit*, *figurative language* types) in the data.

Findings: The results support the proposed ontology of explicit offense and positive implicitness types and a preliminary typology of more refined offensive implicitness categorization criteria to provide more variance among widely recognized types of figurative (metaphorical, metonymic, ironic, etc.) and other languages. The use of the annotation system and the representation of linguistic data will be evaluated in a series of annotators' comments, by means of a questionnaire method and in an open discussion.

Value: The results will be presented, and further developments in the annotation taxonomy creation and practice will be included in a **recommendation package** to be considered in new proposals, i.e., an implicit offense annotation system (Bączkowska et al., 2022; Despot & Ostroški Anić, 2022), and its possible application to other languages represented in the research team towards a subsequent LOD use.

Keywords: *annotation, automatic detection, offensive language taxonomy*

Research type: Research paper

References

Bączkowska, A., Lewandowska-Tomaszczyk, B., Valunaite Oleškevičiene, G., Žitnik, B., & Liebeskind, C. (2022). *Jerusalem workshop: A taxonomy of implicit language*.

Culpeper, J. (2005). Impoliteness and the Weakest Link. *Journal of Politeness Research*, *1*(I), 35–72.

Despot, K., & Ostroški Anić, A. (2022) *Jerusalem workshop: Reflections on Implicitness*.

Haugh, M., & Sinkeviciute, V. (2019). Offence and conflict talk. In *Routledge Handbook of Language in Conflict* (pp. 196–214). Routledge

Lewandowska-Tomaszczyk, B., Liebeskind, C., Žitnik, B., Bączkowska, A., Liebeskind, C., Valunaite-Oleškevičiene, G., & Mitrovic, J. (2021). LOD-connected offensive language ontology and tagset enrichment. *SALLD-2 2021 workshop LDK Proceedings*. Saragossa.

Lewandowska-Tomaszczyk, B., Liebeskind, C., Žitnik, B., Bączkowska, A., Valunaite-Oleškevičiene, G., & Trojszczak, M. (2022). *An offensive language taxonomy and a webcorpus discourse analysis for automatic offensive language identification.* Presentation at 3rd International Conference: Approaches to Digital Discourse Analysis (ADDA 3). St Petersburg, Florida May 13–15, 2022.

Mandl, S., Modha, P., Majumder, D., & Patel, D. (2019). Overview of the HASOCFire 2019: Hate speech and offensive content identification in Indo-European languages. *Proceedings of the 11.Forum for Information Retrieval Evaluation*. India.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 75–86). Minneapolis.

Žitnik, B., Lewandowska-Tomaszczyk, B., Bączkowska, A., Liebeskind, C., Valunaite-Oleškevičiene, G., & Mitrovic, J. (in press). *Detecting Offensive Language: A new approach to offensive language data preparation*.

▶ OntoLex-Morph: Morphology for the Web of Data

Christian Chiarcos,
Maxim Ionov,
Applied Computational Linguistics, Goethe University Frankfurt, Frankfurt am Main, Germany
Katerina Gkirtzou,
Institute for Language and Speech Processing, Athena Research Center, Athens, Greece
Anas Fahad Khan,
Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche, Italy
Penny Labropoulou,
Institute for Language and Speech Processing, Athena Research Center, Athens, Greece
Marco Passarotti,
Matteo Pellegrini,
CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan, Italy

Purpose: OntoLex-Lemon is a widely used community standard for publishing lexical resources in machine-readable form, and is in fact the predominant RDF vocabulary for this purpose. With the growing popularity and increasing adoption of this model for applications in both language technology and lexicography, a number of new modules have been developed in the past year to complement the OntoLex core vocabulary and its lexicographic follow up, lexicog. In this paper, we describe the current status of the development of the OntoLex-Morph vocabulary.

Design/methodology/approach: (1) Describe and motivate OntoLex-Lemon, (2) document its shortcomings with respect to morphology, (3) describe OntoLex-Morph, and (4) illustrate its application for modelling (a) morphological information from traditional print dictionaries, (b) inflection tables from grammar books, (c) inflection and word formation information in machine-readable dictionaries and computational lexicons, (d) morpheme inventories, (e) dictionaries used for morphological generation, (f) morphological generation rules for different languages.

Findings: We demonstrate that OntoLex-Morph is applicable to all these domains for modern and historical languages *on real-world data*.

Research limitations/implications: We show applicability across a wide range of requirements. However, this has so far been applied to well-researched and well-resourced languages, mostly from Europe. OntoLex-Morph has been designed to accommodate typologically different languages (e.g., agglutinative languages), but a large-scale

application to typologically diverse languages has not yet been demonstrated on a broad scale, due to the lack of resources for many morphologically rich languages. So far, examples considered from these languages are artificial and taken from books rather than real-world language resources, as these languages are notoriously under-resourced.

Practical implications: We consider the vocabulary mature enough to codify it into a W3C vocabulary. This talk aims to elicit a final round of feedback from the wider community before we begin to form it into a W3C Community Report. Once published, it will become an integral part of OntoLex and become a stable standard.

Originality/Value: Moderate originality, as we do not describe novel findings, but the consolidation of a model that has been under development for over 5 years. High value, as it marks the conclusion of the modelling process. Before the publication of the community report itself, a publication originating from this talk is likely to serve as a reference publication for modelling morphological resources with RDF and/or LLOD technology. It supersedes all prior publications on the topic.

Keywords: morphology, machine-readable dictionaries, OntoLex, standard development.

Research type: Conceptual paper

An OWL Ontology for ISO-Based Discourse Marker Annotation

Christian Chiarcos,
Applied Computational Linguistics, Goethe-Universität, Robert Mayer Straße 10 60325 Frankfurt am Main, Germany, chiarcos@cs.uni-frankfurt.de
Purificação Silvano,
Faculty of Arts and Humanities of the University of Porto, Centre of Linguistics of the University of Porto, Via Panorâmica, 4150-564 Porto, Portugal, msilvano@letras.up.pt
Mariana Damova,
Mozaika, Ltd., Solunska 52, Sofia 1000, mariana.damova@mozajka.co
Giedrė Valūnaitė-Oleškevicienė,
Mykolas Romeris University, Ateities 20, Vilnius, Lithuania, gvalunaite@mruni.eu
Chaya Liebeskind,
Jerusalem College of Technology, Havaad Haleumi 21, Givat Mordechai 91160, liebchaya@gmail.com
Dimitar Trajanov,
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, dimitar.trajanov@finki.ukim.mk
Ciprian-Octavian Truica,
Department of Information Technology, Uppsala University, Sweden, Lägerhyddsvägen 1, 75105 Uppsala, Sweden
Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Splaiul Independen¸tei nr. 313, sector 6, Bucuresti, Romania, ciprian-octavian.truica@it.uu.se
Elena-Simona Apostol,
Department of Information Technology, Uppsala University, Sweden, Lägerhyddsvägen 1, 75105 Uppsala, Sweden
Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Splaiul Independen¸tei nr. 313, sector 6, Bucuresti, Romania,elena.apostol@upb.ro
Anna Baczkowska,
Institute of English and American Studies, University of Gdansk, anna.k.baczkowska@gmail.com

Purpose: Discourse markers are linguistic cues that indicate how an utterance relates to the discourse context and what role it plays in conversation. The authors are preparing an annotated corpus in nine languages, and specifically aim to explore the role of Linguistic Linked Open Data (/LLOD) technologies in the process, i.e., the application of web standards such as RDF and the Web Ontology Language (OWL) for publishing and integrating data. We demonstrate the advantages of this approach.

Design/methodology/approach: (1) Survey of existing (community) standards for discourse (marker) annotation; (2) decision to follow ISO SemAF Discourse Relations (Core), using ISO SemAF Dialog Annotation standards [1,2] as a plugin; (3) provide annotation in a simplistic tabular format [3,4]; (4) formalize the annotation schema in an ontology [main contribution here]; and (5) convert annotations to RDF, link with ontology, and perform conjoint queries.

Findings: ISO SemAF discourse and dialog annotations can be formalized conjointly in a single OWL ontology. Some aspects needed to be systematically restructured. Most importantly, this concerns how ISO SemAF implements the subclassification of asymmetric discourse relations (as labels on arguments, not as subclasses of discourse relations). In discourse marker annotation, this needs to be converted into a flat hierarchy of relation types (more in line with previous research, e.g., Penn Discourse Treebank). A key advantage of OWL (in comparison to hierarchical annotation schemes) is that it supports fully fledged description logics, so we can annotate communicative functions and their qualifiers using logical (set/class) operators, e.g., *AcceptOffer Conditional*.

Research limitations/implications: Pilot annotations have been conducted for nine languages. The ontology is published under https://purl.org/olia/discourse/discourse.Nexus.owl. The LLOD transformation and linking of the *annotations* have not yet been performed. This paper provides the necessary prerequisites for doing so and discusses modelling challenges.

Practical implications: To the best of our knowledge, this is the first application of ISO SemAF for cross-lingual discourse marker annotation. Using LLOD technologies allows for the seamless integration of *any data* (here: annotations and ontologies), e.g., for querying.

Originality/Value: To the best of our knowledge, this is the first application of ISO SemAF for cross-lingual discourse marker annotation. Furthermore, this is one of the first concrete applications of LLOD technologies for implementing aspects of discourse annotation.

References:

[1] ISO. (2016). Language resource management- Semantic annotation framework (SemAF) - Part 8 - Semantic relations in discourse,core annotation schema (DR-core). Standard, Geneva, CH.

[2] ISO. (2020). Language resource management- Semantic annotation framework (SemAF) - Part 2 - Dialogue acts. Standard, Geneva, CH.

[3] Silvano, Purificação; Damova, Mariana; Oleškevičienė, Giedrė Valūnaitė;

Liebeskind, Chaya; Chiarcos, Christian; Trajanov, Dimitar; Truică, Ciprian-Octavian; Apostol, Elena-Simona & Baczkowska, Anna (2022). "ISO-Based Annotated Multilingual Corpus For Discourse Markers". In Proceedings of the 13th Edition Language Resources and Evaluation Conference (LREC 2022), pages 2739-2749. European Language Resources Association (ELRA). ACL anthology.

(http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.293.pdf

[4] Silvano, Purificação & Damova, Mariana (2022). ISO-DR-core plugs into ISO-dialogue acts for a crosslinguistic taxonomy of discourse markers. DiSLiDaS 2022 workshop, NexusLinguarum, Jerusalem, Israel, May 2022.Web Ontology Language (OWL) https://www.w3.org/OWL/

▶ Word-Formation Patterns of Nomina Loci (Place Names) in German, English, and Lithuanian: A Case Study of Grimms' Fairy Tales

Danguolė Straižytė,
Institute of the Foreign Languages, Faculty of Philology, Vilnius University, danguole.straizyte@flf.vu.lt
Paul Gregor Droessiger,
Faculty of Philology, Vilnius University
paul.droessiger@flf.stud.vu.lt

Purpose: The aim of this study is to conduct a comparative descriptive analysis of the grammatical word-formation systems for place nouns (hereinafter – *Nomina Loci*) in German, English, and Lithuanian. A further aim is to design and apply a potential framework that would assist in categorizing *Nomina Loci* more accurately and consequently reveal the true extent of this particular category of nouns.

Design/methodology/approach: A number of Grimms' fairy tales have been selected for the purpose of producing a data base of *Nomina Loci* from English, German, and Lithuanian. The source material was limited to 35 fairytales out of a total of 200. These 35 fairytales, however, are statistically representative of the entire set of Grimms' accrued tales. Furthermore, by choosing to work with the German, English, and Lithuanian versions of these tales, it is hoped to contribute to bringing Western European and Baltic linguistic traditions closer in a comparative manner.

Research limitations/implications: the study was somewhat limited, as the studied Grimms' fairytales in English and Lithuanian were translations. This created certain minor holes in the data, i.e., certain *Nomina Loci* that were present in German were not translated as nouns or simply had no equivalent in the English and/or Lithuanian translations, and vice versa, in any combination.

Findings: the number of Nomina Loci found across different languages included 124 cases in German, 132 cases in English and 117 cases in Lithuanian. Compounding showed to be the predominant word-formation process for *Nomina Loci* in the German and English versions of the fairytales. Lithuanian, conversely, had a great number of derivations at the expense of compounds, which comprised a mere 4% of all found *Nomina Loci*.

Practical implications: This study will categorize the collected *Nomina Loci* by word-formation process in each language, design a possible framework helping to systematize questionable and peripheral cases of *Nomina Loci*, and draw parallels between the Germanic and the Baltic languages.

Originality/Value This domain is still insufficiently explored, and the findings of this study may merely serve as a starting point for more in-depth research.

Keywords: *place names (Nomina Loci), word-formation, nominal compounds, derivatives, equivalents.*

Research type (choose one): Research paper.

▶ General Characteristics of Language Interference in Social Media in the Context of Russia's Aggression against Ukraine

Daria Koloda,
Skovoroda Kharkiv State Pedagogical University, Ukraine, koloda.daria@gmail.com

The **purpose** of this study is to analyze the main principles of Ukrainian and Russian language interference in modern social media in the context of Russia's aggression against Ukraine through different linguistics levels – phonetics, graphics, lexicology, and text.

Design/methodology/approach: This study adopts a **deductive approach** to the research and raises a hypothesis that the use of Russian-Ukrainian language interference may be found in different linguistic levels – from phonetic and graphic to text – and has a negative connotation in the text. It employs a **mixed methods research design**, whereby quantitative methods are used to represent different language units and test the hypothesis, while qualitative ones serve to detail the context of comments in different social media platforms.

