Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec[1], Kaja Dobrovoljc[3,1], Darja Fišer[4,3,1], Jan Jona Javoršek[1], Simon Krek[2,1], Taja Kuzman[1], Cyprian Laskowski[2], Nikola Ljubešić[1,2], Katja Meden[1]

[1] Institut »Jožef Stefan«
tomaz.erjavec@ijs.si, kaja.dobrovoljc@ijs.si, jan.javorsek@ijs.si, simon.krek@ijs.si,
taja.kuzman@ijs.si, nikola.ljubesic@ijs.si, katja.meden@ijs.si
[2] Center za jezikovne vire in tehnologije Univerze v Ljubljani
cyp@cjvt.si
[3] Filozofska fakulteta Univerze v Ljubljani
darja.fiser@ff.uni-lj.si
[4] Inštitut za novejšo zgodovino

Povzetek

Prispevek povzame storitve slovenske raziskovalne infrastrukture za jezikovne vire in tehnologije CLARIN.SI, ki je članica evropskega konzorcija raziskovalnih infrastruktur CLARIN ERIC. Najprej obravnavamo vodenje, organizacijo in tehnično infrastrukturo CLARIN.SI, nato pa njene spletne storitve, predvsem repozitorij digitalnih jezikovnih virov in orodij ter konkordančnike. Sledi pregled promocije področja jezikovnih tehnologij in digitalne humanistike v Sloveniji, kar vključuje storitve centra znanja za računalniško obdelavo južnoslovanskih jezikov CLASSLA, financiranje projektov in organizacijo, podporo ali sodelovanje na konferencah in delavnicah. Predstavimo tudi sodelovanje CLARIN.SI s CLARIN ERIC in s sorodnima slovenskima infrastrukturama DARIAH-SI in CESSDA/ADP ter vključenost v slovenske in evropske projekte.

The CLARIN.SI Research Infrastructure

The paper summarises the services offered by the Slovenian research infrastructure for language resources and technologies CLARIN.SI, which is a member of the European research infrastructure consortium CLARIN ERIC. We first present the governance, organisation and technical infrastructure of CLARIN.SI, followed by a description of its web applications with a focus on its repository and concordancers. Next comes an overview of support activities that CLARIN.SI offers to the fields of language technologies and digital humanities in Slovenia, which includes services of the knowledge centre for computational processing of South-Slavic languages CLASSLA, financial support of projects, and organisation or support of conferences and workshops. We also introduce the work of CLARIN.SI within CLARIN ERIC, its cooperation with its sister national infrastructures DARIAH-SI and CESSDA/ADP, and involvement in national and European projects.

1. Uvod

Raziskovalna infrastruktura (RI) CLARIN[1] (»Common Language Resources and Technology Infrastructure« oz. »Infrastruktura za skupne jezikovne vire in tehnologije«) zagotavlja digitalne jezikovne vire, orodja in storitve za podporo raziskovalcem na področju humanistike in družboslovja in drugih področij, ki se ukvarjajo z jezikom (Jong et al., 2018). CLARIN je bila ena od infrastruktur, ki so bile predvidene že v prvem načrtu Evropskega strateškega foruma za raziskovalne infrastrukture ESFRI (Váradi et al., 2008). Ustanovljena je bila leta 2012 in je bila ena prvih RI, ki je pridobila status evropske pravne osebe konzorcija raziskovalnih infrastruktur ERIC (European Research Infrastructure Consortium). CLARIN ERIC ima sedež na Nizozemskem in trenutno združuje RI 22 držav članic in 3 opazovalke. Zaposluje vodjo in podporno osebje za koordinacijo in centralne tehnične storitve, medtem ko imajo glavno vlogo pri zagotavljanju storitev nacionalni centri RI.

Glede na pomen slovenskega jezika za Slovenijo je sodelovanje v CLARIN ključnega pomena, saj spodbuja empirično podprto raziskovanje jezika ter razvoj jezikovnih virov in tehnologij, s čimer lahko slovenščina v informacijski družbi nastopa enakopravno z drugimi jeziki, tudi mnogo večjih skupnosti (Krek, 2022). Korist od RI imajo raziskovalci, učitelji in študenti slovenskega jezika ter drugih jezikoslovnih smeri, računalniškega jezikoslovja in umetne inteligence, pa tudi drugi raziskovalci s področja humanistike in družboslovja, ki pri svojem delu uporabljajo jezikovna gradiva. RI nudi podporo tudi slovaropiscem, prevajalcem in podjetjem, ki v svoje produkte vključujejo obdelavo slovenskega jezika, nenazadnje pa tudi laičnim uporabnikom za namene reševanja praktičnih vprašanj.

