Ocene in poročila – Reviews and Reports

Konferenca jezikovne tehnologije in digitalna humanistika 2024

19. in 20. septembra je potekala že štirinajsta konferenca Jezikovne tehnologije in digitalna humanistika, ki jo vsaki dve leti organizira Slovensko društvo za jezikovne tehnologije (SDJT) v sodelovanju s Centrom za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT) ter raziskovalnima infrastrukturama CLARIN.SI in DARIAH-SI. Konferenca, ki ima že več kot dvajsetletno tradicijo, je postala pomembna vez med področjem jezikovnih tehnologij in digitalno humanistiko ter je tudi letos – od razširitve programa konference na področje digitalne humanistike leta 2016 – predstavljala multidisciplinarni dogodek.

Poleg osrednjega dela so v sredo, 18. septembra, v okviru konference JTDH 2024 potekali tudi predkonferenčni seminarji. Prva delavnica je bila *CLASSLA-Express* – iz serije delavnic, na katerih udeleženci raziskujejo korpuse južnoslovanskih jezikov z uporabo konkordančnikov CLARIN.SI. Organizirali in izvedli so jo Ivana Filipović Petrović, Jelena Parizoska, Petya Osenova, Nikola Ljubešić ter Taja Kuzman.

Drugo delavnico – *Brez nočnih mor zaradi urejanja dokumentov: uvod v LaTeX za humaniste* – sta organizirala in vodila Jakob Lenardič in Kristina Pahor de Maiti Tekavčič. Po delavnicah sta potekala še *okrogla miza o velikih jezikovnih modelih v korpusnem jezikoslovju ter mreženje južnoslovanskih raziskovalcev in centrov ReLDI in CLASSLA*.

Letošnja konferenca se je odvijala na Fakulteti za elektrotehniko Univerze v Ljubljani. V dveh dneh so prispevke predstavili vabljeni predavatelji in avtorji sprejetih prispevkov, ker pa je bila udeležba mednarodna, je bil program razdeljen na sekcije v slovenskem in angleškem jeziku.

Prvi dan je po uvodnih govorih konferenco otvoril vabljeni predavatelj Simon Dobnik, ki je predstavil prispevek z naslovom *Beyond pixels and words*. Po njegovi predstavitvi je potekala prva sekcija z naslovom *Speech and UGC resources*, ki se je odvijala v angleškem jeziku. Na njej sta prispevek o korpusu z več kot 170 milijoni objav na Twitterju v slovenskem, hrvaškem, bosanskem, srbskem in črnogorskem jeziku, zbranih med letoma 2017 in 2023, predstavila Filip Dobranić in Nikola Ljubešić. Kristina Pahor de Maiti Tekavčič, Nikola Ljubešić in Darja Fišer so predstavili oblikovanje francoskega dela korpusa FRENK, ki vsebuje družbeno nesprejemljive komentarje, objavljene kot odziv na novice o temah LGBT in migrantov, ki so jih na

Facebooku objavili znani mediji. Nikola Ljubešić, Peter Rupnik in Tea Perinčić so na sekciji govorili o prizadevanjih pri izdaji tiskane in zvočne knjige – prevoda slavnega romana *Mali princ* v čakavsko narečje kot računalniško berljivega, za umetno inteligenco pripravljenega nabora podatkov, pri čemer sta besedilna in zvočna sestavina obeh izdaj zdaj usklajeni na ravni vsake pisne in govorjene besede. Kaja Dobrovoljc je govorila o novi različici Spoken Slovenian Treebank (SST), ki je uravnotežena in reprezentativna zbirka transkribiranega spontanega govora z ročno anotiranimi lemami, oznakami delov govora, morfološkimi značilnostmi in skladenjskimi odvisnostmi. Sekcijo so zaključili Tanja Samardžić, Peter Rupnik, Mirjana Starović in Nikola Ljubešić s prispevkom o novem naboru podatkov, namenjenem reševanju problemov, ki jih predstavljajo objektivni primerjalni modeli.

