Language Technologies and Digital Humanities 2020

The Conference on Language Technologies and Digital Humanities, organised biennially by the Slovenian Society for Language Technologies (SDJT)[1] in cooperation with the Institute of Contemporary History,[2] the Centre for Language Resources and Technologies of the University of Ljubljana (CJVT),[3] and the research infrastructures CLARIN.SI[4] and DARIAH-SI,[5] took place on 24 and 25 September. This was the twelfth iteration of the conference, which boasts more a tradition spanning more than twenty years. The conference has become an important link between the field of language technologies and digital humanities, as this years' conference was also the third multidisciplinary iteration since the conference programme was extended to include the field of digital humanities in 2016.

This years' conference was initially set to take place at the Institute of Contemporary History in Ljubljana but was moved to the virtual environment because of the Covid-19 pandemic. The shift from the traditional to the virtual environment influenced the conference structure and organisation as well as presented new challenges. Typically, the conference would take place over the span of two days. However, unlike the previous years, days and sessions were not divided by languages in which the papers were presented. Instead, they were divided solely based on the themes of the papers. In this year's implementation, the individual student contributions were placed within the thematically relevant sessions and not in a separate student panel. In the course of the Covid-19 pandemic, our way of life, work, and communication have changed drastically, forcing us to transfer our presentations and informative discussions to a virtual environment. Therefore, we asked the authors of the individual papers to pre-record their presentations, which were made public a few days before the start of the conference. This enabled us to focus on the dynamic flow of discourse, as the participants were able to view the recordings and proceedings before the conference, while the authors only presented brief summaries of their lectures, followed by discussions.

The conference was opened by Sara Tonelli, the head of the Digital Humanities research group at the Bruno Kessler Foundation in Trento and associate professor at the Department of Psychology and Cognitive Sciences at the University of Trento. In her invited lecture, titled *Abusive Language Detection: Too Much Digital, Not Enough Humanities?* the lecturer created an overview of the modern approaches to recognising violent speech, emphasised the role as well as comprehension of the language used by various online communities, and presented the current research to understand this

1 *Slovenian Society for Language Technologies (SDJT)*, accessed on 23 November 2020, http://www.sdjt.si/wp/.

2 *Institute of Contemporary History*, accessed on 23 November 2020, https://www.inz.si/.

3 *Center for Language Resources and Technologies of the University of Ljubljana*, accessed on 23 November 2020, https://www.cjvt.si/en/.

4 *Common Language Resources and Technology Infrastructure, Slovenia*, accessed on 23 November 2020, https://www.clarin.si/info/about/.

5 *DARIAH-SI |Digitalna humanistika*, http://www.dariah.si/en/.

phenomenon better.[6] The discussion that followed the lecture was moderated by Filip Dobranić.

The invited lecture was followed by the first session on the topic of language and speech technologies, moderated by Tanja Samardžić. Anka Supej, Matej Ulčar, Marko Robnik Šikonja and Senja Pollak presented their work in which they compared the sexual bias of models (or their embedding) with different configurations and approaches to computing analogies.[7] During the presentations of extended abstracts, Darinka Verdonik presented the development and operation of the research infrastructure of the RI-SI CLARIN project on behalf of the co-authors. In his student paper, Andraž Pelicon presented the perception of sentiment in the news using deep neural networks.[8]

The next section, which dealt with language resources, was moderated by Darinka Verdonik and included presentations of four full contributions, three extended papers, and a student paper. Simon Krek (with co-authors) presented the latest development and achievements of the ssj500k learning corpus, the largest and most frequently used open-source database for Slovenian language processing.[9] Dolores Lemmenmeier-Batinić highlighted the lack of publicly available resources for the Serbian language and presented the methods for converting corpus data into the standardised XML format.[10] Finally, Špela Antloga concluded the session by presenting her student paper on the methodological starting points, development, and guidelines for marking metaphorical words in the KOMET 1.0 metaphor corpus.[11] The first day of the conference was concluded with a special panel of the RSDO project (*Development of Slovene in the Digital Environment*), during which the leaders of the individual work packages presented the primary and intermediate goals to be performed by the end of the project.

The introduction to the second day of the conference, which set the tone for the day, was offered by the digital historian Kaspar Beelen of the Alan Turing Institute, who explores the use of machine learning in humanities research. In his invited lecture, titled *Speaking on Behalf of Others: Why the Digital Humanities Should Care about*

6 Sara Tonelli, "Abusive Language Detection: Too Much Digital, Not Enough Humanities?," in *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 1.

7 Anka Supej, Matej Ulčar, Marko Robnik-Šikonja, and Sanja Pollak, "Primerjava slovenskih besednih vektorskih vložitev z vidika spola na analogijah poklicev," in *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 93–100.

8 Andraž Pelicon, "Zaznavanje sentimenta v novicah z globokimi nevronskimi mrežami," in *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 150–57.

