Uvodnik v tematsko številko o Digitalnem jezikoslovju

Darja FIŠER
Inštitut za novejšo zgodovino

Tomaž ERJAVEC
Institut Jožef Stefan

Pričujoča tematska številka revije *Slovenščina 2.0* se posveča digitalnemu jezikoslovju, hitro rastočemu interdisciplinarnemu področju raziskav na stičišču tradicionalnega jezikoslovja, informacijskih tehnologij in družboslovnih ved. V ospredju digitalnojezikoslovnih raziskav je ohranjanje, analiza in uporaba jezikovnih podatkov, digitalnih artefaktov z jezikom kot nosilcem medčloveškega sporazumevanja. Digitalno jezikoslovje tako pri nas kot po svetu postaja vse pomembnejše ne samo v akademskih in izobraževalnih krogih, temveč tudi v javnem in zasebnem sektorju, ki za uspešno delovanje v sodobni družbi in gospodarstvu vse bolj potrebujeta strokovnjake, vešče upravljanja z digitalnimi jezikovnimi podatki.

Tematska številka vsebuje enajst prispevkov slovenskih in tujih avtorjev, ki so bili v krajši in bolj omejeni obliki najprej predstavljeni na konferenci "Jezikovne tehnologije in digitalna humanistika" leta 2022. Prvi sklop prinaša pester nabor raziskovalnih vprašanj z različnih znanstvenih ved, od jezikoslovja in prevodoslovja pa vse do literarnih ved in zgodovinopisja, ki se jih avtorji lotevajo s korpusnim pristopom, drugi pa združuje prispevke, ki predstavljajo gradnjo eno- in večjezičnih jezikovnih virov in tehnologij za različne naloge.

Fišer, D., Erjavec, T.: Uvodnik v tematsko številko o Digitalnem jezikoslovju/ Introduction to the special issue on Digital Linguistics. Slovenščina 2.0, 11(1): 1–6.

1.20 Uvodnik / Editorial

DOI: https://doi.org/10.4312/slo2.0.2023.1.1-6

V korpusnojezikoslovnem sklopu Špela Arhar Holdt, Iztok Kosem, Eva Pori, Vojko Gorjanc, Simon Krek in Polona Gantar predstavijo inovativne rešitve za prepoznavanje in označevanje sovražnega in grobega besedišča v okviru koncepta odzivnega Slovarja sopomenk sodobne slovenščine. Špela Antloga opiše najpogostejše metode luščenja metaforičnih in metonimičnih izrazov iz jezikovnih korpusov ter na primeru korpusa g-KOMET, ki je ročno označen za metaforične izraze in metonimične prenose, ponazori poskus sistematizacije nekaterih najbolj prisotnih metonimičnih prenosov v slovenskem govorjenem jeziku. Jakob Lenardič in Kristina Pahor de Maiti proučita skladenjsko in pragmatično rabo epistemičnih in deontičnih modalnih izrazov v korpusu slovenskih družbeno sprejemljivih in nesprejemljivih komentarjev na družbenem omrežju Facebook. David Bordon poroča o raziskavi preverjanja razumljivosti nerevidiranih strojno prevedenih spletnih besedil med splošni bralci. Darja Fišer, Tjaša Konovšek in Andrej Pančur analizirajo značilnosti populističnega govora v parlamentarnih razpravah slovenskih poslancev. Andrejka Žejn in Mojca Šorli pa predstavita ročno semantično označevanje imenskih entitet glede na predlagano označevalno shemo, izdelano za korpus modernističnih literarnih besedil Maj68.