Slovenska RI CLARIN.SI je bila ustanovljena leta 2014, članica CLARIN ERIC pa je postala leta 2015, za kar je bilo potrebno, da je bil ustanovljen nacionalni konzorcij in da je Vlada Republike Slovenije podpisala memorandum, s katerim se je zavezala plačevati članarino za članstvo Slovenije v CLARIN ERIC. Do sedaj je bila edina publikacija, ki celostno predstavi CLARIN.SI, objavljena kmalu po njeni ustanovitvi (Erjavec et al., 2014), kjer smo predstavili prve korake RI in načrte za nadaljnje delo. Pričujoči prispevek povzema, kaj je bilo narejenega v minulih osmih letih: v razdelku 2. predstavimo organizacijsko strukturo in upravljanje infrastrukture, v 3. repozitorij jezikovnih virov in orodij, v 4. spletne storitve, v 5. podporne dejavnosti, v 6. vpetost CLARIN.SI v domače in evropske projekte in v aktivnosti CLARIN ERIC, v 7. pa podamo zaključke in načrte za nadaljnje delo.

[1] https://www.clarin.eu/

2. Organiziranost CLARIN.SI

Infrastruktura ima sedež na Institutu »Jožef Stefan« (IJS), kjer tudi domuje večina njene računalniške opreme in kjer se zagotavlja varnost, vzdrževanje in neprestano obratovanje spletnih storitev RI. Pri vodenju in tehničnem vzdrževanju sodelujejo tri organizacijske enote IJS, in sicer Odsek za tehnologije znanja E8, Laboratorij za umetno inteligenco E3 ter Center za mrežno infrastrukturo CMI.

CLARIN.SI je organiziran kot konzorcij, ki nima narave pravne osebe, v njem pa ima članstvo 12 partnerjev.V konzorciju so združene vse glavne institucije, ki se v Sloveniji ukvarjajo z razvojem ali uporabo jezikovnih virov in tehnologij, in sicer:

- Univerze: Univerza v Ljubljani, Univerza v Mariboru, Univerza v Novi Gorici in Univerza na Primorskem. Univerza v Ljubljani je sedež Centra za jezikovne vire in tehnologije (CJVT), ki koordinira delo na področju korpusnega jezikoslovja in jezikovnih tehnologij ter razvija in vzdržuje temeljne digitalne jezikovne vire in jezikovnotehnološka orodja za sodobni slovenski jezik.

- Raziskovalni inštituti: ZRC SAZU, Institut »Jožef Stefan« (IJS), Inštitut za novejšo zgodovino (INZ) in Znanstveno-raziskovalno središče Koper. Znotraj ZRC SAZU Inštitut za slovenski jezik Frana Ramovša zbira jezikovno gradivo in ga uporablja za izdelavo temeljnih del slovenskega jezikoslovja, predvsem slovarjev. IJS kot gostitelj raziskovalne infrastrukture CLARIN.SI koordinira delo infrastrukture, vzdržuje in nadgrajuje njen repozitorij in storitve ter razvija jezikovne vire in orodja.

- Društva oz. zavodi: Slovensko društvo za jezikovne tehnologije (SDJT), ki s konferenco »Jezikovne tehnologije in digitalna humanistika« (JTDH) promovira razvoj jezikovnih tehnologij za slovenski jezik, in Zavod za uporabno slovenistiko Trojina s svetovalno in podporno dejavnostjo ter izdelavo jezikovnih virov in orodij.

- Podjetji Alpineon in Amebis, med katerima prvo podjetje v infrastrukturo CLARIN.SI prispeva predvsem govorne tehnologije, drugo pa se ukvarja z izdelavo programske opreme s področja jezikovnih tehnologij in elektronskega založništva.

Odločitve o vodenju RI sprejema oz. potrjuje Upravni odbor (UO) CLARIN.SI, v katerem ima vsak partner po enega predstavnika in poljubno število namestnikov oz. namestnic. Komunikacija se odvija prek dopisnega seznama upravnega odbora, ki trenutno šteje 34 članov, enkrat letno pa organiziramo sestanek CLARIN.SI UO, na katerem se pogovorimo o delovanju RI v preteklem letu in naredimo načrte za naslednje.

Delovanje raziskovalne infrastrukture CLARIN v Sloveniji se tako kroji na podlagi potreb in konsenza vseh večjih akterjev na področju digitalnega jezikoslovja in jezikovnih tehnologij, kot tudi digitalne humanistike in družboslovja, saj CLARIN.SI tesno sodeluje z dvema sestrskima RI v slovenskem prostoru. To sta DARIAH-SI s sedežem na Inštitutu za novejšo zgodovino (INZ), ki predstavlja nacionalno vozlišče evropske RI za digitalno humanistiko, in CESSDA/ADP v Arhivu družboslovnih podatkov na Fakulteti za družbene vede Univerze v Ljubljani (ADP), ki je nacionalno vozlišče evropske RI za digitalno družboslovje CESSDA (»Consortium of European Social Science Data Archives«). CLARIN.SI je tudi ena od ustanovnih članic Slovenskega nacionalnega superračunalniškega omrežja SLING[2] in preko njega članica federacije računskih in podatkovnih virov EGI[3] ter Partnerstva za napredno računalništvo v Evropi PRACE[4].