Findings: The research findings suggest that language interference is explored in different levels of language usage. However, the brightest and most widely used units are found in the phonetic, graphic, and lexical spheres. Furthermore, the data shows that Russian language usage in Ukrainian social media texts has become the enemy's feature and is used with a negative connotation – it shows extreme dislike or loathing.

Research limitations: The data for the current research were collected from texts published/ available on the three most popular social media platforms in Ukraine, i.e., Telegram, Viber, and YouTube. To deepen our understanding of the investigated phenomena, there is a need for further investigations targeted on the language of other types of social media.

Research implications: At present, the main event and the key focus of any discussion in Ukraine is the war waged by Russia against the country. This omnipresent phenomenon influences all spheres of life, and as a result it also affects language usage. Russian-Ukrainian language interference has indicated, on the one hand, close linguistic connections within the common Slavic branch, but, on the other, a big gap in the mental sphere and principal differences between the languages that must be examined accurately. Moreover, as the war is in process, the connotations of language interference are increasing. As a result, research data and statistics of subsequent explorations may be different in the future. This means that synchronic and diachronic analyses are likely to be different.

Practical implications: The study results can be used while teaching different university-based philology subjects such as General linguistics and Comparative Linguistics, Stylistics, or while working in other spheres such as: Comparative Phonology, Comparative Lexicology, Theory of Translation, etc.

Originality/Value: The work is up to date because it is based on authentic language materials collected from most prevalent types of social media in Ukraine in 2020. Social media is one of the most popular means of communication in the modern world; therefore, the language used in it might be considered the best reflection of reality.

Keywords: *language interference, social media, war discourse, negative connotation.*

Research type: Research paper.

► Multilingual Societies from Twitter Data: Empirical Analysis and Theoretical Modelling

David Sanchez,
IFISC (UIB-CSIC), Spain, david.sanchez@uib.es
Thomas Louf,
IFISC (UIB-CSIC), Spain, thomaslouf@ifisc.uib-csic.es
Jose J. Ramasco,
IFISC (UIB-CSIC), Spain, jramasco@ifisc.uib-csic.es

Purpose: Cultural diversity encoded within the languages of the world is at risk, as many languages have become endangered in the last decades in the context of growing globalization. To preserve this diversity, it is first necessary to understand what drives language extinction, and which mechanisms might enable coexistence.

Design/methodology/approach: Here, we discuss the processes underlying language shift through a conjunction of theoretical and empirical perspectives. We quantify the linguistic diversity with the Earth-mover's distance (EMD), either at full county or region scale. This metric allows us to measure the discrepancy between two distributions embedded in a two-dimensional space and is shown to be a proper distance. To understand how different linguistic states can emerge and, especially, become stable, we propose a theoretical model in which coexistence of languages may be reached when learning the other language is facilitated, and when bilinguals favor the use of the endangered language. Then, we carry out simulations in a metapopulation framework.

Findings: A large-scale study of spatial patterns of languages in multilingual societies using Twitter and census data yields wide diversity. This ranges from an almost complete mixing of language speakers, including multilinguals, to segregation, with a neat separation of the linguistic domains and with multilinguals mainly at their boundaries. We calculate the EMD for both monolingual and multilingual groups. In the former, we find a large segregation for Switzerland and Belgium. In the latter, the segregation is maximal for Java and Estonia. The theoretical model is first analyzed in a single population in mean-field, which uncovers interesting stable states of extinction and coexistence, including with bilinguals alone sustaining a minority language. This stability is discussed with flow diagrams that take into account the language prestige and the bilingual preference at a fixed mortality rate. The metapopulation model highlights the importance of spatial interactions arising from population mobility to explain the stability of a mixed state or the presence of a boundary between two linguistic regions.

Research limitations/implications: Interestingly, the achieved state depends on the ratio between mortality rate and learning rate. When the ratio is low (high), the preferred state is spatial mixing (extinction/dominance or spatial separation). We also

investigate the dynamics of our model. We find that the evolution of the system once it undergoes a transition is highly history-dependent. It is easy to change the status quo, but going back to a previous state may not be simple or even possible.

Practical implications: We have shown that, quite counter-intuitively, increasing the ease to learn the other language may break the existing boundary and lead to extinction, and not to the desired coexistence with mixing of the languages. This calls for caution when designing policies since the final state is strongly history-dependent.

Originality/Value: Overall, our findings shed light on the role of heterogeneous speech communities in multilingual societies, and they may help shape the objectives and nature of language planning in many countries where accelerated changes are threatening cultural diversity.

Keywords: computational sociolinguistics, language dynamics, bilingualism, social media.

Research type: Research paper; see Louf, T., Sanchez D., & Ramasco, J.J. (2021). Capturing the diversity of multilingual societies. *Physical Review Research*, *3*, 043146.

▶ The Use of Free Web-Based Mind Mapping Tools in the Studies of ESP: A Review of Research

Eglė Selevičienė,
Mykolas Romeris University, Lithuania, eseleviciene@mruni.eu

Purpose: This study aims at presenting a research review of the effective use of free web-based mind-mapping tools in the studies of English for Specific Purposes (ESP) in higher education. Mind maps, whether paper-and-pencil or web-based, are one of the best-known logical organization tools. In the Typology of Free Web-based Learning Technologies (Bower & Torrington, 2020), mind mapping tools fall under the cluster of *Image-based* tools and are described as educational tools that support the development of images to represent interrelated concepts in the form of a visual knowledge network that can be shared via URL. They can be used to represent conceptual and even meta-cognitive understanding. The category of mind mapping in the Typology includes such web-based freeware as *Bubbl.us*, *Mindomo*, *WiseMapping*, *MindMup*, *Popplet*, *Mind42*, *Mindmeister*, *Slatebox*, *Coggle* or *DebateGraph*.

Design/methodology/approach: This study is a systematic literature review which summarizes evidence of the effective use of aforementioned free web-based mind mapping tools in the context of teaching and learning ESP in higher education. It incorporates empirical research, published as articles in pre-eminent international scholarly journals, book chapters, and conference presentations in the realm of ESP in higher education within the period of 2017–2022.

Findings: The results of this current literature review imply that web-based mind mapping techniques are considered by the majority of researchers as a power tool for improving ESP students' reading comprehension. They help learners to visualize what they are reading, to extract the most important details of ESP texts, to detect connections between separate pieces of information, and to identify sequences or disclose causes and effects. The use of Web-based mind mapping software seems to be effective for developing ESP vocabulary, as it assists in building up a structure of knowledge for ESP terms by associating them with other related words or concepts. This study also concludes that the use of free web-based mind mapping tools for teaching and learning ESP may be guided by assumptions underpinning *cognitive* or *constructivist* philosophical approaches, however the majority of the research lacks a clear theoretical background. ESP teachers and students mainly have positive attitudes towards the use of Web-based mind-mapping tools in studies of ESP in higher education. On the other hand, some of the tools are described as complicated to use.

Research limitations/implications: As the focus of the study was on the use of free web-based mind mapping tools for teaching and learning ESP in higher education,

it is not inconceivable that different evaluations would have arisen if the focus had also been on the use of commercial products.

Practical implications: This study provides useful insights for both practitioners and researchers interested in the use of web-based mind mapping tools for teaching and learning ESP in higher education.

Originality/Value: The review contributes to the body of knowledge within the field by presenting evidence on the effective use of web-based mind mapping freeware after carefully analyzing available empirical research published in the last five years.

Keywords: *Web-based educational technologies, mind-mapping technologies, English for Specific Purposes, higher education.*

Research type: Literature review.

▶ Motivation, Correction, Foreign Language Acquisition Through the Use of Psycholinguistic Tests

Eugénie Bonner-Bestchastnova,
Mykolas Romeris University, Lithuania, University of Caen Normandy, France, eugenie@mruni.eu
Antoinette Bestachstnova,
Univresity Paris Ouest Nanterre la Défense, France

Purpose: Our task was to develop psycholinguistic tests in order to identify the motivation to study a foreign language, collect all the material for the further correction of the educational process, and create a positive attitude for foreign language learners. We then introduced these tests into our teaching program.

Methodology: The first test is proposed at the beginning of the session and the second at the end. The introductory test consists of 17 questions. It is essentially targeted at the motivation, interest, and objectives of the learner. The final test is composed of 20 questions and allows us to analyze the progress in the learning process, difficulties, possible loss of motivation, and progress. Linguistic and psychological factors were identified through these tests.

Findings: The tests allowed us to draw the conclusion that answering test questions allows learners to realize the importance of learning a foreign language for them, and allows us to understand and increase their motivation. The test made it possible to identify psycholinguistic difficulties in the acquisition of a foreign language among learners in order to correct them later. Thanks to these tests, the level of proficiency in a foreign language has generally improved among learners. General proficiency in a foreign language is progressing after using these tests. We have also found that, often, the difficulties in learning a foreign language are not in the absence of skills, but in the absence of purpose, motivation, and learning system.

Practical implications: A foreign language is becoming one of the main factors in the social, economic, and general cultural progress of society. A foreign language plays a huge role in shaping personality and improving education. Knowledge of a foreign language has become a need. Therefore, the concept of the motive of learning a foreign language comes to the fore. The specifics of the study of foreign language requires the presence of certain knowledge and communicative abilities. Our tests helped to consider motivation more broadly as the main driving force in learning a foreign language.

Originality/Value: Our research allowed us to reveal that the comprehension of the test questions helps the learner to enhance linguistics skills and motivation. These tests were developed and introduced in the learning process and correction for the first

time.

Keywords: *acquisition, learning, foreign language, psycholinguistics, motivation.*

Research type: Research paper

► Tracing Semantic Change with Multilingual LLOD and Diachronic Word Embeddings

Florentina Armaselu,
University of Luxembourg, Luxembourg, florentina.armaselu@uni.lu
Elena-Simona Apostol,
Uppsala University, Sweden, elena-simona.apostol@it.uu.se
Christian Chiarcos,
University of Cologne / Goethe University Frankfurt, Germany, chiarcos@cs.uni-frankfurt.de
Anas Fahad Khan,
Istituto di Linguistica Computazionale "A. Zampolli", Italy, fahad.khan@ilc.cnr.it
Chaya Liebeskind,
Jerusalem College of Technology, Israel, liebchaya@gmail.com
Barbara McGillivray,
King's College London, United Kingdom, barbara.mcgillivray@kcl.ac.uk
Ciprian-Octavian Truică,
Uppsala University, Sweden, ciprian-octavian.truica@it.uu.se
Giedrė Valūnaitė-Oleškevičienė,
Mykolas Romeris University, Lithuania, gvalunaite@mruni.eu

Purpose: The project will combine word embedding techniques and linguistic linked open data (LLOD) with theoretical aspects from lexical semantics, the history of concepts, and knowledge organization to trace the evolution of concepts in a collection of multilingual diachronic corpora of seven extinct and extant languages (Latin, Ancient Greek, Hebrew, French, Old Lithuanian, Romanian, German). The outcome will consist of a sample of diachronic ontologies to be published on the LLOD cloud. It will also comprise reflections on the potential interconnections across different languages that can be built through these knowledge structures.

Design/methodology/approach: The methodology will include the following steps: (1) Train diachronic static and contextual word embeddings (i.e., Word2Vec, FastText, ELMo, BERT) by time slice and sliding window on each dataset. Drawing on state-of-the-art methods on unsupervised lexical semantic change detection, we will model semantic change in the vector space by taking into account theoretical considerations such as semasiological and onomasiological mechanisms, and core – margin and intension – extension – label conceptualizations. (2) Create cross-lingual connections between concepts for overlapping and sequential intervals on the general timeline. (3) Devise transformation pipelines to move from the vector space to a representation of concept evolution through chains of label, core/intension, margin/extension at different time points using Semantic Web formalisms and vocabularies. For instance, we will use the OntoLex-Lemon model and its extension, the Frequency Attestations and Corpus

information (FrAC) module. We will employ the OWL-Time vocabulary to explicitly represent time in our ontologies, and possibly utilize the semantic data derived from our approach to enrich multilingual WordNets with diachronic information.

Findings: The project will contribute new forms of conceiving, detecting, and representing semantic change by combining theoretical, computational, corpus- and knowledge-based approaches for both high and low resource languages.

Research limitations/implications: The main limitations consist of potential data sparsity and the lack of existing LLOD mechanisms that fully support the representation of semantic change. The heterogeneous nature of the available datasets poses an additional challenge in processing the data through a common yet flexible methodology.

Practical implications: Given the complexity of the tasks, we will need to use multiple approaches and resources (corpus-, dictionary-, linked data-based), and to propose extensions to existing LLOD formalisms.

Originality/Value: The novelty of the proposal will consist in bridging research perspectives that have not yet been considered together (e.g., semasiological/onomasiological, core/margin, vector semantics, linked data) to trace concept evolution in a multilingual diachronic setting. It will also reside in broadening current LLOD models to accommodate spatio-temporal dimensions and the dynamics of change, and in reconciling the variety of the available datasets to build cross-lingual space-time connections.

Keywords: *diachronic word embedding, LLOD, semantic change, history of concepts, knowledge organization.*

Research type: Research paper.

► Practices of Online Language Teaching and Learning: A Survey in Lithuania

Giedrė Valūnaitė-Oleškevičienė,
Gražina Čiuladienė,
Lora Tamošiūnienė,
Liudmila Mockienė
Mykolas Romeris University, Ateities g. 20, LT-08303, Vilnius, Lietuva,
gvalunaite@mruni.eu
grazina.ciuladiene@mruni.eu
lora@mruni.eu
liudmila@mruni.eu

Purpose: The findings of the Commission's public consultation observed that the COVID-19 crisis and the subsequent shift to online content have caused an unprecedented move towards digital teaching and learning practices. However, the situation has also revealed a gap in digital technologies fully supporting creative and collaborative student-centered learning and facilitating inclusive education. It should be stressed that language teaching and learning mainly take place through interaction and usage, so the question arises as to how efficient teaching/learning interactions can be maintained at distance. It was observed by White (2006) that the use of digital technology provides a shift away from the classroom and makes the learner experience central as well as digital transition blends in a diverse range of formal and informal practices.