Prvi dan je bil posvečen tudi predstavitvi plakatov. V sekciji se je predstavilo enajst plakatov, od tega šest v angleškem in pet v slovenskem jeziku. Generativno umetno inteligenco za konceptualizacijo računalniške ustvarjalnosti so na plakatu predstavili Boshko Koloski, Senja Pollak, Geraint Wiggins in Nada Lavrač. Ksenija Bogetić, Vojko Gorjanc, Jure Skubic in Alenka Kavčič so govorili o korpusno-lingvističnem pogledu na novo nastajajoče »proti-spolno« besedišče v Sloveniji, na Hrvaškem in v Srbiji. Platformo za transkripcijo govora GOVORI.SI so predstavili Klara Žnideršič, Vid Klopčič, Matevž Pesek in Matija Marolt. Janez Križaj, Jerneja Žganec Gros in Simon Dobrišek so se na sekciji predstavili s plakatom *Uporaba prisilne poravnave za fonetično analizo slovenskega govora*, Lenka Bajčetić, Vuk Batanović in Tanja Samardžić pa so govorili o lematizaciji srbskega in hrvaškega jezika z napovedovanjem urejanja nizov (angl. string edit prediction). Simona Majhenič je predstavila plakat z naslovom *Communicative intent divergence of discourse markers in simultaneously interpreted speech*. Meta Kokalj je govorila o metodi za oblikovanje podatkovne zbirke NLI na ravni odstavka na podlagi večkategorijskih scenarijev Parlay, Mateja Jemec Tomazin pa o Slovenskem terminološkem portalu. Magdalena Gapsa, Špela Arhar Holdt in Iztok Kosem so se predstavili s plakatom *Kako dober je ChatGPT pri umeščanju sopomenk pod pomene*, Janez Štebe je govoril o strojni preverbi internetnih naslovov novičarskih prispevkov v naslov na Wayback Archive, zadnji plakat, predstavljen na sekciji, pa je bil *Na poti k skladenjskim analizam šolskega pisanja: skladenjski vzorci v korpusu Šolar 3.0*, predstavili sta ga Tina Munda in Špela Arhar Holdt.

Prvi dan se je nadaljeval z drugo sekcijo, katere prispevki so se nanašali na temo govornih in parlamentarnih virov ter etike. Potekala je v slovenskem jeziku, otvorili pa so jo Darinka Verdonik, Nikola Ljubešić, Peter Rupnik, Kaja Dobrovoljc in Jaka Čibej s predstavitvijo izbora in urejanja gradiva za učni korpus govorjene slovenščine – ROG. Katja Meden, Tomaž Erjavec in Andrej Pančur so predstavili Slovenski parlamentarni korpus siParl 4.0, o osebnih podatkih v umetnosti pa sta govorila Aleš Vaupotič in Narvika Bovcon. Sekcija se je zaključila s predstavitvijo sistema za zaznavanje sprememb v rabi besed in njegove uporabe za sociolingvistično analizo, ki so jo pripravile avtorice Mateja Martinc, Veronika Bajt, Špela Rot ter Senja Pollak.

Prvi dan se je končal s panelom *Napredki in perspektive v raziskavah govorne komunikacije*, ki je potekal v slovenskem, hrvaškem, srbskem in angleškem jeziku. Panel je združeval aktivne raziskovalce s področij računalniškega jezikoslovja, govornih tehnologij, korpusnega jezikoslovja in tradicionalnih jezikoslovnih disciplin, ki so razpravljali o najnovejših dosežkih in izzivih na svojih raziskovalnih področjih, o motivih, ki so gonilo njihovih raziskav, ter o tem, kako lahko raziskave govorne komunikacije naslavljajo družbene izzive, s katerimi se soočamo danes.[1]

Drugi dan konference se je pričel s predavanjem vabljene predavateljice Barbare McGillivray, ki je predstavila skupni projekt, v katerem sodelujejo digitalni humanisti, računalniški jezikoslovci, inženirji programske opreme in kustosi knjižnic, da bi analizirali učinke mehanizacije na angleški jezik v 19. stoletju. Predavateljica je razpravljala o izzivih in spoznanjih, pridobljenih pri združevanju prostovoljnega množičnega zbiranja podatkov za zgodovinsko jezikovno anotacijo z algoritmi in oblikovalskimi poskusi.

Po vabljenem predavanju se je odvila sekcija tri, ki je potekala v angleškem jeziku, z naslovom *Linguistic annotation, historic language data.* Prvi so se predstavili Nikola Ljubešić, Luka Terčon in Kaja Dobrovoljc s prispevkom o CLASSLA-Stanza, postopku za samodejno jezikovno anotacijo južnoslovanskih jezikov, ki temelji na obdelavi naravnega jezika Stanza. Katja Meden, Ana Cvek, Vid Klopčič, Matevž Pesek, Mihael Ojsteršek, Mojca Šorn in Andrej Pančur so predstavili potek nadgradnje zgodovinarskega portala SIstory, Alice Fedotova, Adriano Ferraresi, Maja Miličević Petrović in Alberto Barrón-Cedeño pa so kot zadnji nastopajoči te sekcije predstavili potek razširitve korpusa Evropskega parlamenta za prevajanje in tolmačenje.