CLARIN.SI vzdržuje dvojezično (slovenščina, angleščina) spletno stran,[5] na kateri je predstavljena RI kot tudi vse njene storitve. Spletno mesto nudi tudi kontaktne podatke, npr. e-poštni naslov, na katerega se lahko obrnejo uporabniki, ki si želijo pomoči ali nasvetov. Poleg tega spletno mesto vključuje z geslom zaščitene interne strani, do katerih imajo dostop člani oz. namestniki UO in ki vsebujejo ustanovne dokumente, zapisnike sestankov, relevantne zapisnike CLARIN ERIC itd.

Za dokumentiranje tehničnega vzdrževanja CLARIN.SI uporablja interno instalacijo platforme WordPress, na kateri dokumentiramo postopke vzdrževanja za vse spletne storitve CLARIN.SI, medtem ko se za zahtevke za reševanje odkritih problemov uporablja instalacijo platform Redmine.

Kritične spletne storitve CLARIN.SI so vedno instalirane tudi na razvojnem strežniku, kjer se najprej preveri delovanje vsake spremembe na programski opremi, na ponujenih jezikovnih virih ali v dokumentaciji. Delovanje spletnih storitev se preverja prek sistema NAGIOS, repozitorij pa tudi neodvisno s strani CLARIN ERIC. V primeru napak so tako skrbniki storitve nemudoma obveščeni in lahko takoj pristopijo k odpravljanju težave.

3. Repozitorij jezikovnih virov

Osnovna storitev CLARIN.SI je vzdrževanje repozitorija jezikovnih raziskovalnih podatkov oz. jezikovnih virov, kot so velike in bogato označene zbirke besedil (korpusi), računalniški leksikoni in modeli, pa tudi strojno berljivi slovarji in računalniška orodja. Računalniška platforma repozitorija je odprtodostopna CLARIN-DSpace,[6] ki so jo razvili posebej za namene CLARIN repozitorijev v okviru češke raziskovalne infrastrukture CLARIN (sedaj CLARIAH, ki je nastala po združitvi češke CLARIN in DARIAH) na Inštitutu za formalno in uporabno jezikoslovje na Karlovi Univerzi v Pragi. Platformo poleg Slovenije, in

[2] https://www.sling.si/
[3] https://www.egi.eu/
[4] https://prace-ri.eu/
[5] https://www.clarin.si/
[6] https://github.com/ufal/clarin-dspace

seveda Češke, uporablja še sedem drugih nacionalnih repozitorijev CLARIN, kar skupaj predstavlja 40 % vseh rednih članic CLARIN ERIC.

Repozitorij CLARIN.SI je poleg ADP edini v Sloveniji akreditiran s certifikatom »Core Trust Seal«,[7] torej kot zaupanja vreden podatkovni repozitorij. Repozitorij v skladu s strategijo CLARIN ERIC implementira načela FAIR[8,9](najdljivost, dostopnost, interoperabilnost in ponovna uporaba). Evropski agendi za odprto znanost in načelom FAIR CLARIN sledi avant la lettre (Jong et al., 2018), in sicer z naslednjimi instrumenti:

- Akademska avtentikacija AAI, ki deluje po sistemu SSO (»Single sign-on«), kjer ločimo ponudnike identitete (Arnes, univerze, druge akademske institucije) in ponudnike storitev (v našem primeru repozitorij), da uporabnikom ni potrebno ustvariti svojega računa na CLARIN.SI, pač pa se v repozitorij prijavijo prek svojega EduGain uporabniškega imena in gesla pri izbranem ponudniku identitete.
- Trajni identifikatorji vnosov po sistemu »handle«, kar omogoča pripis trajnega naslova URL vsakemu vnosu v repozitorij, ki je, enako kot DOI, neodvisen od specifičnega URL-ja tega vira v okviru repozitorija, in s tem tudi odporen na spremembe v platformi oz. lokaciji repozitorija.
- Vpetost v mednarodne spletne agregatorje metapodatkov, kot so OpenAIRE[10], Re3data[11], od 2022 pa tudi European Language Grid. Preko CLARIN ERIC je bil CLARIN.SI med prvimi RI vključen tudi v sistem ponudbe virov in storitev v okviru Evropskega odprtega znanstvenega oblaka EOSC[12] že vse od vzpostavitve portala EOSC leta 2018. V okviru RI CLARIN se za metapodatkovne zapise uporablja priporočila CMDI[13] (»Component MetaData Infrastructure«), izvoz oz. žetev metapodatkov pa je omogočena tudi v standardu Dublin Core.
- Bogata izbira licenc, od odprtih, kot so licence Creative Commons, do bolj omejenih, ki zahtevajo predhodno prijavo v repozitorij in digitalni podpis sporazuma o uporabi vira.
- Eksplicitni pogoji uporabe, ki določajo pravice in dolžnosti tako upravljalcev repozitorija kot uporabnikov.
- Navodila za deponiranje vnosov, ki opišejo postopek oddaje virov s posebnim poudarkom na zahtevanih metapodatkih in njihovi obliki, saj se pri CLARIN.SI trudimo vzdrževati čim bolj popolne in enotne metapodatkovne zapise.
- Navodila za kodiranje deponiranih podatkov, ki navajajo sprejemljive formate zapisa in načine označevanja podatkov, poleg tega pa zajemajo tudi splošna navodila za pripravo kvalitetnih in usklajenih podatkov. Po tem se repozitorij CLARIN.SI razlikuje od večine drugih repozitorijev CLARIN (Lenardič in Fišer, 2022), saj ti tipično ponujajo samo seznam sprejemljivih formatov, ne pa tudi bolj splošnih navodil za pripravo kakovostnih podatkov, kakršna so lahko zelo koristna za avtorje s področja humanističnih znanosti brez poglobljenega znanja računalniških veščin za pravilno pripravo podatkov.
- Seznam pogosto postavljenih vprašanj z odgovori in podobne vsebine.