In this context, this research aims to identify good practices of taking up digital technologies and open pedagogies in language teaching and learning. This is related to the goal of the DIAL4U project to co-develop an innovative approach and digital tools in order to: facilitate mediation in all dimensions of the language learning process (taking into account both formal and informal situations); build the capacity and autonomy of all language learners; and develop the digital/blended pedagogy competences of language educators.

Approach: Within the framework of the DIAL4U project, questionnaires for teachers and students were designed, reviewed, and sent to research participants after finalization. This research particularly targeted two of the most important groups of research participants: teachers working as language instructors and students learning languages. The target groups were diverse in terms of age, language training, position, CEFR level, studied language, and, most importantly, digital literacy level.

Findings: The two sets of questionnaires allowed for good practices in foreign language teaching and learning methods to be collected; they also helped to indicate

trends regarding the digitalization of teaching processes and allowed the evaluation of the manner in which language teachers employ strategies of metacognitive and politeness. The results revealed that good practices encourage creative thinking skills and student/teacher engagement. In addition, the analysis of the survey results allowed the identification of the key elements and issues regarding online politeness and metacognitive strategies.

Value: The impact of this research could be of interest to various organizations (i.e., academic institutions, language centers, etc.). It provides a better understanding of the process of online academic learning/teaching of languages with reference to the apps used, their integration into language classes, associated advantages and possible challenges, as well as the strategies (pedagogical/metacognitive/politeness) that could be used for beneficial training in online language classes. IO1 deliverables are well thought out, innovative, and they encourage creative thinking skills and student/teacher engagement.

Keywords: *language teaching and learning, digital technology, questionnaires, metacognitive strategies, open pedagogies.*

Research type: Research paper

References

White, C. (2006). Distance learning of foreign languages. *Language Teaching*, *39*(4), 247–264.

Acknowledgements

DIAL4U project funded under the program Erasmus+, KA2: Strategic Partnership Reference: 2020-1-FR01-KA226-HE-095526 June 2021–May 2023

▶ DIAL4U: Digital Pedagogy to Develop Autonomy, Mediate and Certify Lifewide and Lifelong Language Learning for (European) Universities

Giedrė Valūnaitė-Oleškevičienė,
Lora Tamošiūnienė,
Gražina Čiuladienė,
Mykolas Romeris University, Ateities g. 20, LT-08303, Vilnius, Lietuva,
gvalunaite@mruni.eu
lora@mruni.eu
grazina.ciuladiene@mruni.eu

Purpose: This presentation will introduce the advances of the work performed by seven European Universities jointly involved in the DIAL4U project, funded under Erasmus+ KA2: Strategic Partnership. The COVID-19 crisis and the unplanned shift to online content have exposed the gap in digital technologies to fully support high-quality and inclusive education facilitating creative and collaborative student-centered learning. The informal language learning validation and recognition process investigated in the DIAL4U project tackles this gap, as the informal learning question is intrinsically linked to the digital transition: many informal language-learning practices are digital, and their successful articulation to formal learning and recognition will be improved with digital tools.

Approach: This project enhances the learning capacity and autonomy of all language learners; develops or upskills the digital/blended pedagogy competences of language educators; and facilitates the recognition and validation of knowledge, skills, and competencies gained through formal, non-formal, and informal language learning.

Practical implications: This project equips teachers and learners (potentially all students from the consortium, as languages refer to transversal competences) with reflexive tools to develop a critical approach and the conceptual tools required for successful mediation of these newly identified types of practice. The DIAL4U project contributes to taking up digital technologies and opening pedagogies in language teaching and learning, participating in the development of high-quality inclusive language education in Europe.

Research implications: This project takes place in the higher education context and focuses on language learners and teachers. On the one hand, the goal is to equip learners (potentially all students from the project consortium, as languages refer to transversal competencies) with reflexive tools to develop a critical approach; on the other hand, language teachers compose a crucial target group as their expertise is required for the mediation process, even when informal resources are considered. The project

aims to equip them with the conceptual tools required for the successful mediation of these newly identified types of practice.

Keywords: *Open and distance learning, recognition, transparency, certification, foreign language teaching and learning*

Research type: Case study

Acknowledgements

DIAL4U project funded under the program Erasmus+, KA2: Strategic Partnership Reference: 2020-1-FR01-KA226-HE-095526 June 2021–May 2023

▶ Processing Multi-Word Discourse Markers in Translation: English to Hebrew and Lithuanian

Giedrė Valūnaitė-Oleškevičienė,
Mykolas Romeris University, Ateities g. 20, LT-08303, Vilnius, Lietuva, gvalunaite@mruni.eu,
Chaya Liebeskind,
Jerusalem College of Technology, 21 Havaad Haleumi str., 9116001, Jerusalem, Israel, liebchaya@gmail.com

Purpose: It has been proved that multi-word expressions are of key importance in language generation and processing. They also could perform a function of discourse organization, and certain multi-word expressions operate as discourse markers. The purpose of the current research is to examine multi-word expressions used as discourse markers in TED talk English transcripts and compare them with their counterparts in their Lithuanian and Hebrew translations, identifying if English multi-word expressions used as discourse markers in social media texts remain multi-word expressions in Lithuanian and Hebrew translation and searching for reasons for the changes of discourse markers in translation. We follow the research question of how English multiword discourse markers are processed in Hebrew and Lithuanian translation.

Approach: To achieve the aim of the research, the set objectives were to create a parallel research corpus to identify multiword expressions used as discourse markers and to analyze their translations in Lithuanian and Hebrew to determine if they are also multiword expressions or one-word translations, or if they acquire any other linguistic forms, and look for the possible reasons for the translator choices. In the research, we combine the alignment model of the phrase-based statistical machine translation and manual treatment of the data in order to examine English multi-word discourse markers and their equivalents in Lithuanian and Hebrew translation by researching their changes in translation. After establishing the full list of multi-word discourse markers in our generated parallel corpus, we research how multi-word discourse markers are treated in translation. We apply the method of Corpus research and phrase-based statistical machine translation/research corpus available at LINDAT/CLARIN-LT repository http://hdl.handle.net/20.500.11821/34

Findings: Our research proves that the examined multi-word discourse markers have different translation tendencies due to the different grammars of the researched languages. There is a trend to remain multi-word in Hebrew translation, but due to the translation choices relying on inflections, they are one-word discourse markers in Lithuanian. There is also possible context-based influence guiding the translator to choose a particle or other lexical item integration in Lithuanian or Hebrew translated discourse markers to express the rhetorical domain, but the observed phenomenon of "over-specification" requires further research. Beyond the empirical research, an extensive parallel

data resource has been created to be openly used.

Value: The valuable outcome of the study was extending the available resources and providing linguistic processing for several languages by creating a multilingual parallel corpus (including English, Lithuanian, and Hebrew) based on social media texts; the created corpus is shared and interlinked via CLARIN open language resources.

Keywords: *Translation, corpus, multi-word expression, discourse relation, discourse marker.*

Research type: Research paper

► Selection, Implementation and Testing of Language Sample Analysis Measures for the Web-Based Application MultiDis

Gordana Hržica,
University of Zagreb – Faculty of Education and Rehabilitation Sciences, Croatia, gordana.hrzica@erf.unizg.hr
Sara Košutar,
University of Zagreb – Faculty of Education and Rehabilitation Sciences, Croatia, sara.kosutar@erf.unizg.hr
Dario Karl,
University of Zagreb, University of Zagreb – Faculty of Humanities and Social Sciences, Croatia, dkarl@ffzg.hr
Matea Kramarić,
University of Zagreb – Faculty of Education and Rehabilitation Sciences, Croatia, matea.kramaric@erf.unizg.hr

Purpose: The MultiDis application is a new, web-based application designed for the analysis of spoken and written language samples, which provides information about the language abilities of children and adults, thus facilitating language assessment. The aim of this paper is to present the selection, implementation, and testing of language measures in the MultiDis application. We will present the application, the process of selecting the measures we implemented, the language resources needed to calculate them, and the results of testing. MultiDis is currently being developed for Croatian, but it could be scaled up for multilingual analysis.

Design/methodology/approach: Language samples can be analyzed according to several dimensions, such as productivity, lexical diversity, and syntactic complexity. A set of (semi-) automatic measures has been selected to assess language abilities (e.g., number of lemmas, mean-average type-token ratio, mean length of communication unit). The next step was the integration of an open-source Python library for lemmatization, part-of-speech tagging, and syntactic parsing (Stanza; Qi et al., 2020). To test whether these tasks and the subsequent calculation of language measures can be successfully performed on spoken language samples, we uploaded 150 short narrative samples produced by children as a result of a storytelling task.

Findings: Lemmatization and part-of-speech tagging are fairly accurate (>85% of cases), as they do not interfere with the calculation of the currently implemented measures of productivity and lexical diversity. The process of syntactic parsing has been an obstacle that is currently being resolved.

Research limitations/implications: The MultiDis web application is still under development, although the current version fulfils its main purpose – it allows for (semi-)automatic spoken language analysis.

Practical implications: There is an increasing awareness of the importance of language sample analysis as a complementary method in language assessment. The time needed for transcription and the linguistic knowledge required for manual analysis are considered to be the main obstacles to its implementation (Pezold et al., 2020). Therefore, the development of a tool for automatic calculation of language measures such as the MultiDis application could make naturalistic language assessment more feasible.

Originality/Value: The value of this study lies in proposing a new application for lemmatization and part-of-speech tagging that allows for more reliable calculation of measures of productivity, lexical diversity, and syntactic complexity. Selecting appropriate measures for language assessment is a challenging task because there are many available. Implementing language technologies developed for large bodies of written texts to spoken language is also challenging. Success in some parts of automated tagging (lemmatization and part-of-speech tagging) allows for the reliable calculation of measures of productivity and lexical diversity. Future work on syntactic parsing will lead to the successful implementation of measures of syntactic complexity.

Keywords: l*anguage sample analysis, microstructural measures, lemmatization, part-of-speech tagging, syntactic parsing.*

Research type: Research paper.

▶ Linking Lexicographic Resources to Language Proficiency-Level Applications

Kris Heylen,
Dutch Language Institute, Netherlands, kris.heylen@ivdnt.nl
Ilan Kernerman,
K Dictionaries – Lexicala, Israel, ilan@kdictionaries.com
Carole Tiberius,
Dutch Language Institute, Netherlands, carole.tiberius@ivdnt.nl

Purpose: We aim to enhance the development of vocabulary teaching and training materials by converging difficulty-graded word lists with lexicographic data. Grading word difficulty is prevalent in both native and additional language learning, in production and reception tasks, and for text readability analysis and vocabulary testing. Our objectives are to upgrade the usability of such resources for creators of vocabulary learning materials – by enriching them with semantic information such as definitions, examples of usage, and multiword expressions (and possibly more) from dictionaries – cross-lingualize the different language sets, and upload the by-products to the Linguistic Linked Open Data cloud.

Design/methodology/approach: The Common European Framework of Reference for Languages (CEFR) promotes the development of empirically based datasets for 30 languages of Europe according to graded proficiency levels. Each of the six CEFR levels – from beginner, A1, through A2, B1, B2, C1, to advanced C2 – refers to specific situations and conditions, and includes corresponding vocabulary in every language. We will associate pedagogical and multilingual lexicographic data with the words in the CEFR lists with the aid of smart matching and linking techniques and monolingual and multilingual sense alignment methods.

Findings: Since their introduction in the early 2000s, CEFR graded lists have been developed for approximately 15 languages, with most lists having gradings only on the lemma level. Only English has comprehensive lists (with Cambridge and Oxford advanced learner's dictionaries) with proficiency gradings on the sense level, but the resulting data is not available for other open applications. The list for Dutch has been linked in a probabilistic way to Dutch WordNet synsets but, as a consequence, it also is quite noisy.

Research limitations/implications: The primary drawback of most existing CEFR lists is that they do not disambiguate polysemous words. Secondly, when linking them to their corresponding dictionary components, it is necessary to assure that the words used within definitions, examples, and expressions are not situated on higher CEFR levels, and likewise for their equivalents in the other languages.

Practical implications: The challenges are to (a) evaluate the words in existing CEFR lists and link their appropriate senses in dictionaries, (b) make sure the additional lexicographic components contain no words from higher levels, (c) create CEFR lists for languages that do not have them yet and link them to lexicographic data, and (d) find an appropriate project framework and a range of competent partners with language learning expertise, lexicographic resources, and link data know-how.

Originality/Value: Previous CEFR lists projects, including Kelly (2009) and CEFR-Lex (2017), as well as a few carried out individually (e.g. Estonian), have had differing results. Our project will use their relevant achievements while gradually expanding the scope to all CEFR languages and attending to all the issues described above.

Keywords: *language learning, proficiency levels, CEFR lists, lexicographic resources, LLOD cloud.*

Research type: Conceptual paper.