9 Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, and Janez Brank, "The ssj500k Training Corpus for Slovene Language Processing," in *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 24–33.

10 Dolores Lemmenmeier-Batinić, Nikola Ljubešić, and Tanja Samardžić, "XML-Encoding of a Spoken Serbian Corpus Targeting Forms of Address", in *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 127–30.

11 Špela Antloga, "Korpus metafor KOMET 1.0," in *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 167–70.

Parliamentary Data, he highlighted the importance and role of parliamentary data, which provide insight into the language and worldview of the MPs and the voters they are supposed to represent, thus offering detailed testimony on almost every topic that has ever been the subject of public debate.

In the third session, moderated by Kristina Štrkalj Despot, contributions in the field of corpus analysis were presented. Kristina Pahor de Maiti presented a study of the morphosyntactic characteristics of comments on Facebook to identify those that are commonly seen in socially unacceptable discourse (SUD).[12] Jakob Lenardič and Darja Fišer analysed epistemic modal adverbs in the 100-million-token corpus of Slovenian doctoral dissertations (the KAS corpus).[13] The student contributions in the context of the relevant session were presented by Zoran Fijavž and Eva Trivunović. In his paper, Zoran Fijavž explored the impact of video content on the presence of socially unacceptable discourse. He detected this from a set of comments related to LGBT communities on Facebook, originating from the leading news sources in Croatia.[14] Eva Trivunović concluded the session on corpus analysis with her presentation on the topic of biblical phrases, their variants, as well as the renewal and non-renewal modifications in the Gigafida 2.0, Janes, and slWaC corpora.[15]

The final session, which combined digital humanities and pedagogy (moderated by Miran Hladnik) was one of the more discussion-oriented sessions, as presenters were able to either focus solely on addressing the questions from the public or to give their opinion on a set of general discussion points that the moderator had provided in advance. The questions touched upon the problem of wiki-sourced tools and their lack of usage; usage of proprietary tools such as Zoom versus open-source ones; as well as discussed virtual communication, which has become the new normal. In the span of this particular session, several presentations were made. Katja Meden and Ana Cvek presented a technical upgrade of the Historiography Citation Index for the systematic listing of cited works in the field of historiography.[16] Andrej Pančur then presented the latest acquisition in the field of victimological demographic research; the digital database of (military) victims of the World War I from the territory of the today's Republic of Slovenia, created based on the cooperation between various research

12 Kristina Pahor de Maiti, Darja Fišer, Nikola Ljubešić, and Tomaž Erjavec, "Grammatical Footprint of Socially Unacceptable Facebook Comments," in *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 48–57.

13 Jakob Lenardič and Darja Fišer, "Epistemic Modal Adverbs in Slovenain Academic Discourse", in *Proceedings of the conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 34–41.

14 Zoran Fijavž, "Ambivalence of Queer Visibility in Video-Based Social Media Content", in *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 144–49.

15 Eva Trivunović, "Variante in modifikacije (iz)biblijskih frazemov," in *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 158–66.

16 Katja Meden and Ana Cvek, "Nadgradnja Zgodovinarskega indeksa citiranosti," in *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 42–47.

and cultural institutions as well as several individuals.[17] Finally, in her student paper, Magdalena Schlintl presented working with digital tools in teacher education in the case of teaching-learning-laboratory.[18]

The conference concluded with the award for the best student contribution, received by Zoran Fijavž for his paper titled *Ambivalence of Queer Visibility and Video-Based Social Media Content*. As there were quite a few outstanding speakers among the pre-recorded presentations, during the conference, the programme committee decided to also give an audience award for the best recorded presentation. It went to Špela Antloga for the presentation of her article *Corpus of Metaphors KOMET 1.0.*

Despite the technical and organisational challenges posed by the pandemic and the virtual environment, the 2020 Conference on Language Technology and Digital Humanities was carried out successfully. Thanks to the virtual environment, we were able to make all the presentations and discussion recordings available on the conference website, along with the PDF versions of the original papers. The presentations at the conference provided extensive insight into the new methods, applications, upgrades, and the development of research fields. The discussions that those presentations sparked gave us an incentive to move our research forward and offered an insight into the interdisciplinarity of the various thematic subfields of language technologies and digital humanities. Simply put, the conference allowed us to work more closely with the related fields and establish new building blocks in the efforts to bridge the gap between language technologies and digital humanities.

Katja Meden

17 Andrej Pančur, Neja Blaj Hribar, Mihael Ojsteršek, and Mojca Šorn, "Projekt Vojaške žrtve prve svetovne vojne na Slovenskem," in *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 136–40.

18 Magdalena Schlintl, Kerstin Pawluch, Mara Rader, and Verena Novak-Geiger, "Working with Digital Devices in Teacher Training Using the Example of the Teaching-Learning-Lab," in *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Tomaž Erjavec (Ljubljana: Inštitut za novejšo zgodovino, 2020), 171–74.