Reviews and Reports

Language Technologies and Digital Humanities 2018, 20–21 September 2018, Faculty of Electrical Engineering, Ljubljana

The conference *Language Technologies and Digital Humanities 2018* took place at the Faculty of Electrical Engineering at the University of Ljubljana on 20 and 21 September 2018. It was organised by the *Slovenian Language Technologies Society,*[1] the *Centre for Language Resources,*[2] the *Faculty of Electrical Engineering,*[3] and the research infrastructures *CLARIN.SI*[4] and *DARIAH.SI.*[5] The conference was the eleventh iteration – as well as the 20th anniversary – of the *Language Technologies* conference series,[6] which was started by the *Slovenian Language Technologies Society* and has been taking place biennually since 1998. In 2016 it successfully expanded its scope to include Digital Humanities as well. The 2018 edition of the conference was very international, with authors from 17 European countries[7] as well as two participants from Brazil and Japan. This is why the conference programme was organized in such a way that talks on Day 1 were in English and on Day 2 in Slovene.

The conference was opened by the first keynote speaker Malvina Nissim, who is Associate Professor of Computational Linguistics and Natural Language Processing at the University of Groningen. In her talk, titled "Too good to be true: Current Approaches to author profiling", she discussed novel approaches to the automatic identification of the gender and age of social media users. In particular, she showed that models which abstract away from the lexical content of social media posts and instead focus on extra-linguistic information such as punctuation and emoticons, whose use is shared across languages to a great extent, offer a robust and reliable way to identify such personal information.

1 *SDJT – Slovensko društvo za jezikovne tehnologije*, http://www.sdjt.si/wp/english/.
2 *CJVT – Centre for language resources and technologies*, https://www.cjvt.si/en/.
3 *Univerza v Ljubljani, Fakulteta za elektrotehniko*, http://www.fe.uni-lj.si/en/.
4 *CLARIN Slovenia*, http://www.clarin.si/info/about/.
5 *Dariah-SI | Digitalna humanistika*, http://www.dariah.si/en/.
6 *SDJT – Slovensko društvo za jezikovne tehnologije*, http://www.sdjt.si/wp/dogodki/konference/strani/.
7 In addition to Slovenia, the following European countries were represented: Austria, Belgium, Bulgaria, Denmark, Finland, Germany, Greece, Ireland, the Netherlands, Norway, Poland, Portugal, Serbia, Spain, Sweden, and Switzerland.

The keynote talk was followed by two morning sessions devoted to topics in machine translation and language resources. The machine translation session was chaired by Tomaž Erjavec and comprised two talks. Gregor Donaj and Mirjam S. Maučec compared traditional statistical machine translation with the use of neural networks for translating between Slovenian and English, while Mihael Arčan compared the two approaches by using translations between three Slavic languages – Slovenian, Croatian and Serbian.

In the subsequent session devoted to language resources, which was chaired by Simon Krek, six papers were presented, introducing on-going work on language corpora and lexical resources in Slovenian, Croatian and Portuguese. For example, Filip Dobranić presented joint work with Nikola Ljubešić, Darja Fišer and Tomaž Erjavec on the creation of the *Parlameter* corpus, which contains contemporary Slovenian parliamentary proceedings from 2014 to 2018 with rich speaker metadata on the gender, age, education and party affiliation of the members of the Slovenian parliament. Filip also showcased how the resource facilitates in-depth exploration of institutionalised language use and interpersonal behaviour patterns, which is important for an interdisciplinary approach to the analysis of parliamentary discourse that involves collaboration between researchers working in disciplines like sociology, discourse analysis, history, sociolinguistics, and political science.

The poster session presented nine posters on various applications of quantitative approaches to data analysis within digital humanities and social sciences. For instance, Katja Mihurko Poniž and colleagues introduced a tool that aids in the research of the historical representation of women's authorship, which is an important topic in sociohistoric approaches to literary theory, while Damjan Popič and Darja Fišer presented a corpus-driven analysis of the attitudes toward language in Slovenian, Croatian, and Serbian computer-mediated communication.

The first afternoon session, which was chaired by Jurij Hadalin, was devoted to Digital Humanities. Dan Podjed and Ajda Pretnar analysed the use of social media by the Slovenian President Borut Pahor for self-promotion. On the basis of qualitative and quantitative approaches to data analysis, they identified three distinct categories of the President's Instagram posts that prove to be the most popular among his followers; namely, (i) photographs in which he is seen together with celebrities and his family, (ii) posts in which he gives the impression of being approachable, and (iii) photographs in which he is depicted in an unusual situation. Tobias Weber and Jeremy Bradler discussed a novel approach of integrating computational methods, digital resources and computer literacy skills into the curriculum of Finno-Ugric linguistics, stressing the importance of tailoring the materials to the students' non-computational backgrounds in humanities and social sciences.