▶ Lithuanian Translation of the Discourse Marker *And* in Social Media Texts

Livija Puodžiūnaitė,
Giedrė Valūnaitė-Oleškevičienė,
Mykolas Romeris University, Ateities g. 20, LT-08303, Vilnius, Lietuva,
lipuodziunaite@stud.mruni.eu
gvalunaite@mruni.eu

Purpose: Discourse markers spark the attention of many linguists and researchers, so linguistics applies computer science in order to study and understand discourse marker use. The problem with discourse markers is their possible ambiguity and polysemy, as the same discourse marker can perform distinctive functions and express different types of discourse interfaces (Sweetser, 2002; Aijmer, 2002; Zufferey & Degand, 2017). Therefore, this research exhibits that and is not only used solely for connecting idea units, but also expresses conditional, causal, temporal, and other types of discourse relations. The aim of the research is examine the translation of the discourse marker and from English to Lithuanian language in TED-ELH parallel corpus. To achieve the research aim, the following objectives are set: identify sentences with and in TED-ELH parallel corpus; manually annotate sentences with and in English and ir in Lithuanian, and indicate instances where they are used as a stand-alone marker or in a group of multiple discourse markers; and analyze how discourse marker is used in the source language and target language.

Approach: The TED-ELH Parallel Corpus was used for conducting the research of the DM (discourse marker) and. It contains parallel aligned scripts of spoken language data from TED Talks in English, Lithuanian, and Hebrew. This research focuses on the English and Lithuanian languages. The methods of research used for this work are manual annotation and qualitative analysis. Manual annotation is performed both in English and Lithuanian languages to indicate whether and in English and its translation in Lithuanian would be considered as discourse markers, and then different translations of and in the Lithuanian language are analyzed.

Findings: The results reveal that in most cases and did not function as a discourse marker in the English language (79%). Concerning multiple discourse markers, there are very little of them in the Lithuanian language (1%) as opposed to English (10%). In relation to the usage of a discourse marker in source (English) and target (Lithuanian) languages, there are instances where a discourse marker is absent in the Lithuanian language. This absence is explained by one of the two translation strategies that are applied by the translator, either omission or grammatical transformation. In addition, and has 11 different translations, and is most frequently translated as ir (53.3

%), o (35.2%), and bet (3.4 %) in Lithuanian. This also raises the issue of ambiguity and polysemy of translating discourse markers as the translation of and varies significantly.

Value: This research is relevant for analyzing discourse markers and their translations in order to gain knowledge in developing machine translation. Moreover, this research is beneficial for conducting similar future research with different discourse markers and language combinations.

Keywords: *discourse markers, translation, manual annotation, parallel corpus.*

Research type: Research paper

References

Aijmer, K. (2002). English Discourse Particles: Evidence from a Corpus. John Benjamins Publishing Company.

Sweetser, E. (2002). From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure. Beijing; Cambridge, England: Peking University Press; Cambridge University Press, 2002.

Zufferey, S., & Degand, L. (2017). Annotating the meaning of discourse connectives in multilingual corpora. Corpus Linguistics and Linguistic Theory, 13(2), 399–422.

▶ Issues in Building the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin

Marco Passarotti,
Francesco Mambrini,
Università Cattolica del Sacro Cuore, Milan, Italy, marco.passarotti@unicatt.it

Purpose: This abstract presents the architecture and the current state of the LiLa Knowledge Base (https://lila-erc.eu), i.e., a collection of multifarious linguistic resources for Latin described with the same vocabulary of knowledge description, by using common data categories and ontologies developed by the Linguistic Linked Open Data (LLOD) community according to the principles of the Linked Data paradigm.

Design/methodology/approach: LiLa uses the lemma as the most productive interface between lexical resources, annotated corpora, and NLP tools. The core of the LiLa Knowledge Base consists of a large collection of Latin lemmas (called Lemma Bank): interoperability is achieved by linking, via (L)LOD data categories and ontologies, all those entries in lexical resources and tokens in corpora that point to the same lemma.

Findings: The textual resources currently interlinked through LiLa include: two dependency treebanks (*Index Thomisticus* Treebank, UDante), Fibonacci's *Liber Abbaci*, and a large corpus of Classical Latin texts (LASLA corpus). Lexical resources include: a manually checked subset of the Latin WordNet, a sentiment lexicon (Latin Affectus), a derivational lexicon (Word Formation Latin), a Latin-English Dictionary (Lewis & Short), a list of Greek loans in Latin, and an etymological dictionary.

Research limitations/implications: Limitations of representation of linguistic (meta)data in the currently available LLOD models and ontologies; issues of harmonization of different lemmatization strategies and PoS tagging; problems of automatic linking of lexical and textual resources; and problems related to extracting information from linguistic resources published as LLOD.

Practical implications: The LiLa Knowledge Base can be queried through a SPARQL endpoint at https://lila-erc.eu/sparql/, where a few pre-compiled queries are available. The Lemma Bank of LiLa can be accessed at https://lila-erc.eu/query/. Lemmas can be searched by string of characters, Part of Speech, affix, lexical base, inflectional category, and gender (for nouns). Results are provided both in data sheet fashion and in a network-like graphical visualization. The entries in lexical resources and the tokens in corpora linked to each lemma in LiLa are reported as well. The Turtle files of the resources interlinked in LiLa are available at https://github.com/CIRCSE.

Originality/Value: The LiLa Knowledge Base makes it possible to perform que-

ries on interoperable, distributed linguistic resources for Latin published on the web: this represents a terrific advancement in the way such resources can be used, particularly by Classicists, who need a steady confrontation with the empirical evidence provided by textual and lexical resources. Building the LiLa Knowledge Base represents a large-scale use-case where to apply and test the currently available vocabularies developed by the LLOD community, as well as developing new ones.

Keywords: *linguistic resources, Linguistic Linked Open Data, Latin*

Research type: Research paper.

▶ Verbo-Visual Pun in Memetic Warfare Against Russia's Aggression in Ukraine

Maryna Bielova,
G.S. Skovoroda Kharkiv National Pedagogical University, Ukraine,
marynavasylieva2005@gmail.com

Purpose: In a time when war-torn Ukraine faces a grave crisis amid Russia's invasion attempt, social media chooses humor over aggression to fight back. Humor has long been known as a kind of coping mechanism in extreme crises wherein the world of jokes is quite similar with the world of memes, which instantly reacts to resonating events. Puns and memes, armed with words and images, become weapons of mass disruption, influencing the hearts and minds of a target audience. This invites the examination of how memes detailing the Russian-Ukrainian military crisis employ verbo-visual puns as a discursive strategy of information warfare to counter Putin's propaganda.

Design/methodology/approach: This study illustrates a specific mechanism of processing verbo-visual puns representing the Russian-Ukrainian war in English-language memes, and explores how these puns are strategically used in crisis communication. It proposes an interdisciplinary template of the analysis of puns and draws on multimodal social semiotics, cognitive linguistics, a critical study of multimodality, and crisis communication analysis.

Findings: This paper reveals cognitive and semiotic triggers of puns as well as their ability to serve as a powerful tool to generate strong public reaction to Russia's military aggression in Ukraine reaching the intended goals of crisis communication.

Research limitations/implications: The description of all memes depicting the Russian-Ukrainian war exceeds the scope of this article, so it focuses exclusively on the verbo-visual response to the crisis and is limited to six cases.

Practical implications: The research is especially timely in the face of the advanced spread of Russian propaganda against world democracies, including Ukraine. Its major practical contribution is that it provides much needed insight into how English-language memes, as the new frontier of information warfare against Russia's aggression in Ukraine, can pursue a competitive advantage over the enemy.

Originality/Value: This paper has not been previously published, in whole or in part, in any other journal or scientific publishing company.

Keywords: *pun, multimodality, Internet-meme, strategic communication.*

Research type: Research paper.

▶ Theoretical Basics of Stylized Colloquial Speech Functioning in Works of Fiction

Olena Kholodniak,
Mykolas Romeris University, Lithuania, olenakholodniak@mruni.eu

Purpose: The purpose of this research lies in an extended analysis of stylized colloquial speech functioning in the language of fiction, identifying its differences from colloquial speech and the clarification of their functions.

Design/methodology/approach: The methodological basis of this work is the traditional methods of linguistic research. The method of linguistic analysis and the descriptive method make it possible to analyze and characterize the distinctive features of colloquial speech and stylized colloquial speech. The method of component analysis makes it possible to study the content component of the object of study. Using the method of logical comparison, it is possible to draw and substantiate the theoretical conclusions of this work.

Findings: Live and artistically reproduced colloquial speech are not the same thing, but they are not completely asymmetric and have a certain number of cumulative features. Summing up the results of research by various scientists, it was possible to establish the linguistic means and stylistic devices with the help of which colloquial speech is stylized, and what functions it performs in works of fiction. The main theoretical questions related to the processes which cause the transformation of colloquial speech into stylized colloquial speech, differences in syntactic constructions, semantic features of stylized elements, and the sphere of use were systematized.

Research limitations/implications: When studying this issue, domestic and foreign studies of recent decades were studied and analyzed, which related to the style of colloquial speech, its structural and semantic features, and the concept of stylized colloquial speech.

Practical implications: The practical value of the study is due to the possibility of using materials and conclusions in university practice, namely in courses on stylistics, analytical reading, and artistic text analysis. The results of the study can be applied when writing teaching aids and lecture courses, and in the study of individual authors' stylistic features. The theoretical statements can be used in further study of this topic.

Originality/Value: The originality of the work is due to the fact that for the first time in domestic philology a comprehensive comparison of colloquial speech and stylized colloquial speech in the language of fiction was carried out.

Keywords: *stylized colloquial speech, work of fiction, elements of non-artistic*

speech, semantic features of stylized elements.

Research type: General review.

English and Lithuanian Terms in the Parallel Migration Corpus

Olga Usinskiene,
Sigita Rackevičienė,
Mykolas Romeris University, Lithuania
olga.usinskiene@mruni.eu
sigita.rackeviciene@mruni.eu

Purpose: The aim of this paper is to present the results of conceptual/thematic and linguistic (synonymous variability) analysis of migration terminology extracted from the parallel corpus composed of the EU legal acts and other documents, as well as legal acts of international organizations – the United Nations and the Council of Europe. This presentation will provide the principles of compilation of the parallel corpus, terminology extraction methodology, and results of their analysis.

Design/methodology/approach: The English terminology is extracted from the corpus using tools of the Sketch Engine software tool "Keywords". The tool extracts keywords (single-word and multi-word terms) which are typical of the focus corpus; it does that by comparing the focus corpus with the reference corpus selected by the software (in our case – English Web 2020 (enTenTen 2020)). The extracted lists of terms have been exported as MS Excel files and, subsequently, the migration terms are being annotated manually in accordance with the thematic groups/conceptual categories the terms belong to, and synonymous variants are being established and clustered. Lithuanian equivalents of the English terms are being established using Sketch Engine tool "Parallel Concordance". Finally, quantitative analysis of synonymous variability in English and Lithuanian is performed.

Findings: One thousand English multi-word terms have been extracted by the Sketch Engine. The terms are being tagged according to the migration domain that they belong to. In addition, they are ascribed to thematic groups – 3 broad thematic categories based on the causal chain, comprising: cause (reasons of migration), effects (resulting events and activities of participants in migration process), and consequences (impact of migration processes on social, economic and political life), which are further subdivided into more specific thematic groups. Numerous terms have synonymous variants both in English and Lithuanian (e.g. *irregular migrant, illegal migrant – neteisėtas migrantas, nelegalus migrantas*), and the results of quantitative synonymous variability analysis will be provided during the presentation.

Research limitations/ implications: The scope of the dataset is limited to legal acts. Therefore, they reflect the use of migration terms only in legislative discourse.

Practical implications: Terminology of migration has become immensely significant not only for migrating people, but also for the officials, legislative authorities, border officers, non-governmental organizations, and volunteers constantly using migration-related terms in dealing with migration issues.

Originality/ value: To our knowledge, this is the first research of such scope that included Lithuanian migration terminology. Its results can be used for management of terminology in this domain, the compilation of a migration termbase, and linking terminological data with other resources using LLOD technologies.

Keywords: *migration domain, parallel corpus, terminology analysis, thematic groups.*

Research type (choose one): Research paper.

▶ The LexMeta Metadata Model for Lexical Resources: Theoretical and Implementation Issues

Penny Labropoulou,
Institute for Language and Speech Processing, Athena Research Center, Athens,Greece, penny@athenarc.gr
David Lindemann,
UPV/EHU University of the Basque Country, Spain, david.lindemann@ehu.eus
Christiane Klaes,
TU Braunschweig/Univ. Library, Germany, c.klaes@tu-braunschweig.de
Katerina Gkirtzou,
Institute for Language and Speech Processing, Athena Research Center, Athens,Greece, katerina.gkirtzou@athenarc.gr

Abstract: The paper presents LexMeta, a metadata model catering for descriptions of human-readable and computational lexical resources included in library catalogues and repositories of language resources. We present the main concepts of the model, its implementation, and discuss current findings and future plans.

Purpose: The paper presents LexMeta, a metadata model catering for descriptions of human-readable and computational lexical resources included in library catalogues and repositories of language resources. It, therefore, takes into account requirements for bibliographical metadata (for citation purposes) as well as for features of contents and form, addressing user needs of lexicographers and, in a broader context, of scholars of social sciences and humanities (SSH) that wish to deploy these resources in their research workflows (for findability and usability purposes).