Poleg prilagojenosti za opis jezikovnih virov je za razliko od splošnih repozitorijev za samoarhiviranje, kot je npr. Zenodo, pomembna odlika repozitorija CLARIN.SI zagotavljanje visoke kvalitete deponiranih jezikovnih virov in njihovih metapodatkov, saj vsak vnos pred objavo skrbno pregleda eden od urednikov repozitorija, ki preveri, ali vnos ustreza merilom CLARIN.SI. Če jim ne, urednik vnos zavrne z obrazložitvijo napak, v vnaprej dogovorjenih primerih pa tudi pomaga pri popravljanju vira.

V osmih letih, kolikor jih je minilo od prvega vnosa, je število deponiranih jezikovnih virov in orodij naraslo na več kot 300, ki so rezultat dela prek 700 avtorjev, pri čemer je v mnoge bilo vloženih več let dela. V letu 2021 je repozitorij beležil okoli 40.000 ogledov in 4.000 prenosov. V tem letu so bili najpogosteje preneseni viri zbirka 751 emodžijev z avtomatsko pripisanim sentimentom, ki je bil izračunan na podlagi 70.000 tvitov v 13 evropskih jezikih, označenih za sentiment in s strani 83 anotatorjev (Kralj Novak et al., 2015) ter jezikovni modeli (besedne vložitve) tipa BERT (Devlin et al., 2018) za slovenske besede (Ulčar in Robnik-Šikonja, 2021), ki so koristni za marsikatero nalogo obdelave slovenskega jezika.

S spodbujanjem deponiranja jezikovnih virov in pomočjo pri njihovem oblikovanju in opisu je CLARIN.SI bistveno pripomogel k uveljavljanju koncepta odprte, preverljive, ponovljive in odgovorne znanosti na področju jezikoslovnih raziskav v Sloveniji ter številne jezikovne vire, nastale kot rezultat slovenskih raziskovalnih projektov, obvaroval pred izginotjem in jim omogočil mednarodno vidnost in vpliv.

4. Spletne storitve

Poleg repozitorija CLARIN.SI trajno vzdržuje več spletnih storitev, od katerih so najpomembnejši konkordančniki, tj. orodja za analizo korpusov, in sicer ponuja CLARIN.SI uporabo konkordančnika KonText in dveh različic konkordančnika noSketch Engine (Crystal in Bonito). Vsi trije uporabljajo isti zaledni program, in sicer Manatee (Rychlý, 2007), ki omogoča

[7] https://www.coretrustseal.org/
[8] https://www.go-fair.org/fair-principles/
[9] https://www.clarin.eu/fair
[10] https://www.openaire.eu/
[11] https://www.re3data.org/
[12] https://eosc-portal.eu/
[13] https://www.clarin.eu/content/component-metadata

hitre poizvedbe po bogato označenih korpusih, vendar se razlikujejo v čelnem delu. NoSketch Engine je odprtokodna različica komercialnega konkordančnika Sketch Engine (Kilgarriff et al., 2014),[14] medtem ko je bil KonText razvit na oddelku Češkega nacionalnega korpusa Karlove univerze v Pragi (Machálek, 2020). Poleg izgleda konkordančnikov so glavne razlike med njimi v tem, da noSketch Engine ponuja nekaj več funkcionalnosti kot KonText (predvsem možnost izračuna ključnih besed korpusa oz. podkorpusa), medtem ko KonText podpira prijavo prek sistema AAI (enako kot repozitorij), kar nato omogoča personalizirane nastavitve zaslona, hranjenje zgodovine poizvedb, itd.