Sekcija štiri je prav tako potekala v angleškem jeziku, predstavili so se prispevki na temo razvoja in uporabe LLM (angl. large language model; slov. obsežni jezikovni model). Generativni model za jezik z manj viri z eno milijardo parametrov so predstavili Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič in Marko Robnik-Šikonja, Jaka Čibej pa je govoril o prvih korakih k sestavi varnostnega nabora podatkov za slovenske velike jezikovne modele. Sekcija se je zaključila s predstavitvijo velikih jezikovnih modelov pri podpori leksikografiji s poudarkom na konceptualni organizaciji hrvaških idiomov avtorjev Slobodana Belige in Ivane Filipović Petrović.

Sledila je študentska sekcija, na kateri so se predstavili trije prispevki. Prvi je imel naslov *Efficient fine-tuning techniques for Slovenian language models*, predstavili pa so ga Camile Lendering, Manfred González in Joaquín Figueira, sledil je prispevek Luke Terčona, ki je predstavil uporabo šestih mer skladenjske kompleksnosti za primerjavo jezika v govornem in pisnem korpusu, sekcijo pa je zaključil Matej Klemen, ki je govoril o testu poznavanja splošnih besed v slovenščini med udeleženci Mladinske poletne šole.

Zadnji sekciji (pet in šest) sta se odvijali vzporedno. Peta je potekala v angleškem jeziku, predstavili so se trije prispevki. Anna Kryvenko je govorila o študiji na temo stopnje pripadnosti Evropi v parlamentarnem diskurzu, ki jo je avtorica izvedla s

1 »Konferenca jezikovne tehnologije in digitalna humanistika 2024,« *SDJT – Slovensko društvo za jezikovne tehnologije*, https://www.sdjt.si/wp/jtdh-2024/, pridobljeno 27. 9. 2024.

pomočjo korpusa. Ajda Pretnar je predstavila korpusno-lingvistično karakterizacijo sPeriodike, sekcija pa se je zaključila s predstavitvijo Jakoba Lenardiča na temo skladenjskih kategorij.

Šesta sekcija je potekala v slovenskem jeziku in tudi tu so se zvrstile tri predstavitve. Mojca Stritar Kučuk je govorila o korpusu KOST 2.0 in poteku označevanja jezikovnih napak; Jaka Čibej in Tina Munda sta predstavila metodo polavtomatskega popravljanja lem in obliko skladenjskih oznak na primeru učnega korpusa govorjene slovenščine ROG; sekcijo pa sta zaključila Diana Košir in Tomaž Erjavec s predstavitvijo izdelave, opisa in analize zbirke starejših besedil v verski periodiki.

Konferenca se je zaključila s podelitvijo nagrade za najboljši študentski prispevek, ki jo je prejel Matej Klemen. Po uspešnem uradnem zaključku konference JTDH 2024 je sledil še redni letni občni zbor Slovenskega društva za jezikovne tehnologije.

Ana Cvek

Poročilo o konferenci »Družine v Alpah«

Konferenca z naslovom »Družine v Alpah« je potekala med 29. in 31. avgustom v prostorih Inštituta za novejšo zgodovino v Ljubljani, ki je bil skupaj s Fakulteto za humanistične študije Univerze na Primorskem in Inštitutom za slovensko izseljenstvo in migracije ZRC SAZU tudi eden od organizatorjev. Partnerja konference sta bila tudi Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS) in Mednarodno združenje za zgodovino Alp (AIHA/AISA/IGHA). Namen konference je bila poglobljena predstavitev tematik, povezanih z družinskimi odnosi, dediščinskimi vzorci, ekonomskim razvojem in družbenimi razmerji v alpskem prostoru ter nekaterih drugih ruralnih predelih evropskih držav.

Konferenca se je pričela s podrobnim spoznavanjem delovanja in razvoja kraške vasi Tomaj in vpogleda vanju. V tem delu so se zvrstili prispevki z dveh panelov, katerih avtorji so bili dr. Alberto Mauchigna, dr. Meta Remec, dr. Jurij Hadalin, dr. Lev Centrih, dr. Polona Sitar in mag. Leonida Borondič. V njih so obravnavali uveljavljanje družbene elite v povojni tomajski družbi in spremembe v njeni sestavi, v njeni konceptualizaciji s sočasnim dvigom življenjskega standarda in prehranske potrebe ter prehrambno samooskrbo v vasi v 19. stoletju. Nekateri avtorji so se dotaknili tudi odraščanja, vpliva šolskega sistema in z njim povezanih ideologij na vaško mladino ter življenjske poti in vloge duhovnika Albina Kjudra na historiografsko obravnavo vasi in kolektivni vaški spomin.

V višje ležeči alpski prostor nas je z analizo vzorcev dedovanja in velikosti družin v slovenskem alpskem svetu popeljal dr. Aleksander Panjek. Njegova predstavitev z naslovom »Vzorci dedovanja in velikost družine v slovenskem alpskem svetu (15.–19. stol.)« je obravnavala raznolikost (ne)deljivosti kmetij, velikosti družin in