Predgovor

S pojavom digitalnega medija se je metodologija (tudi) na področju uporabnega jezikoslovja pomembno razvila. Skupnosti so na voljo referenčni in številni specializirani besedilni korpusi, učne množice ter drugi informacijsko bogati jezikovni viri za sodobno slovenščino. Razvili so se strojni postopki za pripis jezikoslovnih informacij v digitalna besedila in napredno pridobivanje jezikovnih podatkov iz raznovrstnih besedilnih zbirk. Dostopni so korpusni konkordančniki in druga orodja za izvedbo empirično osnovanih kvantitativnih in kvalitativnih jezikoslovnih analiz. Vedno več podatkovne infrastrukture za slovenščino je na voljo povsem odprto in to spodbuja nastanek novih izdelkov in storitev.

Skupaj z novimi možnosti se pojavljajo tudi novi raziskovalni izzivi in razvojne potrebe. Ko danes razmišljamo o naslednji posodobitvi slovničnega opisa sodobne slovenščine, že vemo, da ta ne bo le vsebinska, ampak bo morala biti predvsem metodološka in konceptualna: osnovana na empiričnih, strojno berljivih, medsebojno povezljivih, večnamensko zasnovanih in odprto dostopnih slovničnih podatkih. Tej nalogi se posveča delo, ki je pred vami.

Monografija *Nova slovnica sodobne standardne slovenščine: viri in metode* predstavlja rezultate istoimenskega raziskovalnega projekta, ki je potekal med leti 2017 in 2020 s finančno podporo ARRS. V projektu smo sodelovali predstavnice in predstavniki Instituta »Jožef Stefan« ter Univerze v Ljubljani: Filozofske fakultete in Fakultete za računalništvo in informatiko. Interdisciplinarna ekipa je pod vodstvom Simona Kreka združila znanja s področja digitalne slovenistike ter strojnega procesiranja naravnega jezika in postavila metodološke temelje celostne računalniške analize sodobnega jezika, kakršen je zajet v referenčnih korpusnih virih. Na podlagi nove metodologije smo izdelali odprto dostopne podatkovne baze, uporabne za različne namene, v končni fazi – upamo – tudi

za korpusno osnovani slovnični opis sodobne slovenščine. Pripravo podatkovnih baz, virov in orodij osvetljuje devet monografskih prispevkov, ki jih je spisalo osem sodelujočih raziskovalcev in raziskovalk.

V prvem prispevku Špela Arhar Holdt in Jaka Čibej predstavita analize za nadgradnjo **učnega korpusa ssj500k,** ki je eden od temeljnih virov za nadzorovano strojno učenje jezikoslovnega označevanja sodobne pisne slovenščine.

V drugem prispevku Jaka Čibej, Špela Arhar Holdt in Marko Robnik Šikonja predstavijo zasnovo in delovanje **programa LIST,** s katerim je mogoče v relativno kratkem času izvoziti raznolike jezikovne podatke iz (referenčnih ali specializiranih) besedilnih korpusov.

V tretjem prispevku Špela Arhar Holdt opiše nastanek **baze oblikoslovnih podatkov,** v kateri je 96.290 enotam leksikona besednih oblik Sloleks (samostalnikom, pridevnikom, glagolom in prislovom) pripisana koda oblikoslovnega vzorca, po katerem se pregibajo.

V četrtem prispevku Jaka Čibej predstavi metodologijo strojnega povezovanja leksikonskih enot glede na njihovo besedotvorno sorodnost. Z metodologijo, ki je v prispevku jezikoslovno evalvirana, je pripravljena **baza povezanih leksikonskih enot** v obsegu 66.347 povezav.

V petem prispevku Simon Krek, Polona Gantar, Iztok Kosem in Kaja Dobrovoljc opišejo metodologijo **izboljšanega luščenja kolokacijskih podatkov,** s katero iz skladenjsko označenega korpusa Gigafida 2.1 izluščijo 4.002.918 kolokacijskih kandidatov v 81 skladenjskih strukturah.

V šestem prispevku Polona Gantar predstavi pravila za zapis **kanonične oblike frazeoloških enot** v novo izdelanem Leksikonu večbesednih enot in na podlagi izluščenih podatkov prikaže tudi konkretne rešitve povezovanja variantnih in pretvorbeno povezanih frazeoloških enot v leksikonu.

V sedmem prispevku Tadej Škvorc, Polona Gantar in Marko Robnik Šikonja preizkusijo in ocenijo pristope za **strojno prepoznavanje idiomov** na podlagi globokih nevronskih mrež, ki uporabljajo vektorske vložitve.

V osmem prispevku Polona Gantar opiše izdelavo strojno berljivega **Vezljivostnega leksikona slovenskih glagolov** s postopki avtomatskega luščenja vezljivostnih vzorcev iz oblikoslovno, skladenjsko in semantično označenega korpusa Gigafida 2.1.

V devetem prispevku Kaja Dobrovoljc predstavi izdelavo in vsebino **leksikona formulaičnih besednih nizov v pisni in govorjeni slovenščini**, ki prinaša podatek o skladenjski zgradbi, pragmatični funkciji in potencialni slovarski relevantnosti posameznega niza.

Prispevki popisujejo vire in metode, nastale v projektu, kar je podlaga za ustrezno uporabo projektnih rezultatov, prav tako pa opozarjajo na šibka mesta trenutnega stanja, ki jih bo mogoče nasloviti v nadaljnjem delu. Slednje že poteka pod okriljem projekta Razvoj slovenščine v digitalnem okolju, ki ga med leti 2020 in 2023 financirata Ministrstvo za kulturo Republike Slovenije in Evropski sklad za regionalni razvoj. Spoznanja, pridobljena v raziskovalnem projektu, se torej tekoče prelivajo v aplikativno razvojno prakso, kar je izrednega pomena za razvoj področja in v širšem smislu celotne družbe, saj je digitalna jezikovna infrastruktura kot temelj jezikovne opremljenosti pomembna za prav vse dejavnosti, ki vključujejo jezikovno rabo v digitalnem svetu.

Priprava monografije je potekala v času, ko smo se sodelujoči že globoko zakopali v naloge na novih projektih, zato mi je njen uspešen izid v posebno veselje. Avtorjem in avtoricam se zahvaljujem za vztrajno delo in vsebinsko bogate prispevke. Zahvaljujem se recenzentkam in recenzentom, ki so posredovali natančne in konstruktivne recenzije: Tomaž Erjavec, Mateja Jemec Tomazin, Boris Kern, Nina Ledinek, Nikola Ljubešić, Nataša Logar, Tadeja Rozman, Mojca Stritar Kučuk, Darinka Verdonik in Slavko Žitnik. Enaka zahvala gre za ekipne kolegialne recenzije, ki so jih pripravili Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Iztok Kosem in Marko Robnik Šikonja, ter pregledno branje, ki sta ga opravili Senja Pollak in Mojca Smolej. Za prijazno pomoč pri tehničnem urejanju se zahvaljujem Tini Munda, za čudovito podobo monografije pa ekipi Znanstvene založbe Filozofske fakultete Univerze v Ljubljani. Zahvala gre seveda Javni agenciji za raziskovalno dejavnost Republike Slovenije,

ki je projekt Nova slovnica sodobne standardne slovenščine: viri in metode (ARRS J6-8256) sofinancirala iz državnega proračuna, nenazadnje pa vsem bralkam in bralcem, ki boste podarili čas pregledu novosti, morda pa tudi uporabi in nadaljnjemu razvoju projektnih rezultatov in idej.

Špela Arhar Holdt,
Kopenhagen, 8. 8. 2021