Vsi konkordančniki na CLARIN.SI ponujajo isti nabor korpusov, ki jih je sedaj že preko 40, od referenčnih do specializiranih, pa tudi govorjenih in večjezičnih. Tu izpostavimo novi korpus metaFida, ki združuje 34 obstoječih korpusov in vsebuje skupaj 4,5 milijarde pojavnic, s čimer je največji in najbolj raznovrsten korpus za slovenščino, po katerem je mogoče iskati s pomočjo konkordančnikov.

Konkordančniki CLARIN.SI se uporabljajo pri izvajanju študijskih programov na več univerzah, v sklopu jezikoslovnih raziskav ali pri različnih raziskovalnih projektih, kot tudi v prevajalskih podjetjih.

Naslednja spletna storitev, ki jo ponuja CLARIN.SI, je platforma za ročno označevanje korpusov WebAnno (Yimam et al., 2013), ki so jo razvili v okviru CLARIN-DE. V okviru CLARIN.SI smo razvili pretvorbo iz zapisa korpusov TEI v format TSV3, ki ga uporablja WebAnno, in združevanje izvornega korpusa TEI z ročnimi oznakami iz datoteke TSV, s čimer smo omogočili dodajanje oz. spreminjanje obstoječih oznak v TEI kodiranih korpusih z oznakami, ki so bile ročno vstavljene oz. popravljene na platformi WebAnno (Erjavec et al., 2016)[15]. Naša instalacija in pretvorba je bila do sedaj uporabljena pri prek 10 projektih, npr. za ročno označevanje normaliziranih besednih oblik, lem in oblikoslovnih oznak uporabniško generiranih vsebin v okviru projekta Janes »Jezikoslovna analiza nestandardne slovenščine« (Fišer et al., 2020),[16] za označevanje dvojezičnih terminov v okviru projekta KAS »Slovenska znanstvena besedila: viri in opis« (Erjavec et al., 2021)[17] ali za označevanje definicij terminov v besedilih v okviru projekta TermFrame »Terminologija in sheme znanja v medjezikovnem prostoru« (Vintar in Martinc, 2022).

Za kontrolirano in kolaborativno vzdrževanje je postala zelo popularna platforma Git, ki jo v okviru CLARIN.SI prav tako uporabljamo, ne samo za programsko opremo, temveč tudi za jezikovne vire. Za spletno dostopne repozitorije Git, ki vključujejo tudi množico drugih funkcij, kot so zahtevki in izvajanje programov, sta najbolj uporabljana GitHub in GitLab. Na GitHubu ima CLARIN.SI svojo virtualno organizacijo,[18] ki združuje sedaj že okoli 60 odprtokodnih projektov. Za razliko od GitHuba, ki obstaja samo kot spletna storitev v lasti podjetja Microsoft, je mogoče platformo GitLab tudi instalirati, kar ima to prednost, da so projekti locirani na lokalni računalniški opremi, dostopnost projektov pa je mogoče tudi omejiti, kar je v posameznih primerih potrebno, npr. zaradi avtorskih pravic nad besedili nekega jezikovnega vira, ki se ga razvija. Instalacija GitLab na CLARIN.SI[19] vsebuje okoli 20 projektov, tako javnih (kot npr. že omenjena pretvorba TEI za WebAnno) kot tudi zasebnih.

CLARIN.SI v okviru centra znanja CLASSLA, ki ga obravnavamo v naslednjem razdelku, ponuja tudi spletno storitev ReLDIanno za jezikoslovno označevanje besedil v slovenskem, hrvaškem in srbskem jeziku.[20] Storitev podpira oblikoskladenjsko označevanje, lematizacijo, označevanje imenskih entitet in skladenjsko razčlenjevanje, dostopna pa je tako prek spletnega vmesnika kot prek vtičnika API, pri čemer lahko rezultate prikaže na zaslonu ali pa označeno besedilo prenesemo na lastni računalnik.

5. Strokovna podpora in diseminacija

5.1. Središča znanja

CLARIN.SI je aktiven pri promociji in spodbujanju razvoja računalniškega jezikoslovja, ne le za slovenščino, ampak tudi za druge južnoslovanske jezike, kot so hrvaščina, srbščina, makedonščina in bolgarščina, s čimer si je RI bistveno povečala mednarodno odmevnost. CLARIN.SI namreč skupaj z bolgarsko raziskovalno infrastrukturo CLARIN CLADA-BG in hrvaškim Institutom za hrvaški jezik in jezikoslovje upravlja središče znanja CLARIN za južnoslovanske jezike CLASSLA, v okviru katerega ponuja strokovno podporo pri uporabi jezikovnih virov in tehnologij za južnoslovanske jezike. Središče znanja podpira raziskovalce z dokumentacijo o prosto dostopnih jezikovnih virih, orodjih za ustvarjanje in obdelavo besedilnih korpusov ter drugih jezikovnih tehnologijah. Poleg tega CLASSLA razvija lastne jezikovne tehnologije in korpuse, s katerimi pokriva velike potrebe južnoslovanskih jezikov, ki so tehnološko manj podprti. Tako je na primer v letu 2020 v sklopu projekta zbiranja korpusov besedil iz Wikipedije središče ustvarilo prvi jezikoslovno označeni makedonski korpus, CLASSLAWikimk (Ljubešić et al., 2021).

