Konferenca Jezikovne tehnologije in digitalna humanistika 2022

David BORDON
Filozofska fakulteta, Univerza v Ljubljani

1 O konferenci

Septembra 2022 je v prostorih Fakultete za družbene vede Univerze v Ljubljani potekala konferenca Jezikovne tehnologije in digitalna humanistika (JTDH), ki jo je priredilo Slovensko društvo za jezikovne tehnologije (SDJT) v soorganizaciji s Centrom za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT), Inštitutom za novejšo zgodovino (INZ) ter raziskovalnima infrastrukturama CLARIN.SI[1] in DARIAH-SI[2].

Po spletni izvedbi leta 2020 je konferenca, ki se sicer odvija vsaki dve leti, letos ponovno potekala v živo. SDJT organizira konferenco že od leta 1998; do leta 2014 sicer pod drugim imenom – konferenca Jezikovne tehnologije, leta 2016 pa je izvedlo tematsko širitev še na polje digitalne humanistike. Splošna tematska področja konference so jezikovne tehnologije, digitalno jezikoslovje in digitalna humanistika.

Konferenca je mednarodna – od 120 avtorjev prispevkov je bila skoraj tretjina tujih – večina prispevkov pa je bila predstavljena v angleščini, ki je poleg slovenščine tudi uradni jezik konference. V sklopu programa je bilo mogoče prisluhniti študentski sekciji v slovenščini in angleščini, dvema slovenskima in trem angleškim rednim sekcijam, predstavitvi plakatov tako v angleščini kot v slovenščini in dvema vabljenima predavanjema.

1 https://www.clarin.si/info/o-projektu/
2 http://www.dariah.si/

Bordon, D.: Konferenca Jezikovne tehnologije in digitalna humanistika 2022. Slovenščina 2.0, 10(1): 131–135.

1.19 Recenzija, prikaz knjige, kritika / Review, book review, critique

DOI: https://doi.org/10.4312/slo2.0.2022.1.131-135

Predstavitve so bile posnete in so na voljo na spletni strani konference JTDH 2022.[3]

2 Predkonferenčne delavnice

Konferenčno dogajanje se je začelo že dan pred uradnim odprtjem. Na Inštitutu za novejšo zgodovino (INZ)[4] so se odvijale praktične delavnice uporabe različnih orodij, ki so namenjena raziskovalkam in raziskovalcem. Ajda Pretnar Žagar je predstavila tematsko modeliranje parlamentarnih razprav na korpusu ParlaMint pred in med epidemijo covida-19 z uporabo orodja Orange. Na drugi delavnici sta Jakob Lenardič in Kristina Pahor de Maiti predstavila raziskovalno infrastrukturo in orodja CLARIN.SI, ki nudi podporo pri ustvarjanju, obdelavi, arhiviranju in ponovni uporabi jezikovnih podatkov.

3 Vabljeni predavanji

Letos sta kot vabljena predavatelja na konferenci gostovala Eetu Mäkelä in Benoît Sagot. Mäkelä, izredni profesor na Univerzi v Helsinkih in Univerzi Aalto ter tehnološki direktor infrastrukture DARIAH-FI[5], se ukvarja z interakcijo med računalništvom in humanistiko, obenem pa vodi raziskovalno skupino, ki si prizadeva ugotoviti tehnološke, procesne in teoretične temelje uspešnih računalniških raziskav v humanistiki in družboslovju. V predavanju *Designing computational systems to support humanities and social sciences research* je predstavil izsledke raziskav prej omenjene raziskovalne skupine. Med glavnimi dognanji njegove predstavitve gre izpostaviti, da sodelovanje med humanisti in računalničarji pogosto ne obrodi želenih sadov, kar je posledica neskladij med disciplinami, denimo različnih tradicij in znanstvenih pristopov, ter različno koncepcijo tega, kateri podatki so pomembni in relevantni za obdelavo. Poudaril je, da je pred začetkom interdisciplinarnih projektov pomembno vse te vidike vzeti v poštev in jih prilagoditi simbiotičnemu sodelovanju.