V jezikovnotehnološkem sklopu Iztok Kosem, Jaka Čibej, Kaja Dobrovoljc, Taja Kuzman in Nikola Ljubešić opišejo prvi spremljevalni korpus za slovenščino Trendi ter razvoj, evalvacijo in aplikacijo algoritma za avtomatsko kategorizacijo besedil z novičarskih portalov. Kaja Dobrovoljc, Luka Terčon in Nikola Ljubešić predstavijo nove smernice, odprto dostopne ročno označene podatke ter razčlenjevalni model za označevalnik CLASSLA Staza v formalizmu Universal Dependencies za slovenščino. Petra Bago in Virna Karlić poročata o novi različici odprto dostopnega korpusa DirKorp (Korpus direktivnih govornih činova hrvatskoga jezika), namenjenega raziskavam v pragmatiki, ki vsebuje simulirano komunikacijo, implementirano v vnaprej danih pogojih stotih hrvaških govorcev. Uroš Šmajdek Matjaž Zupanič, Maj Zirkelbach in Meta Jazbinšek opišejo ter evalvirajo sistem za odgovarjanje na vprašanja v slovenskem jeziku. Gregor Donaj in Mirjam Sepesy Maučec pa predstavita uporabo podbesednih enot za nevronsko strojno prevajanje iz slovenščine v angleščino.

Posebno številko so recenzirali Zoran Bosnić, Václav Cvrček, Jaka Čibej, Helena Dobrovoljc, Kaja Dobrovoljc, Polona Gantar, Vojko

Gorjanc, Jurij Hadalin, Matej Klemen, Jakob Lenardič, Nikola Ljubešić, Matija Marolt, Maja Miličević Petrović, Matija Ogrin, Matevž Pesek, Dan Podjed, Tanja Samardžić, Marko Robnik Šikonja, Mojca Šorn, Simon Šuster, Daniel Vasić in Aleš Žagar.

Urednika posebne številke se iskreno zahvaljujeva avtorjem in recenzentom za njihovo predano delo.

V primerjavi s sorodno tematsko številko na temo jezikovnih tehnologij in digitalne humanistike iz leta 2021, kjer je bil poudarek na uvajanju naprednih tehnik in metod strojnega učenja, večjezikovnim pristopom, kritičnemu ocenjevanju obstoječih tehnologij ter razvoju storitev za končnega uporabnika, v pričujoči številki opazimo pomembno širitev korpusnega pristopa na vede izven jezikoslovja, ki pri svojem delu prav tako izhajajo iz pretežno besedilnih podatkov, ter razvoj vse bolj specializiranih jezikovnih virov in tehnologij. Oba premika nakazujeta na dozorevanje in razmah področja v Sloveniji v zadnjih nekaj letih. Z vstopom v obdobje, v katerem se metode umetne inteligence povsod po svetu intenzivno uveljavljajo v praktično vseh znanstvenih disciplinah, pa tudi v našem vsakdanjem življenju, pa bo pomen zanesljivih in visoko kakovostnih jezikovnih virov, preverljivih jezikovnih tehnologij ter čezdisciplinski prenos metodološkega znanja samo še naraščal, zato je ključno, da skupnosti zagotovimo ustrezno raziskovalnoinfrastrukturno podporo in izobraževalni okvir.

Ljubljana, julij 2023

Introduction to the special issue on Digital Linguistics

Darja FIŠER
Institute of Contemporary History

Tomaž ERJAVEC
Jožef Stefan Institute

The current special issue of the journal Slovenščina 2.0 focuses on Digital Linguistics, a growing interdisciplinary field at the crossroads of traditional linguistics, information technology and social sciences. Digital Linguistics preserves, analyses and utilises language data, i.e. digital artefacts that use language as a means of human expression. In Slovenia as well as abroad, Digital Linguistics is attracting increasing attention not only from the academic and educational communities but also from the public and private sectors, since skills in handling digital language data are considered essential in the modern economy and society.

The special issue presents eleven papers of Slovenian and international authors that were originally presented, in a shorter and more limited form, at the 2022 Language Technologies and Digital Humanities Conference. Part 1 contains a broad range of research questions from different scientific disciplines such as linguistics and translation studies but also literary studies and historiography which are approached within the corpus linguistics framework. Part 2 comprises papers that present the development of mono- and multilingual language resources and technologies for a variety of tasks.