Design/methodology/approach: The model builds upon broadly used existing models; these are mainly the FRBR (Functional Requirements for Bibliographic Resources) model from librarians' perspective, the META-SHARE ontology for language resources and technologies, and the LexVoc Vocabulary of Lexicographic Terms, a structured controlled list of terms related to lexicographical and metalexicographical concepts. New classes and properties are introduced in the LexMeta namespace when these are not covered by existing models. The model is structured around three main classes: *Lexicographic Work*, equivalent to the abstract notion of a lexicographical creation, *Lexical/Conceptual resource*, representing the realization of a single work (e.g., a certain version or edition of a lexicographic work) and *Distribution*, representing the physical form in which a work is issued (e.g., as a printed book or a digital file). For each class, the appropriate set of properties is defined accompanied with cardinality features, and, for classification properties, the relevant controlled vocabularies. Development of the model is underway in two forms: as an ontology of Wikibase entities,

and as an OWL ontology. The former implementation enables its adoption in the LexBib Wikibase Knowledge Graph of Lexicography and Dictionary Research, a research infrastructure that follows the Wikibase data model, while the latter supports works in the Linked Data community. Classification vocabularies are represented in the form of SKOS.

Findings: The model looks promising for the envisaged purposes.

Research limitations/implications: LexMeta is used in the population of the LexBib catalogue, which helps identify potential gaps and errors. We are also soliciting feedback from the target communities.

Practical implications: The model can be extended and enriched with recommendations and user feedback. We foresee issues in keeping the two representation forms synchronized, and we are looking into a viable solution.

Originality/Value: To the best of our knowledge, this is the only attempt at the dual implementation of a metadata model. The metadata model itself is of particular interest to SSH researchers and lexicographers, and is already applied in a real use case.

Keywords: *lexicography, metadata model, lexical resources, linked data, wikibase.*

Research type: Research paper

▶ Developing a Cybersecurity Termbase

Sigita Rackevičienė,
Mykolas Romeris University, Lithuania, sigita.rackeviciene@mruni.eu
Andrius Utka,
Vytautas Magnus University, Lithuania, andrius.utka@vdu.lt
Liudmila Mockienė,
Mykolas Romeris University, Lithuania, liudmila@mruni.eu
Agnė Bielinskienė,
Vytautas Magnus University, Lithuania, agne.bielinskiene@vdu.lt

Purpose: The aim of the paper is to present the principles of compilation of a bilingual cybersecurity termbase and to discuss possibilities to publish the terminological data included in the termbase as open linked data.

Design/methodology/approach: The development of the bilingual termbase is based on the combination of prescriptive and descriptive terminology management principles. The best practices of both approaches have been used for selection and representation of the terminological data. Selection of terms for inclusion in the termbase has involved human and machine annotation of cybersecurity terminology in specially designed bilingual resources: parallel and comparable English-Lithuanian corpora composed of texts from a variety of discourses (legal, administrative, academic, and media) and in different genres (legal acts, reports, scientific papers, textbooks, media articles, etc.). The termbase will provide frequential data of the selected terms and their typical usage examples in context based on the empirical dataset. This information is important for term standardization, which is essential for smooth specialized communication.

The macrostructure of the termbase has been developed on the basis of the analysis of thematic groups of cybersecurity concepts and their classification. The microstructure of the termbase is based on the concept-oriented approach enabling terminological data and metadata to be organized around the concepts that the data and metadata pertain to.

Results: The termbase is developed in the open-source cloud-based terminological management platform *Terminologue*, designed and administered by the Gaois research group at Dublin City University, and will be freely available online. The termbase data can be exported in the TBX format and converted to RDF to enable its LLOD compatibility for linking with other terminological data.

Practical implications: The termbase is expected to provide necessary information for cybersecurity specialists and translators, as well as to contribute to the development of Lithuanian national terminology, which is used in national legislation,

administrative documentation, educational materials, etc. Potentially, the termbase will contribute to the understanding of cyber threats and will raise cyber awareness in the general public. LLOD technology application will enable interoperability of the terminological data to be improved, increasing its reuse.

Originality/Value: To our knowledge, the Lithuanian-English cybersecurity termbase is the first termbase in which Lithuanian and English data is based on bilingual corpora combining descriptive and prescriptive approaches.

Keywords: *cybersecurity, termbase, descriptive and prescriptive terminology management, LLOD.*

Research type: Research paper.

Acknowledgments

The research is carried out under the project "Bilingual Automatic Terminology Extraction" funded by the Research Council of Lithuania (LMTLT, agreement No. P-MIP-20-282). The project is also included as a use case in COST action "European Network for Web-Centred Linguistic Data Science" (CA18209).

Towards the Integration of Sign Languages Data in the Linguistic Linked Open Data Cloud

Thierry Declerck,
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus D3 2, Germany
declerck@dfki.de

Purpose: In the field of electronic lexicography, there is an increasing interest in offering ways to represent and interlink lexical data originating from different modalities. This topic is particularly discussed within initiatives and projects concerned with the representation of lexical information in a Linked Data (LD) compliant format,[1] so that they can be published within the Linguistic Linked Open Data (LLOD) cloud. In this context, we can observe that Sign Language (SL) lexical data are not currently represented in the datasets included in the LLOD cloud.[2] Looking at the "Overview of Datasets for the Sign Languages of Europe", published by the "Easier" European project,[3] we also do not see any mention of a dataset being available in an LD-compliant format. We therefore investigate ways of representing SL data in the LLOD cloud and linking them to other types of language data already available in an LD-compliant format.

Design/methodology/approach: A first step in our work consisted in an extensive study of the literature on Sign Languages, including websites and online demos. We noticed a huge disparity in the use of vocabularies for describing constitutive elements of Sign Languages. Therefore, we began with the building of a harmonized taxonomy of such descriptors. This resulted in a (tentative) structured listing of more than 260 harmonized elements, which can be used for linking to lexical resources already available in the Linked Data format (mainly lexical resources encoded in the context of the OntoLex-Lemon framework4). The current version of this harmonized resource is available on GitHub.[5] In a second step, we looked at current transcription approaches dealing with Sing Languages. This included the SignWriting and the HamNoSys models. We concentrated, for now, on the second model.[6] These transcription data are helping to establish a link between elements of the harmonized taxonomy and lexical entries or forms available in OntoLex-Lemon encoded data sets. As HamNoSys itself is not directly machine readable, we make use of a conversion tool,[7] which is transforming Ham-

[1] See the past projects Prêt-à-LLOD (https://pret-a-llod.github.io/) and ELEXIS (https://elex.is/), and the running Cost Action "NexusLinguarum" (https://nexuslinguarum.eu/), in the context of which the current investigation is being conducted.

[2] See http://linguistic-lod.org/llod-cloud for more details.

[3] See https://www.project-easier.eu/wp-content/uploads/sites/67/2021/08/EASIER-D6.1-Overview-of-Datasets-for-the-Sign-Languages-of-Europe.pdf

[4] See https://www.w3.org/community/ontolex/wiki/Final_Model_Specification for more details.

[5] See https://github.com/Declerck/sl-onto

[6] See https://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/HamNoSys_2018.pdf for more details

[7] This tool is described, among others, at https://github.com/carolNeves/HamNoSys2SiGML

NoSys into SiGML – an XML representation which is also often used for the animation of avatars. Current work is dedicated to establishing the best possible way of linking SiGML data to lexicographic data of the spoken/written languages that are encoded in OntoLex-Lemon.

Findings: The complexity and richness of Sign Language data require a sound approach to the form of linking that we are aiming to achieve. We are in the process of integrating members of the Sign Languages community into discussions that are taking place in the Linked Data community.

Research limitations/implications: A limitation is provided by the relatively small number of open data for Sign Languages, but there are enough to propose a general approach for interlinking Sign Language data and the LD-compliant resources available for spoken/written lexical data. Some initiatives are investigating the use of Open Multilingual WordNet datasets8 as a possible (concept based) bridge between Sign Language data, using their transcription in SL glosses.

Practical implications: Towards the creation of multimodal lexical resources in the Linguistic Linked Data framework. The resources generated can possibly support automated translations between modalities.

Originality/Value: We are not aware of related work in this field (Linked Data and Sign Languages), but we see that there is a growing interest in using WordNet data within the SL community, and some of those data are already available in LD-compliant format.

Keywords: *Linguistic Linked Data, Sign Languages, WordNet.*

Research type: Research paper.

8 See http://compling.hss.ntu.edu.sg/omw/

▶ Lithuanian Poetry Translation to Spanish: A Review of Two Decades (2000–2021)

Vilhelmina Vaičiūnienė,
Akvilė Šimėnienė,
Mykolas Romeris University, Lithuania,
vvaiciun@mruni.eu
akvisim@mruni.eu

Purpose: Translation studies has always had a historical component, which has been coming increasingly to the fore in recent years. The discourse on translation has been spinning around accuracy, fluency, meaning, equivalence, and fidelity, which, in fact, do not have any direct links with the history of translation. The current article aims to review the scope of translation of Lithuanian poetry into Spanish over two decades (2000–2021).

Design/methodology/approach: We take a historical perspective to analyze the translations of Lithuanian poetry into Spanish over a period of 20 years. The historiographic research of translations is conducted according to the following criteria: firstly, the time span (two decades), secondly, poets whose poetry is translated in to Spanish, and thirdly, translators of Lithuanian poetry into Spanish, among them Birutė Ciplijauskaitė and Carmen Caro Dugo.

Findings: There are specific trends in literary translation including poetry translation. Since the regaining of independence in 1990, translation of literary works from other languages, including Spanish, prevailed rather than translation from the Lithuanian language. The dominance of women's poetry translation is noticeable in the first decade (2000–2010) under analysis. The research findings on the second decade (2011–2021) indicate a bigger scope of translations and a greater variety of poetry genres and authors. Apart from the scope of poetry translation, a very important point to be discussed is the translator's identity in the light of two cultures (Spanish and Lithuanian) and their role as intercultural communicators and/or cross-cultural mediators in different historical contexts. Moreover, the responsibility of the translator for the choices they make in the translation process, as well as the invisibility phenomenon discussed by Venuti (1995), are also analyzed in view of the historical and social context.

Research limitations/implications: The analysis of translations focuses exclusively on two decades (2000–2021) of Lithuanian poetry translation into Spanish.

Originality/Value: Translation of Lithuanian authors to other languages has been a problematic issue since its status as a minority language, and Spanish is one of the most popular languages in the world. Moreover, translations of Lithuanian poets

into Spanish have not been investigated from a historiographic perspective so far.

Keywords: *poetry translation, historiographic approach, the Lithuanian and Spanish languages.*

Research type: Research paper.

The Effectiveness of Video Creation in the ESL Classroom in Lithuania: A Case Study

Vitalija Jankauskaitė-Jokūbaitienė,
Vilnius University, Lithuania,
vitalija.jankauskaite-jokubaitiene@flf.vu.lt

Purpose: The growing number of research investigating the efficacy of Web 2.0 tools in the ESP classroom (Selevičienė, 2020; Almuhaisen et al, 2020; Rodgers & Dhonnchadha, 2018; Šulovská, 2013) has revealed the positive effect of technology integration on the development of students' language learning in higher education. However, the implementation of web-based technologies, namely digital video creation for language learning, has been scarcely discussed through the prism of a K-12 environment. The purpose of this study is to examine the effectiveness of video making in L2 acquisition of English from the perspective of secondary learners and explicate the potential strengths and weaknesses of its utilisation in the ESL classroom.

Design/methodology/approach: A qualitative research methodology was employed to analyze an open-ended survey. Tenth and ninth graders were asked to produce their responses to a two-fold questionnaire aimed at evaluating the use of video making for revisiting target vocabulary of the lesson.

Findings: The current research indicates a polarized view of video making practice in the ESL classroom. Regarding the potential of video creation, learners admitted it being beneficial for the acquisition of vocabulary, when it comes to memorizing it, forming strong associations, comprehending and communicating the meaning more properly. Yet the results of the study also demonstrated that creating content via a video editing tool proved to be relatively demanding and time consuming due to the lack of ICT skills in video making, insufficient internet connection speed or restrictions imposed on time, and vocabulary limit.

Research limitations/implications: One of the limitations of this study is its small sample of respondents participating in the survey, i.e. 20 tenth graders and 8 ninth graders sharing their experiences on the matter. In addition, the teacher provided only a basic instruction to the video making software Biteable.com, which was unfamiliar to learners beforehand and made the learning process more complicated than expected. Despite these limitations, the majority of teams completed the task successfully by producing short film/book trailers within the framework of two lessons of 45 minutes.

Practical implications: The study may provide English language instructors with further insights as to how the use of video creation may facilitate the learners' acquisition of vocabulary and how possible hindrances can be prevented to achieve more

effective language learning outcomes with the aid of digital video making tools in the ESL classroom.

Originality/Value: The major contribution to the research in this field is its focus on the adoption of a video making method in the English as the second language classroom considering the context of secondary education in Lithuania.

Keywords: *video creation, Web 2.0 technologies, ESL, secondary education, language acquisition.*

Research type: Case study.

▶ Stylistic Means of Verbalizing Images of the Russian-Ukrainian War

Yuliia Shpak,
H.S.Skovoroda Kharkiv National Pedagogical University, Ukraine, starlingyu@gmail.com
Ganna Krapivnyk,
H.S.Skovoroda Kharkiv National Pedagogical University, Ukraine, krap302@gmail.com

Purpose: The paper aims at considering the ways of verbalizing the stylistic devices used to describe the images of the Russian-Ukrainian war in the selected Ukrainian and British mass media.

Design/methodology/approach: The study comprises the stylistic and linguocultural analysis of the articles taken from the selected online Ukrainian and British periodicals. The above analysis was focused on the major images of the ongoing war and the linguistic means of their description. The methods applied in the work include the sampling and systematic selection, stylistic, and linguocultural analysis of the text, as well as partially intertextual analysis.