V 2021 je CLARIN.SI postal tudi član CLARIN centra znanja za obdelavo uporabniško generiranih vsebin CKCMC,[21] ki ga vodi Eurac Research, Bolzano.

5.2. Financiranje projektov

CLARIN.SI finančno podpira projekte, letno izbrane na odprtem razpisu za člane konzorcija, ki pripomorejo k uresničitvi strategije CLARIN.SI. Ta de-

[14] https://www.sketchengine.eu/
[15] https://gitlab.clarin.si/clarinsi/webanno_tei
[16] https://nl.ijs.si/janes/
[17] https://nl.ijs.si/kas/
[18] https://github.com/clarinsi
[19] https://gitlab.clarin.si/
[20] http://clarin.si/services/web/
[21] https://cmc-corpora.org/ckcmc/

javnost je bila zelo odmevna in je tudi pomembno doprinesla k zanimanju za raziskave in razvoj jezikovnih virov med mladimi. Od leta 2018, ko smo z iniciativo začeli, je bilo uspešno izvedenih 19 projektov, v sklopu katerih so med drugim nastali korpus parlamentarnih razprav Državnega zbora Republike Slovenije siParl (Pančur et al., 2020), nadgradnja korpusa akademske slovenščine KAS 2.0 (Žagar et al., 2022) in govornega korpusa Gos Videolectures (Verdonik et al., 2019), orodje za učinkovito analizo slovenskih korpusov LIST (Krsnik et al., 2019) in drugi jezikovni viri in programska oprema. Med drugim je CLARIN.SI financiral tudi projekt »Razvoj učnega gradiva na korpusu siParl 2.0: Korpusni pristop k raziskovanju parlamentarnega diskurza« (Fišer in de Maiti, 2021).

5.3. Organizacija dogodkov

CLARIN.SI sodeluje pri organizaciji in izvedbi dogodkov s področja računalniškega jezikoslovja in sorodnih tematik v Sloveniji, npr. »XVIII EURALEX Intl. Congress« (Ljubljana, 2018) ali »22nd Intl. Conf. on Text, Speech and Dialogue« (Ljubljana, 2019), predvsem pa glavne konference za to področje v Sloveniji, »Jezikovne tehnologije in digitalna humanistika«, ki ima prek 20-letno tradicijo in z organizacijo katere je začelo društvo SDJT. SDJT od leta 2005 organizira občasna predavanja JOTA (Jezikovnotehnološki abonma), kjer je CLARIN.SI podprl snemanje in arhiviranje 12 predavanj na VideoLectures.NET[22], do sedaj z 10.000 ogledi.

5.4. Obveščanje in promocija

Nenazadnje, delovanje CLARIN.SI in njegovih središč znanja redno predstavljamo na domačih in tujih delavnicah in konferencah, kot so konferenca Evropskega strateškega foruma za raziskovalne infrastrukture (ESFRI), konference CLARIN idr., ter na predavanjih v okviru študijskih programov slovenskih univerz.

CLARIN.SI organizira tudi delavnice o uporabi korpusov in jezikovnih tehnologij za namene znanstvenih raziskav. Tako smo npr. izvedli delavnice[23] za uporabo konkordančnika noSketch Engine, platform WebAnno in Git, središče znanja CLASSLA pa je sodelovalo pri delavnici o uporabi korpusov za analizo regionalne variacije spolne zaznamovanosti jezika[24].

O dejavnostih partnerjev konzorcija CLARIN.SI in njegovih središč znanja javnost obveščamo tudi prek ažurnih novic, objavljenih na spletni strani infrastrukture, poštnega seznama ter objav s profila CLARIN.SI na Twitterju. Delo CLARIN.SI in njegovega središča znanja CLASSLA je bilo izpostavljeno tudi v več publikacijah »CLARIN ERIC Tour de CLARIN« (Fišer et al., 2019).

[22] https://videolectures.net/jota/
[23] https://www.clarin.si/info/dogodki/
[24] https://www.clarin.si/info/k-center/delavnice/

6. Vpetost v projekte in infrastrukture

CLARIN.SI je vpet v domače in evropske projekte, s čimer zagotavlja večjo izkoriščenost in vidljivost ter seveda tudi dodaten dotok sredstev za svoje delovanje.