Drugi konferenčni dan je predaval Benoît Sagot, vodja programske skupine ALMAnaCH iz pariškega raziskovalnega centra Inria (Institut

3 https://www.sdjt.si/wp/dogodki/konference/jtdh-2022/
4 https://www.inz.si/
5 https://www.dariah.fi/

national de recherche en sciences et technologies du numérique).[6] Je specialist na področju procesiranja naravnega jezika in deluje na mnogoterih področjih digitalnega jezikoslovja. V predavanju *Large-scale language models: challenges and perspective* je predstavil nastanek večjezičnega korpusa OSCAR,[7] predvsem z vidika prečiščevanja ogromnih količin podatkov, na katerih sloni korpus (podatke pridobivajo iz *dumpov* ameriškega združenja Common Crawl). Govoril je tudi o jezikovnem modelu za francoščino CamemBERT, prvem modelu take velikosti za jezik, ki ni angleščina, o težavah, s katerimi so se srečevali v teku projekta, in načinih, kako so jih premostili. Pri CamemBERTu je vredno izpostaviti dejstvo, da 4 GB (dovolj raznolikih) podatkov zadostuje za doseganje *state-of-the-art* nivoja kakovosti.

4 Vzporedne sekcije

Jedrni del konference JTDH 2022 so bile zagotovo vzporedne sekcije v slovenskem in angleškem jeziku. Format vzporednih sekcij se je na konferenci prvič izvedel leta 2016, po širitvi na področje digitalne humanistike. V zadnjih dveh izvedbah pred letošnjo, leta 2018 in 2020, pa se je program delil na tematske sklope. Na letošnji konferenci smo lahko v osrednjem delu poslušali predstavitve 70 avtorjev, od tega 12 prispevkov v slovenščini in kar 17 prispevkov v angleščini. Zaradi števila in jezikovne narave prijavljenih prispevkov se je sistem vzporednih sekcij pokazal kot smiseln, razporeditev udeležencev pa zaradi visoke mednarodne udeležbe precej homogena.

Na področju korpusnega jezikoslovja smo lahko spoznali tri nove korpuse – korpus Trendi[8], prvi spremljevalni korpus za slovenščino, ki uporabnikom nudi podatke o aktualni jezikovni rabi in omogoča diahrone jezikovne analize, korpus študentskih besedil KOŠ, namenjen pridobivanju empiričnih podatkov o pisni jezikovni zmožnosti študentske populacije, in hrvaški korpus DirKorp, specializiran za govorna dejanja. Z uporabniškega vidika je bila predstavljena raba *Kolokacijskega slovarja sodobne slovenščine* (KSSS)[9] pri prevajanju kolokacij iz angleščine v slovenščino. Analiza je bila izvedena na vzorcu dodiplomskih študentov Oddelka za

6 https://www.inria.fr/fr
7 https://oscar-project.org/
8 https://sled.ijs.si/korpus-trendi/
9 https://viri.cjvt.si/kolokacije/slv/#

prevajalstvo FF UL – izsledki kažejo, da je sposobnost uporabe slovarja sorazmerna s kakovostjo rešitev, do katerih uporabnik lahko pride, hkrati pa sama raba jezikovnih virov ni zagotovilo, da bo prevodna rešitev ustrezna. Obratno, raba jezikovnih virov ni zagotovilo, da bo prevodna rešitev ustrezna. V debati je bilo izpostavljeno, da slovenski visokošolski univerzitetni programi študente dobro učijo, kako uporabljati jezikovna orodja.

Pri govornih tehnologijah so bili med drugim predstavljeni najnovejši napredki pri samodejni slovenski grafemsko-fonemski pretvorbi ter projekt poravnave zvočnih posnetkov s transkripcijo narečnega govora in petja. Izpostaviti velja prispevek, vezan na izgradnjo govorne baze Artur. Avtorji so se posvetili primerom dobre prakse pri poenotenju metapodatkov med združevanjem različnih govornih korpusov in predlagali načine, kako se v bodoče izogniti neskladjem med metapodatki.