Findings: The conducted study showed similarities and differences between the ways British and Ukrainian authors have depicted the situation in Ukraine since February 24, 2022. In this respect, British media provided a more reserved and pragmatic position compared to the Ukrainian ones. They sought to be observers trying to describe the developing events from outside, whereas Ukrainian authors tended to use more expressive means and stylistic devices in their texts. It is to be noted that diachronically, Ukraine war discourse had been changing, though predominantly in the British texts under analysis, which may well be related to potential war fatigue.

Research limitations/implications: The conducted survey is not exhaustive since the empirical material includes only a number of articles from the mass media resources (in particular, *NV*, *Ukrainian Pravda*, *The Guardian*, and *The Independent*); therefore, the research findings may be considered the foundation for further study on the topic.

Practical implications: The research may be well used to study psychological, sociological and philosophical aspects of the war as a phenomenon in human life, as manifested in the verbalization of the war in Ukraine, taking the perspective of the authors being directly involved in the crisis and those staying outside.

Originality/Value: The body of selected media texts under analysis is up-to-date and reflects the current usage of the language material to express the images referring to

the Russian-Ukrainian war.

Keywords: *media, image, stylistic means, epithet*.

Research type: Research paper.

WORKSHOP ABSTRACTS

▶ Machine Learning Methods for Discourse Marker Detection in Italian

Emma Angela Montechiari,
University of Trento, Trento, Italy, emma.angela.0812@gmail.com,
Stanko Stankov,
Mozaika, Ltd. Solunska 52, Sofia 1000, stanko.stankov@mozajka.co
Kostadin Mishev,
Department for Information Systems and Network Technologies,
Faculty of Computer Science and Engineering Ss. Cyril and Methodius University ul. RugerBoskovik 16 1000 Skopje, North Macedonia, kostadin.mishev@finki.ukim.mk,
Mariana Damova
Mozaika, Ltd. Solunska 52, Sofia 1000, mariana.damova@mozajka.co,

The latest advances in NLP, more precisely NLP Transformers, show great performance in building universal language representations. The trained vectors of words or sentences can provide unique representation for multiple languages, exclusively extracting semantic information from texts that is mapped into shared embedding space. This semantic information can be leveraged to train a model for specific downstream tasks, such as text classification, clustering, and others, while also leveraging semantic information for language understanding. The resulting model from the training phase can be universally used for all languages whose shared vector space is encompassed, thus avoiding the need to train separate models for each language individually.

Attitudinal discourse markers play a central role in the sense that they are pointers towards the speaker's attitudes. They are single-word or multiword expressions (MWE), and are mainly drawn from syntactic classes of conjunctions, adverbials, and prepositional phrases (Fraser, 2009), as well as expressions such as *you know*, *you see*, and *I mean* (Schiffrin, 2001; Hasselgren, 2002; Maschler & Schiffrin, 2015). Discourse markers are regarded as significant discourse relation triggers, and, consequently, are broadly studied (e.g. Sanders et al., 1992; Knott & Dale, 1994; Wellner et al., 2006; Taboada & Das, 2013; Das, 2014; Das & Taboada, 2019; Silvano, 2011). Recently, discourse relations and discourse marker research has gained certain impetus with corpora annotation for exploring discourse structure in texts – for example: RST-DT English corpus (Carlson, Marcu & Okurowski, 2003); Penn Discourse Tree Bank (PDTB) (Prasad et al., 2008); and the SDRT Annodis French corpus (Afantenos et al., 2012). This paper presents several experiments regarding discourse marker detection with a number of machine

learning methods carried out on an Italian corpus containing 2,000 manually annotated contexts and a larger number of contexts that have not been annotated. We present the results of experiments with the following machine learning methods: a) FastText, b) XML Roberta, and c) LABSE (Language agnostic BERT Sentence Embedding).

The annotated Italian corpus is included in a larger parallel corpus of 14,785 examples of Italian-English correspondence, of which 2,000 have been analyzed. The corpus will be included in a parallel corpus containing data from 7 languages, using publicly available TEDTalk transcripts (Silvano, 2022). The multilingual corpus contains alignments of the Italian, Lithuanian, Bulgarian, Portuguese, Macedonian, and German languages, with English as the pivot language and with a size of 1.3 million sentences. This itself is an ongoing expansion of the TED-EHL parallel corpus published in the LINDAT/CLARIN-LT repository (link: http://hdl.handle.net/20.500.11821/34). The corpus is divided into parallel English and Italian parts, and both are then divided into several parts, e.g. sentence chunks (A) and discourse marker (DM) chunks (B). These are the sentences in which the expression of the marker (A) and the particular expression of the discourse marker, if there is one (B), must be identified (Hutchinson, 2004; Mishev, 2022; Hasselgren, 2002). There is then a third area (C) which reports the textual expression that constitutes the discourse context in which the sentences are included. The last columns concern the validation of the presence of the discourse marker – manual (D) and automatic (E) (Crible, 2014) (see Table 1).

Table 1

Sentence Chunk EN	Context EN	DM EN	Sentence Chunk IT	Context IT	DM IT
BR: So you see us here cutting up	for all confectionary products and sodas, and we can replace it with all- natural fresh fruit. BR: So you see us here cutting up some watermelon. Theidea with this is that we're going to eliminate tons of	you see	BR: Qui civedete mentre tagliamo	in tutti i prodotti confezionati e nelle bibite, sostituendolo con frutta fresca naturale. BR: Qui civedete mentre tagliamo l'anguria. L'idea è di eliminaretonnellate	Qui ci vedete

The manual annotation process assigns the presence or absence of a discourse marker in the sentence chunk and in the sentence context in an extra column (see Table 2).

Table 2

Sentence ChunkIT	Context IT	DM IT	DM Presence IT	DM Presence AUT
perché almeno, **sisa,** un ago è qualcosa su cui	In realtà, trovare un ricordo nel cervello non è poi così semplice. XL: Anzi, è ancora più difficile che trovare un ago in un pagliaio, perché almeno, **si sa**, un ago è qualcosa su cui si possono mettere le mani fisicamente. Ma il ricordo non lo è.	, si sa,	1	

Language-agnostic BERT Sentence Embedding (LaBSE) is a BERT-based transformer model, which maps a sentence into a fixed-length vector representation and supports 109 languages. The model is trained on 17 billion monolingual sentences and 6 billion bilingual sentence pairs using Masked Language Model (MLM) and Translation Language Modeling (TLM) objectives, resulting in a model that is effective even on low-resource languages for which there is no data available during training. (Hutchinson, 2003; Joulin, 2016; Mishev, 2021) In our paper, we use the sentence representation produced by LaBSE and apply them in a binary text classification (BTC) task. The main objective of the task is to detect the presence of a DM in the sentence (Bunt, 2016; Damova, 1998); it returns 1 if a DM is present, and 0 if a DM is absent in the sentence provided as an input. Initially, the downstream task is trained on the English datasets, and evaluation is performed using the Italian dataset.

The annotated corpora has been used to train machine learning models to predict the existence of discourse markers in a text. Because we had a multilingual dataset, we chose FastText (Joulin et al., 2016) and XLM-Roberta-Large (Conneau et al., 2019) as the base models. The model was fine-tuned using the k-train library (Maiya, 2022), a low-code Python library built on top of the state-of-the-art Transformers library (Wolf et al., 2020). The dataset was divided 80-20 for train and test datasets, and the model was trained using a learning rate of 0.00001 for three epochs. The dataset was slightly unbalanced (53% records without a discourse marker and 47% with a discourse marker), so we used class balancing weights to compensate. The model fine-tuning was run ten times, and the average performance is reported in Table 3 and Figure 1. LaBSE is fine-tuned on the English dataset to detect the presence of Discourse Markers. Next, the same model was used for the evaluation on the Italian dataset without additional fine-tuning. The whole English dataset was used as a training set, and the whole Italian dataset as a test set.

Table 3

	FastText	XLM-RoBERTa-Large	LaBSE
Accuracy	0.5695	0.6890	0.7051
Precision	0.7394	0.8000	0.7768
Recall	0.5083	0.6862	0.7573
Specificity	0.6791	0.6940	0.6119
F1-Score	0.6025	0.7387	0.7669
MCC	0.1810	0.3667	0.3659

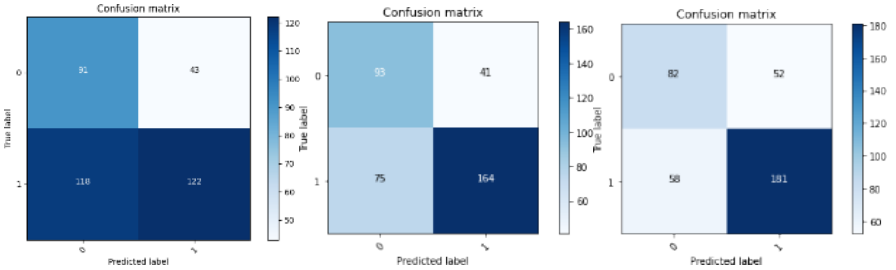


Figure 1

The presented methods will be applied to a parallel corpus of 7 languages from TED talks with English as a pivot language, and the results will be compared with the results for Italian. The presented annotated Italian corpus will be included into the parallel corpus of 7 languages of TED talks and will be published in CLARIN. Further comparative linguistic analysis will be carried out on the differences in occurrence of discourse markers in Italian and the other languages.

References

Afantenos, S., Asher, N., Benamara, F., Cadilhac, A., Dégremont, C., Denis, P., Guhe, M., Keizer, S., Lascarides, A., Lemon, O., Muller, P., Paul, S., Popescu, V., Rieser, V., & Vieu, L. (2012). Modelling strategic conversation: Model, annotation design and corpus. In Brown-Schmidt, S., Ginzburg, J., Larsson, S. (Eds.), *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*.

Bunt, H., & Prasad, R. (2016). ISO DR-Core (ISO 24617-8): Core Concepts for the Annotation of Discourse Relations. In: *Proceedings 10th Joint ACL-ISO Workshop on Interoperable Semantic Annotation* (pp. 45–54).

Buschbeck-Wolf, B., Butt, M., Damova, M., Dorna, M., Eberle, K., Emele, M.,

Nubel, R., Reinhard, S., Ripplinger, B., Schiehlen, M., & Schmid, H. (1995). *Transfer in the Verbmobil Demonstrator*. Verbmobil-Report 147. IAI Saarbrucken. IBM Deutschland Informations systeme GmbH. Universität Stuttgart; Eberhard-Karls- Universitat Tubingen.

Crible, L. (2014). *Identifying and describing discourse markers in spoken corpora. Annotation protocol v.8*. Technical report. Louvain-la-Neuve: Université catholique de Louvain.

Damova, M. (1998). *Tense and Aspect in Discourse*: *A Study of the Interaction between aspect, discourse relations and temporal reference within Discourse Representation Theory with special attention to Bulgarian*. PhD dissertation, University of Stuttgart.

Das, D. (2014). *Signalling of coherence relations in discourse*. PhD dissertation, Simon Fraser University.

Das, D., & Taboada, M. (2019). Multiple signals of coherence relations. *Discours*, *24*, 3–38.

Carlson, L., Marcu, D., & Okurowski, M. E. (2003). Building a discourse-tagged corpus in the frame work of Rhetorical Structure Theory. In J. van Kup Pevelt & R. Smith (Eds.), *Current Directions in Discourse and Dialogue* (pp. 85–112). Kluwer Academic Publishers.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics.

Fraser, B. (2009). An Account of Discourse Markers. *International Review of Pragmatics*, *1*(2), 293–320.

Hasselgren, A. (2002). Learner corpora and language testing: Small words as markers of learner fluency. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 143–174). John Benjamins Amsterdam.

Hutchinson, B. (2003). Automatic classification of discourse markers on the basis of their co-occurrences. In M. Stede & H. Zeevat (Eds), *Proceedings of the ESSLLI Workshop The Meaning and Implementation of Discourse Particles* (pp. 1–8).

Hutchinson, B. (2004). Acquiring the Meaning of Discourse Markers. In *Proceedings of the 42nd Annual Meeting* (pp. 684). Association for Computational Linguistics.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter*

of the Association for Computational Linguistics: Volume 2, Short Papers (pp. 427–431). Association for Computational Linguistics

Knott, A., & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, *18*(1), 35–62.

Maiya, A. S. (2022). Ktrain: A low-code library for augmented machine learning. *Journal of Machine Learning Research*, *23*(158), 1–6.

Maschler, Y., & Schiffrin, D. (2015). Discourse markers: Language, meaning, and context. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), *The handbook of discourse analysis* (2nd ed., pp. 189–221). Wiley Online Library.

Mishev, K., Valunaite-Oleškeviciene, G., Liebeskind, C., Trajanov, D., Silvano, P., Truica, C.-O., Apostol, E.-S., Chiargo,s C., & Damova, M. (2021). Speaker Attitudes Detection through Discourse Markers Analysis. In *Proceedings of Workshop "Deep Learning and Neural Approaches for Linguistic Data"*. Skopje: NexusLinguarum.

Mishev, K., Valunaite-Oleškeviciene, G., Liebeskind, C., Trajanov, D., Silvano, P., Truica, C.-O., Apostol, E.-S., Chiargos, C., & Damova, M. (2022). Evaluation of Cross-Lingual Methods for Discourse Markers Detection. In *DiSLiDaS 2022 workshop: Using cross-lingual methods to detect and predict DMs' presence in texts, over contexts of annotated corpora*. Jerusalem: NexusLinguarum.

Piurko, E. (2015). *Discourse markers: their functions and distribution in the media and legal discourse*. Master's thesis, Lithuanian University of Educational Sciences.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA).

Sanders, T., Spooren, J. W., Leo, & Noordman, G. M. (1992). Toward a Taxonomy of Coherence Relations. *Discourse Processes*, *15*, 1–35.