6.1. Sredstva Evropske kohezijske politike

V okviru projekta kohezijskih sredstev MIZŠ 2018–2021 so partnerji konzorcija IJS, UM in UL nadgradili strojno opremo, s čimer je omogočeno hitrejše in proti okvaram odpornejše delovanje spletnih storitev CLARIN.SI, pridobljena gruča GPU strežnikov na Univerzi v Mariboru pa služi za raziskave globokega strojnega učenja obdelave jezikovnih podatkov, npr. na področju govora. S temi nadgradnjami lahko CLARIN.SI slovenski raziskovalni skupnosti zagotavlja odlično raziskovalno infrastrukturo, ki mdr. pripomore k privlačnosti slovenskih partnerjev v mednarodnih raziskovalnih in inovacijskih projektih ter podpira doseganje znanstvene odličnosti in mednarodno vrhunskih rezultatov. Tako npr. projekt EU MaCoCuuporablja gručo računalnikov CLARIN.SI za zajem in obdelavo spletnih velepodatkov, v okviru projekta EU InTaviapa se jezikoslovno označuje Slovenski biografski leksikon z modeli, razvitimi na gruči GPU. Več velikih projektov EU, kot sta ELEXIS in EMBEDDIA, je deponiralo razvite jezikovne vire v repozitorij CLARIN.SI.

6.2. Vpetost v evropske projekte

Med evropskimi projekti posebej izpostavimo ELEXIS,[25] saj je bila za potrebe tega projekta v repozitoriju CLARIN.SI narejena nova zbirka CLARIN.SI ELEXIS, v kateri so zbrani metapodatki in povezave do spletnih vmesnikov 143 digitalnih slovarjev. Ob koncu projekta ELEXIS v okviru CLARIN.SI oz. IJS načrtujemo tudi vzpostavitev novega centra znanja CLARIN za digitalno leksikografijo.

6.3. Vpetost v domače projekte

Sodelujemo tudi v več domačih projektih. Največji je »Razvoj slovenščine v digitalnem okolju«[26], ki mu CLARIN.SI zagotavlja svoje storitve za pregled in deponiranje v projektu izdelanih jezikovnih virov ter definicijo shem za usklajeno označevanje jezikovnih virov slovenskega jezika. V načrtu je tudi izdelava seznamov kontroliranih besedišč za jezikoslovno označevanje slovenskih besedil na ravni oblikoskladnje, skladnje, imenskih entitet, udeleženskih vlog itd.

6.4. Sodelovanje z drugimi RI

CLARIN.SI sodeluje s slovenskimi centri sestrskih infrastruktur CESSDA/ADP in DARIAH-SI. V projektu »RDA Node Slovenia« (2019–2020), ki ga je koordiniral ADP (FDV UL), smo pregledali in analizirali slovenske repozitorije raziskovalnih podatkov (Meden in Erjavec, 2021). Z INZ oz. DARIAH-SI pa smo sodelovali na področju standardizacije zapisa in izdelave korpusov parlamentarnih podatkov.

[25] https://elex.is/
[26] https://www.cjvt.si/rsdo/

6.5. Sodelovanje v delu CLARIN ERIC

CLARIN.SI je ena od aktivnejših nacionalnih RI v CLARIN ERIC. Pridobili smo sredstva za dva manjša projekta, ki sta vključevala mednarodni delavnici, in sicer 2016 v Ljubljani in 2019 v Amersfoortu. Slednja je bila v sodelovanju z DARIAH-SI posvečena izdelavi priporočil za standardizirano kodiranje korpusov parlamentarnih razprav z imenom Parla-CLARIN[27] (Erjavec in Pančur, V tisku), ki je postala priljubljena izbira za kodiranje parlamentarnih korpusov. Na tej osnovi je CLARIN.SI pridobil ključno vlogo v dveh večjih »CLARIN Flagship« projektih, ParlaMint I (2020–2021) in ParlaMint II (2022–2023).

Namen projektov ParlaMint je ustvariti primerljive, interpretativne in enotno kodirane korpuse parlamentarnih razprav. V že zaključenem projektu ParlaMint I je CLARIN.SI vodil zbiranje in kodiranje 17 korpusov nacionalnih parlamentov (Erjavec et al., 2022), ki so odprto dostopni na repozitoriju CLARIN.SI, kot tudi na konkordančnikih RI. V okviru projekta ParlaMint II, katerega namen je razširitev in obogatitev obstoječih korpusov ter dodajanje korpusov novih partnerjev, prav tako pa tudi razvoj izobraževalnih gradiv in primerov dobrih praks uporabe parlamentarnih korpusov za raziskave v humanistiki in družboslovju, člani CLARIN.SI vodijo štiri izmed petih delovnih sklopov[28].