The subsequent session, which was chaired by Simon Dobrišek, concluded the first day of the conference. Nine papers were presented on topics related to language technologies and their application. For instance, in a cross-disciplinary approach to phonetics and medicine, Tatjana Marvin introduced joint work with Jure Derganc,

Samo Beguš and Saba Battelino on a novel Slovenian Sentence Matrix Text for measuring speech intelligibility in patients suffering from hearing loss. To give another example in a different field of application, Milan van Lange and Ralf Futselaar presented their use of word embeddings in the analysis of parliamentary debates on war criminals in The Netherlands.

The second day of the conference began with a keynote talk delivered by Martijn Kleppe, who is Head of the Research Department at the National Library of the Netherlands. In his talk, titled "Bringing Digital Humanities to the wider public: libraries as incubator for DH research results", Martijn presented one of the main aims of the National Library of the Netherlands, which is to support researchers in the Digital Humanities and social sciences and incorporate their research results in its services and products. To this end, Martijn showcased *LAB,* and online toolchain of the Library which offers researchers an interoperative environment for working with richly annotated texts and state-of-the-art tools for processing handwritten documents. He also discussed the Institute's collaborations with other national and international research infrastructures, such as CLARIN ERIC.

The next session, chaired by Andrej Pančur, brought two talks on issues related to Slovenian research infrastructures. Maja Dolinar, Janez Štebe and Sonja Bezjak, presented a new set of guidelines for the acquisition and archiving of qualitative research in the Slovenian Social Science Data Archives. Tomaž Erjavec presented joint work with Darja Fišer and Jakob Lenardič on how linguistic data, such as those found in language corpora, are cited in Slovenian research publications, and proposed recommendations and solutions for more consistent and rigorous citation practices in line with the Austin Principles of Data Citation in Linguistics.

In the next session, chaired by Darja Fišer, six talks were given on topics related to corpus linguistics. For instance, Nataša Logar presented the main morphosyntactic characteristics of academic Slovenian, which she analysed together with Tomaž Erjavec on the basis of the Slovenian balanced corpus Kres and the corpus KAS, which consists of Slovenian BA, BSc and PhD theses. Iztok Kosem presented joint work with Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, and Cyprian Laskowski on the user interface of the first Collocations Dictionary for Modern Slovenian, which was compiled on the basis of state-of-the-art lexicographic methods.

The student session, which was chaired by Iza Škrjanec, included four talks. Urška Bratoš discussed the compilation and analysis of a corpus of tweets written by Slovenian politicians. Isolde van Dorst then presented a statistical analysis of Shakespeare's use of pronominal expressions, specifically his usage of the second-person pronoun *you* and its two now-obsolete informal variants – nominative *thou* and accusative *thee*. Gabi Rolih presented an implementation of a K-means clustering method applied to computer-mediated communication and discussed how it can be used to further improve a state-of-the-art part-of-speech tagger for Slovenian. Finally, Klara Eva Kukovičič compared the concordancer Sketch Engine with the tool CollTerm from the point of view of terminology extraction. The Best Student Paper Award was awarded to Isolde van

Dorst by the selection committee Iza Škrjanec (chair of the Student Session), Tomaž Erjavec (on behalf of the Programme Committee) and Kaja Dobrovoljc (on behalf of the *Slovenian Language Technologies Society*).

The final session, chaired by Matija Ogrin, focused again on Digital Humanities and concluded the conference. Five papers were presented on topics related to cultural heritage, historical studies and geography. For instance, Andrej Pančur presented the SIstory web portal, which offers a sustainable repository for digital editions of historical texts, while Alenka Kavčič presented joint work with Ivan Lovrić and Vera Smole on the development of an interactive online map of the seven major Slovenian dialect groups, which includes geocoded text examples enriched with audio materials that exemplify the salient phonological features of the dialects.