The corpus-linguistic section opens with the paper by Špela Arhar Holdt, Iztok Kosem, Eva Pori, Vojko Gorjanc, Simon Krek and Polona Gantar who introduce innovative solutions for the recognition and mark-up of hate and offensive lexis in the framework of the Reactive Dictionary of Contemporary Slovenian Synonyms. Špela Antloga then describes the most common methods for extracting metaphorical and

metonymic expressions from language corpora and, on the case of the g-KOMET corpus, which is manually labelled for metaphoric expressions and metonymic transfers, exemplifies an attempt to systemise some of the most common metonymic transfers in the Slovenian spoken language. Jakob Lenardič and Kristina Pahor de Maiti investigate grammatical and pragmatic usage of epistemic and deontic modal expressions in the corpus of acceptable and unacceptable comments in Slovenian Facebook comments. David Bordon gives an account of an investigation to assess the comprehensibility of un-edited machine translated Web texts among general users. Darja Fišer, Tjaša Konovšek and Andrej Pančur explore the characteristics of populist discourse in parliamentary speeches of Slovenian members of parliament. Andrejka Žejn and Mojca Šorli introduce manual semantic labelling of named entities in accordance with the annotation scheme developed for the corpus of Slovenian modernist literary texts May68.

The language-technology section starts with a description of Trendi, the first monitor corpus for Slovenian by Iztok Kosem, Jaka Čibej, Kaja Dobrovoljc, Taja Kuzman and Nikola Ljubešić who also describe the development, evaluation and application of their method for the automatic categorisation of texts from news portals. Kaja Dobrovoljc, Luka Terčon and Nikola Ljubešić introduce the new guidelines, openly available manually annotated datasets and the analysis model for the CLASSLA Stanza annotation tool, all for the Universal Dependencies formalism for Slovenian. Petra Bago and Virna Karlić report on the new version of the openly available DirKorp corpus, meant for research on pragmatics, which contains simulated communication implemented in pre-set conditions by one hundred Croatian speakers. Uroš Šmajdek Matjaž Zupanič, Maj Zirkelbach and Meta Jazbinšek describe and evaluate their system for question answering in Slovenian. Finally, Gregor Donaj and Mirjam Sepesy Maučec discuss and evaluate the use of sub-word units for neural machine translation from Slovenian to English.

The special issue was reviewed by Zoran Bosnić, Václav Cvrček, Jaka Čibej, Helena Dobrovoljc. Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Jurij Hadalin, Matej Klemen, Jakob Lenardič, Nikola Ljubešić, Matija Marolt, Maja Miličević Petrović, Matija Ogrin, Matevž Pesek, Dan Podjed, Tanja Samardžić, Marko Robnik Šikonja, Mojca Šorn, Simon

Šuster, Daniel Vasić in Aleš Žagar. The editors of the special issue would like to thank the authors and the reviewers for their dedicated work.

Compared to the related special issue on Language Technologies and Digital Humanities published in this journal in 2021 where the focus of research was on the implementation of state-of-the-art machine learning methods, multilingual approaches, critical evaluation of technologies, and development of services for the end user, we observe an important spread of the corpus approach to disciplines beyond linguistics which are also based on predominantly textual data, as well as the emergence of increasingly specialised language resources and technologies. This suggests that in Slovenia the field has matured and advanced significantly in the last couple of years. In the time when Artificial Intelligence is playing an increasingly important role in the methodological approaches in virtually all scientific disciplines, as well as our everyday lives all over the world, the importance of reliable and high quality language resources, verifiable language technologies and transdisciplinary knowledge transfer will only increase, which is why it is crucial that our community is equipped with an adequate research infrastructure support and a robust educational framework.

Ljubljana, July 2023