Schiffrin, D. (2001). Discourse markers: Language, meaning, and context. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), *The handbook of discourse analysis* (1st ed., pp. 54–75). Oxford: Blackwell Publishing.

Silvano, P. (2011). *Temporal and rhetorical relations: the semantics of sentences with adverbial subordination in European Portuguese*. PhD dissertation, University of Porto.

Silvano, P., Valunaite-Oleškevicien, G., Liebeskind, C., Chiargos, C., Trajanov, D., Truica, C.-O., Apostol, E.-S., Baczkowska A., & Damova M. (2022). *ISO-based Annotated Multilingual Parallel Corpus for Discourse Markers* [Poster]. Presentation of an

ISO-based parallel multilingual corpus for DMs (validation process) (deep learning for discourse relations and communicative functions and LinguisticLink Open Data). LREC 2022, Marseille, France.

Taboada, M., & Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, *4*(2), 249–281.

Wellner, B., Pustejovsky, J., Havasi, C., Rumshisky, A., & Saurí, R. (2006). Classification of Discourse Coherence Relations: An Exploratory Study using Multiple Knowledge Sources. In *Proceedings of the7th SIGdial Workshop on Discourse and Dialogue* (pp. 117–125). Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen P., Ma C., Jernite, Y., Plu, J., Xu, C., Le Scao T., Gugger. S., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics.

► Evaluating Synonym and Antonym Acquisition from a Portuguese Masked Language Model

Hugo Gonçalo Oliveira,
CISUC, Department of Informatics Engineering
University of Coimbra, Coimbra, Portugal
hroliv@dei.uc.pt

As for many other tasks in the domain of Natural Language Processing, transformer-based models have been explored in the acquisition of semantic relations, and are useful for the automatic creation or enrichment of knowledge bases [1], e.g., represented in RDF. Towards this, they can be used for completing lexico-syntactic patterns, in a shortcut to earlier methods of relation acquisition from corpora [12]. When focused on lexico-semantic relations, they can be useful for enriching lexical knowledge bases (LKBs) like WordNet [4].

For Portuguese, BERT [3] has been used for detecting hyponymy pairs [14] and discovering arguments of lexico-semantic relations [9]. Here, we adopt the previous method, but focus on synonymy and antonymy, both symmetric relations that may pose different challenges. In what can be seen as a zero-shot learning approach, we prompt BERTimbau [15] with masked lexical patterns that transmit the target relations, and rely on the model predictions for the masks. For evaluation, we propose two available tests of Portuguese lexico-semantic knowledge: B2SG [16] and TALES [10].

B2SG is similar to the WordNet-Based Synonymy Test [5], but based on the Portuguese part of BabelNet [13] and partially evaluated by humans. It contains Portuguese words (source) followed by four candidates, out of which only one is related, and is organised in three relations: synonymy, antonymy, and hypernymy between nouns and verbs, respectively. An example for noun-synonymy is: cataclismo desastre_noun talha_noun obesidade_noun alusão_noun (cataclysm disaster obesity allusion).

TALES [10] was created from the contents of ten lexical resources for assessing lexico-semantic analogies in Portuguese, and its format is similar to BATS [7]. For each relation, TALES includes 50 entries with two columns: a word (source) and a list of related words (targets). An example for antonym-of is: `novo velho/idoso/entradote` (young old/aged/oldish).

Given the source words, several handcrafted lexical patterns indicating synonymy and antonymy were tested with two versions of BERTimbau (base and large) for selecting the related words in B2SG, and predicting the target words in TALES. For BSG, the process is simplified with FitBERT,[1] a tool that, given a masked sentence and a

[1] https://github.com/Qordobacode/fitbert

list of options, selects the most suitable option for the mask, based on pre-softmax logit scores [8] computed on BERT. Performance in both tests can be measured by accuracy. In BSG, we also look at the average ranking of the correct answer (1–4), and in TALES at the presence of correct words in the top-10 predictions, i.e., Accuracy@10. The tested patterns included those from the relation-validation service VARRA [6], which were among the best performing.

Table 1 displays the best performing patterns for each relation and test. From the accuracy in B2SG, we confirm that, despite being far from perfect, answers are also far from random, which would yield 0.25. A preliminary analysis suggests that BERT-large is preferable for the majority of relations, and that the procedure may suit antonymy better. The restricted number of antonyms of a word may contribute to this.

When compared to the performance for other relations, the best performance in BSG is for noun-antonymy, even more accurate than noun-hypernymy (0.64), while verb-synonymy is less accurate, but still higher than verb-hypernymy (0.52). In TALES, comparison is more difficult, because accuracies are typically lower and the test covers different types of hypernymy and hyponymy.

Due to its connection with semantic similarity, an initial hypothesis was that synonymy is better captured by similarity-based methods than with a single lexical pattern. So, we tested three different approaches for answering B2SG by selecting the candidate that maximises similarity with the source, based on embeddings from: the `CLS` token of the pretrained BERTimbau; BERTimbau fine-tuned for Natural Language Inference[2]; and GloVe pretrained for Portuguese [11].[3] Accuracy was worse with the pretrained model, except for noun-synonymy (0.67); with the NLI model, there were minor improvements for all but verb-antonymy; while GloVe clearly outperformed the pattern-based approach, with accuracies between 0.7 (noun-synonymy) and 0.83 (noun-antonymy). This supports the initial hypothesis, not only for synonymy, but also antonymy.

In the future, we will deepen this analysis and use the same datasets for evaluating the acquisition of lexico-semantic relations from neural language models. A similar analysis, based on the completion of patterns, may also be performed in generative models like GPT-3 [2] or similar ones that recently became available [17].

[2] `ricardo-filho/bert-portuguese-cased-nli-assin-assin-2` in HuggingFace.
[3] <1% of the words in the test were not covered by the vocabulary of this model.

Table 1: Best performing patterns for each relation in each test.

B2SG						
Relation	**PoS**	**Pattern**	**BERT-base**		**BERT-large**	
			Acc	**Rank**	**Acc**	**Rank**
Synonym-of	N	é o mesmo que [MASK]	0.57	1.71	0.64	1.58
Synonym-of	V	, isto é, [MASK]	0.50	1.80	0.56	1.67
Antonym-of	N	nem [MASK], nem	0.76	1.64	0.77	1.36
Antonym-of	V	nem [MASK], nem	0.63	1.64	0.61	1.61
TALES						
Relation	**PoS**	**Pattern**	**BERT-base**		**BERT-large**	
			Acc	**Acc@10**	**Acc**	**Acc@10**
Synonym-of	N	é sinónimo de [MASK]	0.28	0.64	0.20	0.70
Synonym-of	V	é o mesmo que [MASK]	0.12	0.80	0.34	0.90
Synonym-of	ADJ	estar é o mesmo que estar [MASK].	0.06	0.46	0.24	0.54
Antonym-of	ADJ	ser [MASK] é o contrário de ser	0.26	0.40	0.38	0.48

Acknowledgements: This work was partially supported by: the COST Action CA18209 Nexus Linguarum (European network for Web-centred linguistic data science); national funds through the FCT – Foundation for Science and Technology, I.P., within the scope of the project CISUC – UID/CEC/00326/2020 and by the European Social Fund, through the Regional Operational Program Centro 2020.

References

1. AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M., Ghazvininejad, M.: A review on language models as knowledge bases. https://arxiv.org/abs/2204.06031 (2022)

2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)

3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc 2019 Conf of North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186. ACL (2019)

4. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)

5. Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., Wang, Z.: New experiments in distributional representations of synonymy. In: Procs 9th Conference on Computational Natural Language Learning (CoNLL-2005). pp. 25–32. ACL, Ann Arbor, Michigan (2005)

6. Freitas, C., Santos, D., Gonçalo Oliveira, H., Quental, V.: VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. In: Pesquisas e perspetivas em linguistica de corpus (Livro do IX Encontro de Linguística de Corpus, 2010), pp. 199–232. ELC 2010, Mercado de Letras, Rio Grande do Sul, Brasil (2015)

7. Gladkova, A., Drozd, A., Matsuoka, S.: Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In: Procs of NAACL 2016 Student Research Workshop. pp. 8–15. ACL (2016)

8. Goldberg, Y.: Assessing BERT's syntactic abilities. https://arxiv.org/abs/1901.05287 (2019)

9. Gonçalo Oliveira, H.: Drilling lexico-semantic knowledge in Portuguese from BERT. In: Computational Processing of the Portuguese Language – 14th International Conf, PROPOR 2022. LNCS, vol. 12037, pp. 387—397. Springer (2022)

10. Gonçalo Oliveira, H., Sousa, T., Alves, A.: TALES: Test set of Portuguese lexical-semantic relations for assessing word embeddings. In: Procs of ECAI 2020 Workshop on Hybrid Intelligence for Natural Language Processing Tasks (HI4NLP 2020). CEUR Workshop Proceedings, vol. 2693, pp. 41–47. CEUR-WS.org (2020)

11. Hartmann, N.S., Fonseca, E.R., Shulby, C.D., Treviso, M.V., Rodrigues, J.S., Aluísio, S.M.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: Proc of 11th Brazilian Symposium in Information and Human Language Technology (STIL 2017) (2017)

12. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proc 14th Conference on Computational Linguistics. pp. 539–545. COLING 92, ACL (1992)

13. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250 (2012)

14. Paes, G.E.: Detecção de Hiperônimos com BERT e Padrões de Hearst. Master's thesis, Universidade Federal de Mato Grosso do Sul (2021)

15. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: Pretrained BERT models for

Brazilian Portuguese. In: Proc of Brazilian Conf on Intelligent Systems (BRACIS 2020). LNCS, vol. 12319, pp. 403–417. Springer (2020)

16. Wilkens, R., Zilio, L., Ferreira, E., Villavicencio, A.: B2SG: a TOEFL-like task for Portuguese. In: Procs 10th International Conference on Language Resources and Evaluation (LREC 2016). ELRA (may 2016)

17. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L.: OPT: Open pre-trained transformer language models. https://arxiv.org/abs/2205.01068 (2022)

▶ Lexico-Semantic Relation Classification With Multilingual Finetuning

Lucía Pitarch
Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain
Lacramioara Dranca
Centro Universitario de la Defensa (CUD), Zaragoza, Spain
Jorge Bernad
Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain
Jorge Gracia
Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain
{lpitarch, licri, jbernad, jogracia}@unizar.es

Following the research line of lexico-semantic relation induction based on Language Models (LM) [4, 1, 5], this work analyses how the multilingual fine-tuning of a LM might improve its performance on a relation classification task. The model is trained on a set of word pairs which share a known relation (e.g., *Big–small*, antonymy). Then, it is used to classify the relation between an unseen word pair. Our hypothesis is that enriching this training with multilingual information benefits the classification of lexico-semantic relations that are common in several languages.

Two different experiments were conducted, based on: (i) crosslingual transfer and (ii) multilingual verbalisation. Both consist of fine-tuning the multilingual BERT (mBERT) pre-trained model [2] with multilingual data. LMs in general need a verbalised input [1], and although different verbalisations of the lexico- semantic relations are possible, analysing them is out of our scope. We focus on two simple verbalisations: *v1* for one relation pair (e.g., *'big' is related to 'small'*) and *v2* for two relation pairs (e.g., *'big' is related to 'small' as 'grande' is related to 'pequeño'*, in this case for two English words and its Spanish translations). The considered relations in the classification task are: *hypernymy, antonymy, synonymy, meronymy*. In total, 50 word pairs were extracted from the BATS data set [3] in English and Japanese, per each relation category, and manually translated to Spanish. We also created an *others* category for cases in which none of the four relations are present, extracting 50 pairs of words not explicitly related in BATS, thus finally having five possible categories of relations and 250 word pairs per language. We performed 5 runs of each experiment. We report the median and a 95% confidence interval. We chose the median since it is more informative than the average for non-symmetric distributions, as in our case.

In the first experiment, *crosslingual transfer* [6] is explored by fine-tuning the model with one (English) and two (English and Spanish) languages and then testing the results on English, Spanish and Japanese. Verbalisation *v1* is used. The results are summarized in Table 1. Crosslingual transfer to Japanese improves when training the model with both English and Spanish. The median of the accuracy was 0.42 when the model

was just only on English data, and improves up to 0.52 when the fine-tuning is multilingual. This technique led to an improvement of the results, in line with our hypothesis.

Table 1. Crosslingual transfer experiment: accuracy results

Training language EN; training size: 125			
Test language	Median	CI 95%	Test size
EN	0.64	[0.6, 0.79]	100
ES	0.58	[0.52, 0.74]	250
JP	0.42	[0.3, 0.48]	250
Training language EN + ES; Training size: 450			
Test language	Median	CI 95%	Test size
JP	0.52	[0.51, 0.52]	250

The second experiment consisted of testing a *multilingual verbalisation*. Both *v1* (one word pair either in Spanish or English) and *v2* (two word pairs, one in English and the other in Spanish) were tested. The results are summarized in Table 2. Monolingual training and shorter verbalisations seem to perform better in this case.

In conclusion, while crosslingual transfer seems to improve the results, multilingual verbalisation is not as promising. However, results are inconclusive due to the small amount of data, leading high variability in accuracy results. In future research we plan to extend the dataset to include more languages and examples.