Številni prispevki so se posvečali diskurzu, med zanimivejšimi so bili prispevki o sovražnem in grobem besedišču v odzivnemu *Slovarju sopomenk sodobne slovenščine* (SSSS)[10] ter dva, vezana na parlamentarno okolje; prvi se je posvetil pregledu mednarodnih raziskav parlamentarnega diskurza v zadnjih desetih letih, drugi pa je bil vezan na populistični diskurz v slovenskem parlamentu med letoma 1992 in 2018.

Velikega napredka je bilo deležno področje označevanja – na konferenci so bili predstavljeni primeri dobre prakse in optimalne rešitve, uporabljene pri projektu oblikoskladenjskega označevanja korpusa SentiCoref[11], ki bo vključen v nov učni korpus za slovenščino (trenutni ssj500k[12]), v sklopu projekta Razvoj slovenščine v digitalnem okolju (RSDO)[13]. Poleg tega so se v okviru projekta RSDO izvajale aktivnosti v povezavi s shemo Universal Dependencies (UD)[14] – raziskovalci so obstoječo infrastrukturo nadgradili in ustvarili dokumentacijo označevalnih smernic UD za slovenščino.

Pri strojnem prevajanju je izstopal predvsem prispevek, ki je predstavil človeško evalvacijo prevodnih rešitev strojnega prevajalnika za jezikovno kombinacijo slovenščina-angleščina, ki nastaja na projektu RSDO. Na področju terminologije pa so bili prispevki vezani predvsem na modele samodejnega luščenja terminov.

10 https://viri.cjvt.si/sopomenke/slv/
11 https://www.clarin.si/repository/xmlui/handle/11356/1285
12 https://www.clarin.si/repository/xmlui/handle/11356/1434
13 https://slovenscina.eu/
14 https://universaldependencies.org/

5 Študentska sekcija in predstavitve plakatov

Velika dodana vrednost konference JTDH je samostojna študentska sekcija, uvedena že leta 2016, in sekcija s plakati, ki obstaja od leta 2018. Mladim raziskovalkam in raziskovalcem ter študentkam in študentom je tako omogočeno, da se lahko (brezplačno) prijavijo na konferenco – če je njihov prispevek sprejet, ga lahko predstavijo v eni izmed omenjenih sekcij in objavijo v zborniku. Za mnoge mlade raziskovalke in raziskovalce je objava v zborniku JTDH prva resnejša objava znanstvenega prispevka, kar predstavlja velik doprinos, saj so jim na začetku poklicne poti tovrstne možnosti običajno povsem (predvsem finančno) nedostopne.

Samostojna študentska sekcija je po izvedbi ekvivalentna ostalim jedrnim sekcijam, letos so denimo predstavitve trajale 10 minut, sledila pa so vprašanja občinstva. Dinamika sekcije s plakati je nekoliko drugačna, saj se odvija v preddverju, avtorji prispevkov pa so na voljo za predstavitev, vprašanja ali pogovor.

6 Občni zbor SDJT – predstavitev vmesnih rezultatov projekta RSDO in zaključek

Po formalnemu zaključku konference je sledil občni zbor SDJT in predstavitev orodij, ki so nastala v sklopu projekta RSDO, ki je v času pisanja poročila v zaključni fazi. Predstavitve so si sledile po delovnih sklopih, v katerih so bili izpostavljeni strojni označevalnik, metakorpus slovenskega jezika, terminološki portal, modeli strojnega prevajalnika, orodje za prepoznavanje imenskih entitet in koreferenčnosti, ekstrakcijo povezav, baza znanja, orodja za povzemanje besedil ter orodje za semantične premike in diahrone analize.

Zelo plodna izvedba konference – po štirih letih ponovno v živo – je pokazala, da se konferenca JTDH vedno bolj uveljavlja na mednarodnem parketu, obenem pa ohranja svojo dostopnost mlajši generaciji. Tematska raznolikost in bogatost programa ji utrjujejo položaj kot eni pomembnejših pri nas. Čestitke za odlično izvedbo; še na mnoga leta.