The *Language Technologies and Digital Humanities 2018* conference successfully presented on-going and completed work on state-of-the-art language tools and resources, as well as their application. The presentations that used computational tools and methodologies to answer qualitative research questions were especially illustrative in showing how language technologies facilitate and open new grounds for research in fields like translation studies, political science, historical studies, phonetics and phonology, and literary theory. Perhaps crucially, the work presented by master's and doctoral students was an inspiring showcase of how young researchers use innovative computational approaches to tackle complex research problems in such interdisciplinary fields. The conference thus gave both novice and experienced researchers from Slovenia and abroad a chance to strike up collaborations and get involved in research projects that bridge the gap between language technologies on the one hand and humanities and social sciences on the other.

Jakob Lenardič*

* **Department for Translation, Faculty of Arts, University of Ljubljana, Aškerčeva 2, SI-1000 Ljubljana, jakob.lenardic@ff.uni-lj.si**

Sources and Literature

- Arčan, Mihael. 2018. "A comparison of Statistical and Neural Machine Translation for Slovene, Serbian and Croatian." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 3–10. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Bratoš, Urška. 2018. "Gradnja korpusa tvitov slovenskih politikov Janes-TwePo." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 269–73. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Dolinar, Maja, Janez Štebe, and Sonja Bezjak. 2018. "Razvoj smernic za predajo in arhiviranje kvalitativnih podatkov v Arhivu družboslovnih podatkov." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 55–61. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Donaj, Gregor, and Mirjam S. Maučec. 2018. "Prehod iz statističnega strojnega prevajanja na prevajanje z nevronskimi omrežji za jezikovni par slovenščina-angleščina." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 62–68. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- van Dorst, Isolde. 2018. "You, Thou and Thee: A Statistical Analysis of Shakespeare's Use of Pronominal Address Terms," In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 274–80. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Fišer, Darja, Jakob Lenardič, and Tomaž Erjavec. 2018. "Citiranje jezikoslovnih podatkov v slovenskih znanstvenih objavah: stanje in priporočila." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 77–84. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Kavčič, Alenka, Ivan Lovrić, and Vera Smole. 2018. "Karta slovenskih narečnih besedil." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 121–25. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Kleppe, Martijn. 2018. "Bringing Digital Humanities to the Wider Public: Libraries as Incubator for DH Research Results." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 2. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Kosem, Iztok, Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, and Cyprian Laskowski. 2018. "Kolokacijski slovar sodobne slovenščine," In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 133–39. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Kukovičič, Klara Eva. 2018. "Uporabnost luščilnikov terminologije Sketch Engine in CollTerm z vidika (študenta) prevajalca." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 281–87. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- van Lange, Milan, and Ralf Futselaar. 2018. "Debating Evil: Using Word Embeddings to Analyze Parliamentary Debates on War Criminals in The Netherlands." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 147–53. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, and Filip Dobranić. 2018. "The Parlameter corpus of contemporary Slovene parliamentary proceedings." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 162–67. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Logar, Nataša, and Tomaž Erjavec. 2018. "Strokovnoznanstvena slovenščina: besednovrstne in oblikoskladenjske značilnosti." In *Proceedings of the Conference on Language Technologies & Digital*

Humanities 2018, edited by Darja Fišer and Andrej Pančur, 175–80. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.

- Marvin, Tatjana, Jure Derganc, Samo Beguš, and Saba Battelino. 2018. "Word Selection in the Slovenian Sentence Matrix Test for Speech Audiometry." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 181–87. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Mihurko Poniž, Katja, Amelia Sanz, Marie Nedregotten Sørbø, Suzan van Dijk, Viola Parente-Čapková, Narvika Bovcon, and Aleš Vaupotič. 2018. "Teaching Women Writers with NEWW Virtual Research Environment." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 254–55. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Nissim, Malvina. 2018. "Too Good to Be True: Current Approaches to Author profiling." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 1. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Pančur, Andrej. 2018. "Trajnost digitalnih izdaj: Uporaba statističnih spletnih strani na portal Zgodovina Slovenije – SIstory." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 203–10. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Podjed, Dan, and Ajda Pretnar. 2018. "Samopromocija na Instagramu: Primer predsednikovega profila." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 221–26. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Popič, Damjan, and Darja Fišer. 2018. "Odnosi do jezika v slovenski, hrvaški in srbski računalniško posredovani komunikaciji." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 256–59. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Rolih, Gabi. 2018. "K-means Clustering of CMC Data for Tagger Improvement." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 288–91. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.
- Weber, Tobias, and Jeremy Bradley. 2018. "Exploring Finno-Ugric Linguistics Through Solving IT Problems." In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018*, edited by Darja Fišer and Andrej Pančur, 248–53. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.