Table 2. Multilingual phrasing experiment: accuracy (training size: 125, test size: 125)

	Training languages EN + ES		Training language EN		Training language ES	
verbalisation	Median	CI 95%	Median	CI 95%	Median	CI 95%
v1	0.36	[0.28, 0.53]	0.64	[0.56, 0.65]	0.55	[0.49, 0.57]
v2	0.20	[0.00, 0.21]	0.6	[0.57, 0.63]	0.45	[0.31, 0.71]

References

1. Bouraoui, Z., Camacho-Collados, J., Schockaert, S.: Inducing relational knowledge from BERT (2020). https://doi.org/10.1609/aaai.v34i05.6242

2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidi- rectional transformers for language understanding. CoRR abs/1810.04805 (2018), http://arxiv.org/abs/1810.04805

3. Gladkova, A., Drozd, A., Matsuoka, S.: Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. (2016). https://doi.org/10.18653/v1/n16-2002

4. Nayak, T., Majumder, N., Goyal, P., Poria, S.: Deep neural approaches to relation triplets extraction: a comprehensive survey. vol. 13 (2021). https://doi.org/10.1007/s12559-021-09917-7

5. Oliveira, H.G.: Acquiring lexico-semantic knowledge from a portuguese masked language model (2021)

6. Zhao, M., Zhu, Y., Shareghi, E., Vulic, I., Reichart, R., Korhonen, A., Schütze, H.: A closer look at few-shot crosslingual transfer: The choice of shots matters (2021). https://doi.org/10.18653/v1/2021.acl-long.447

► Extracting and Linking Morphological Data From the Pre-Standard Croatian Grammars Using TEI

Martina Kramarić,
Institute of Croatian Language and Linguistics, Zagreb, Croatia
mkramar@ihjj.hr

Digital humanities offer the possibilities and tools to overcome the limitations of print technology (Robinson, 2002, p. 43). This is especially useful for conducting any kind of linguistic analysis on written texts. DH can also give new insights in processing older linguistic data, historical dictionaries, or grammar books.

In the literature, the common title "Croatian pre-standard grammars" refers to a heterogeneous group of grammar books with common structural features and methods of description, and whose metalanguage or object language[1] is one of the Croatian literary languages that precede the modern Croatian standard language based on Štokavian dialect: Čakavian, Kajkavian, and Štokavian.

The first Croatian grammar was written in 1604, but in the Latin language describing the Croatian language. The other grammars are also diverse when it comes to their object language and the language in which they were written. There are different combinations of languages involved – Croatian with its three dialects, Latin, Italian, and German. The digitization of these grammar books is being conducted in the Institute of Croatian Language and Linguistics within the project Retro-Digitization and Interpretation of Croatian Grammar Books before Illyrism – RETROGRAM.

The main goal of the Retrogram project is to create a model for the digitization of older grammar books, which will enable research of the linguistic material they provide and their mutual linking. Therefore, the digitization in our case does not apply only to the written grammar text, but also to the morphological paradigms of the eight selected old Croatian grammars written from the 17th to the 19th century.[2] All of the morphological paradigms, declensions, and conjugations will be annotated using TEI tags (Text Encoding Initiative – TEI), which will enable the linking of morphological data from all eight diverse grammars. TEI is one of many XML languages and can be easily modified to edit historical documents of any kind (Pierazzo, 2014, p. 14).

The final result of the project will be a web portal that will present Croatian grammar books including facsimiles of selected grammar books with their formal description, encoded transcription or translation of the grammar book text, and an index

[1] With the term object language, we refer to the language which is the object of grammatical description.

[2] To see the description of the project, members of the project, and the list of grammars visit https://retrogram.jezik.hr/ and Horvat – Kramarić 2021.

of grammatical and linguistic terminology. The grammar text will be searchable and connected to the facsimiles. The portal will be equipped with the possibility of morphological thematic search according to the pre-agreed parameters, which are in this case declensions and conjugations.

There were several obstacles in the annotation work on the linguistic material which was written in the grammar books. We overcome the fact that the grammar books were written in different languages and their object language is different by translation or transcription of the grammar text into the Croatian language. Only the Croatian morphological paradigms were annotated. All of those eight selected grammars were written according to the Latin language grammar model, and that is the reason why they are very different from the standard Croatian grammar books. Therefore, before the annotation we also had to conduct deep philological and comparative analyses of all grammar books at all linguistic levels. The next question concerned which grammar model we would choose for the annotation work. Such a model should enable us to perform the extraction and linking of the morphological data and comparison between different grammar books. For these reasons, we decided to annotate only morphological data which can fit into the Standard Croatian grammar model. Thus, we will not annotate the attempts of the Croatian grammarians to translate the Latin subjunctives and optatives and fit them into Croatian verbal paradigms. This sometimes results in absurd verbal constructions.

When annotating verb paradigms, special attention must be given to the distinction of the syntactical and morphological language levels. Such doubtful and other demanding examples will be thoroughly presented in the oral presentation.

Since the TEI module for digitizing grammar books does not exist, we had to use the tag set from Module 9: Dictionaries,[3] and adapt it to cover all the linguistic features of the Croatian morphological paradigms. The Croatian language, like all Slavic languages, belongs to the group of highly flective languages with a rich morphological system. So, for example, for each morphological form, the whole set of tags is combined and united with the main element <form>, while the morphological features are described by the element <gram> and different attributes. In the case of the example noun *vojnik* ('soldier'), these describe POS, type of the noun, gender, number, case, inflection type, and animacy.

```
<form type="inflectedForm" xml:lang="hr">
    <orth>vojnik</orth>
    <gramGrp>
      <gram type="pos" corresp="#imenica"/>
```

[3] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/examples-gram.html TEI Guidelines.

```
<gram type=”nounType” corresp=”#I_opca”/>
<gram type=”gender” corresp=”#muski”/>
<gram type=”number” corresp=”#jednina”/>
<gram type=”case” corresp=”#nominativ”/>
<gram type=”inflectionType” corresp=”#I_a_sklonidba”/>
<gram type=”animacy” corresp=”#I_zivo”/>
</gramGrp>
</form>
```

For each part of the speech, its own grammatical group set of tags is defined, which can be expanded by other attributes (@type) defining other features relevant to the morphological description (for verbal paradigms besides the attributes defining the verb type, tense, number, person, and voice, we can add attributes for the verbal aspect, valence, reflexive or nonreflexive verbs, etc.). In the attempt to adjust the Latin grammar model to Croatian, early grammarians introduced gerunds and participles into the grammatical description of the Croatian language. With gerund, they named verbal adverbs and participles verbal adjectives. For the participles, grammarians provided the whole paradigms, since they are declinable, so in their annotation with element <gram> and attributes @ type=gender, @type=number, @type=case, we annotated those specific features.

The TEI annotation of all existing examples from eight selected pre-standard Croatian grammar books will enable highly structured collection and extraction of morphological data. Those data will complete existing knowledge about the morphological development of the Croatian language and its normative description. This will lay the foundation for a reliable historical grammar of the Croatian language.

In a similar way, using TEI tags we will extract Croatian linguistic terms from all selected grammar books and create an index of pre-standard Croatian linguistic terminology.

References

Horvat, M., & Kramarić, M. (2021). Retro-digitization of Croatian Pre-standard Grammars. *Athens Journal of Philology*, *8*(4), 297–310.

Pierazzo, E. (2014). *Digital scholarly editing: Theories, models and methods*. Routledge. Retrieved from: https://hal.univ-grenoble-alpes.fr/hal-01182162/document

RETROGRAM. Retrieved from: https://retrogram.jezik.hr/

Robinson, P. (2002). What is a Critical Digital Edition? *Variants: The Journal of the European Society for Textual Scholarship*, *1*, 43–62.

Text Encoding Initiative. *TEI: Guidelines*. Retrieved from: https://tei-c.org/guidelines/

W3C. *Extensible Markup Language (XML)*. Retrieved from: https://www.w3.org/XML/

Accuracy of Slovak Language Lemmatization and MSD Tagging – MorphoDiTa and SpaCy

Radovan Garabík,
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Slovakia,
radovan.garabik@kassiopeia.juls.savba.sk
Denis Mitana,
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Slovakia,
denis.mitana@korpus.juls.savba.sk

Introduction

The Slovak language, as a "typical" Slavic language, belongs to the group of moderately inflected languages, with three or four genders, two grammatical numbers, all interacting with the inflections in somewhat complicated and unpredictable ways. The inflections are realized primarily by suffixes, but with many irregularities; one suffix encodes several relevant grammatical categories and the same suffix often reflects unrelated features in other words, a typical inflectional language not amenable to a heuristic analysis. Following these limitations, lemmatization is often an indispensable step in all kinds of text processing (starting with full-text search), and full morphosyntactic analysis or description (MSD) is the core of corpus linguistic research. Given the core importance of lemmatization and MSD in Slovak corpus linguistics, it is important to realize its limitations and recognize achievable accuracy. Since modern approaches aim to utilize deep learning and huge language models, we evaluate the accuracy of lemmatization + MSD in several common usage scenarios by comparing the state-of-the-art "classical" lemmatizer and MSD tagger *MorhoDiTa*, based on perceptron; and *spaCy*, using a multilingual BERT language model.

1. Dataset

Models were trained on a manually annotated corpus *r-mak-6.0* containing 720 documents, 77,671 sentences, 1,199,793 tokens, 55,090 unique lemmas, and 1,354 unique morphosyntactic tags. The composition of the corpus is 30.5% journalistic, 50.5% fiction, and 19.0% professional. To train the taggers, the corpus was divided randomly with stratification over documents into *train*, *dev*, and *test* segments in the ratio 8:1:1. The morphological database used for training MorphoDiTa contains 3,816,295 entries (i.e. distinct word-lemma-tag combinations), 114,634 unique lemmas, and 1,330,039 unique wordforms.

2. MorphoDiTa and SpaCy

MorphoDiTa [5] is a language-independent, open-source tool for morphological

analysis of natural language texts. MorphoDiTa has been extensively used in Slovak language corpora, and the tagger issued in the web interface provides access to lemmatization and tagging [4]. Part of MorphoDiTa is a statistical guesser for out-of-vocabulary (OOV) words, trained on suffixes. We use suffixes of at most 3 length, with 8 rules per suffix. To improve the accuracy of the guesser on real-world texts, we postprocess the guesser output and filter the list of possible lemmas to prefer tokens that appeared (as raw wordforms) in the corpora *prim-9.0-juls-sane* [1] and *Araneum Slovacum IV Maximum* [2]. We also include several heuristic rules to filter out implausible combinations of lemmas and tags and directly assign tags for numbers, punctuation, and other symbols. SpaCy is a language-independent, open-source, and production-ready Natural Language Processing (NLP) library. It comes with a wrapper package of the state-of-the-art Hugging Face's transformers architecture. For the Slovak language, spaCy officially supports only stop words and lexical attributes for general numerals. Although there is one online tool using spaCy available [6], it is not described further and is not available for unrestricted use. We use only MSD tagging and lemmatization components. As an architecture for the morphological analysis component, we use Transformer architecture based on the pre-trained multilingual BERT language model [3]. The lemmatization component is rule-based only. The rules applied are as follows: (1) if a given pair of word form and MSD tag is in the morphological database, then use the assigned lemma; (2) try to lemmatize a given pair of word form and MSD tag using the morphological suffix database. Postprocessing in the form of directly assigning tags for numbers, punctuation, and other symbols is the same as in MorphoDiTa.

3. Training and Evaluation

We summarise the accuracy of MorphoDiTa and spaCy output in the Table 1, where we take the *lemma+tag* accuracy to be the baseline. The "no OOV" row refers to MorphoDiTa accuracy calculated only on sentences containing only words known to the morphological database, thus describing a sort of "ideal" goal if the underlying morphological database has 100% coverage. As we can see, spaCy achieves higher accuracy in tagging. We suppose that more complex Transformer architecture handling a large number of output tags is better. On the other hand, spaCy is worse in lemmatization, apparently due to the rule-based approach. The most obvious and relevant single improvement of spaCy over MorphoDiTa is in disambiguating between singular masculine inanimate nominative and accusative (the word forms are identical), where apparently relatively free word order of Slovak requires better use of the context (or bigger context) to find out the correct case, where spaCy cuts the number of errors by two thirds compared to MorphoDiTa.

Conclusion

We calculated the accuracy of two state-of-the-art Slovak language lemmatizers and MSD taggers, one based on MorphoDiTa and the other one on spaCy. Over all, for

the combination of lemma+tag, spaCy with an accuracy of 95.6% overcame MorphoDiTa with 93.5%, with the most relevant single improvement in disambiguating between otherwise identical masculine inanimate singular nominative from the accusative, where better use of the context apparently helps to find out the correct case.

Table 1. Accuracy of various token annotations. CI means case insensitive.

Model name	Lemma	Lemma CI	MSD	POS	Lemma+tag	Lemma CI + tag
MorphoDiTa	98.24	98.95	94.19	98.06	93.50	94.03
spaCy	98.23	98.79	96.54	98.47	95.61	96.05
MorphoDiTa†	99.09	99.35	95.05	98.48	94.76	94.98

†no OOV

References

1. Slovenský národný korpus – prim-9.0-juls-sane. Bratislava: Jazykovedný ústav Ľ. Štúra SAV. https://korpus.juls.savba.sk (2020), accessed: 2022-03-07

2. Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček & K. Pala (Eds.), *Text, Speech and Dialogue. 17th International Conference* (pp. 257–264). Springer.

3. Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.

4. Garabík, R., & Bobeková, K. (2021). Lematizácia, morfologická anotácia a dezambiguácia slovenského textu – webové rozhranie. *Slovenská reč*, *86*(1), 104–109.

5. Straková, J., Straka, M., & Hajič, J. (2014). Open-source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 13–18). Association for Computational Linguistics.

6. Wencel, M. Spracovanie prirodzeného jazyka – SpaCy Web App. https://spacy.tukekemt.xyz/analyze_sk (2021), accessed: 2022-03-07

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 breaktrough 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