Člani UO CLARIN.SI sodelujejo v delu CLARIN delovnih teles za pravna vprašanja (Mateja Jemec Tomazin, ZRC SAZU), za standardizacijo (Tomaž Erjavec, IJS) in za uporabniška vprašanja (Jakob Lenarčič, FF UL) ter na letnih konferencah CLARIN (T. Erjavec je predsednik programskega odbora konference v 2022 v Pragi). J. Lenarčič je prejel CLARIN »Stewen Krawer award« za mladega raziskovalca leta 2019, mdr. za svoje delo (skupaj z Darjo Fišer) pri vzpostavitvi iniciative »CLARIN Resource Families«[29], T. Erjavec pa je prejel »Steven Krauwer Award for CLARIN Achievements 2021« za svoje delo na projektu ParlaMint. Darja Fišer in Kristina Pahor de Maiti (FF UL) sta leta 2021 prejeli nagrado »Teaching with CLARIN Award« za najboljši učni material, povezan z uporabo virov CLARIN. Kaja Dobrovoljc (FF UL) je predstavila RI CLARIN na konferenci ob 20. obletnici ESFRI v Parizu leta 2022[30]. Darja Fišer je bila med leti 2016 in 2020 direktorica področja za uporabniška vprašanja, z letom 2023 pa naj bi postala generalna direktorica CLARIN ERIC.

7. Zaključki

CLARIN.SI je izjemno uspešno vzpostavljena infrastruktura, ki pokriva široko interdisciplinarno področje od humanističnih in družboslovnih raziskav do razvoja sistemov in tehnologij znanja in umetne inteligence. Podpira temeljne in aplikativne raziskave ter razvoj aplikacij, informacijskih sistemov in orodij na vseh ravneh tehnološke pripravljenosti.

Z izjemnim nacionalnim in regionalnim pomenom, spodbujanjem področja, pritegovanjem mladih, povezovanjem z industrijo ter široko vključenostjo deležnikov, močno vlogo pri uvajanju načel odprte znanosti ter izjemno odmevno in uspešno ter tudi nagrajeno vlogo na ravni evropskega in mednarodnega sodelovanja CLARIN.SI s povezanimi projekti predstavlja zgled za vzpostavitev uspešne vrhunske sodobne interdisciplinarne znanstveno-raziskovalno tehnološke infrastrukture.

V naslednjem obdobju bo CLARIN.SI poleg vzdrževanja obstoječih storitev še bolj intenzivno spodbujal ponovno uporabo raziskovalnih podatkov, s čimer bo omogočal raziskovalcem na področju humanistike in družboslovja povečanje produktivnosti, in kar je še pomembneje, vzpostavljanje novih raziskovalnih smeri, ki obravnavajo eno ali več družbenih vlog jezika. Drug pomemben cilj je izvajanje smernic za zagotavljanje interoperabilnosti CLARIN ERIC[31], ki je ključni predpogoj za učinkovito podporo raziskovalnemu delu skozi interoperabilnost orodij, virov, metapodatkov, standardov za zapis, kot tudi na organizacijski ravni (Jong et al., 2020). Hkrati bo treba okrepiti podporo uporabnikom, saj univerze in agencije od raziskovalcev v doktorskih in raziskovalnih programih vse intenzivneje zahtevajo načrte za ravnanje z raziskovalnimi podatki in njihovo trajno hrambo.

ESFRI kažipot 2021[32] za RI poudarja pomen podatkov FAIR, pri čemer smo na tem področju v okviru repozitorija CLARIN.SI storili že več pomembnih korakov, se pa bomo vidikom FAIR posvečali tudi naprej. Tako je v povezavi z RDA Node Slovenia v načrtu priprava delavnice o certifikaciji CTS in načelih FAIR za slovenske repozitorije raziskovalnih podatkov. ESFRI kažipot 2021 poudarja tudi vedno večjo prisotnost velepodatkov in pomembnost infrastruktur, da jih ustrezno hranijo in obdelujejo. Zaradi vedno večje količine dostopnih besedil, premika s pisnih na govorne in vizualne jezikovne vire ter vse bogatejšega označevanja besedil tudi na področju jezikovnih virov prehajamo v obdobje velepodatkov, kot se že sedaj izkazuje v projektih ParlaMint II, RSDO in MaCoCu. Zato bo CLARIN.SI v naslednjem obdobju podprl uporabo strojne in programske kapacitete za hrambo in predvsem obdelavo velepodatkov. Kažipot tudi poudarja pomembnost raziskovalnih infrastruktur za zajem, hrambo in obdelavo podatkov z družbenih omrežij in spleta. CLARIN.SI je že sedaj posvečal posebno pozornost takšnim jezikovnim virom, v prihodnosti pa bo te aktivnosti še okrepil, ne samo za slovenski, temveč (v okviru centra znanja CLASSLA in projekta MaCoCu) tudi za druge južnoslovanske jezike.

Kažipot poudarja tudi instrumentalizacijo in dostopnost podatkov ter storitev, pomembnih za posamezne skupnosti. Konzorcij CLARIN.SI trenutno vključuje

[27] https://clarin-eric.github.io/parla-clarin/
[28] https://www.clarin.eu/parlamint
[29] https://www.clarin.eu/resource-families
[30] https://www.esfri.eu/esfri-events/esfri-20years-conference
[31] https://www.clarin.eu/content/interoperability
[32] https://www.esfri.eu/esfri-roadmap-2021

12 članic, s čimer sicer pokriva veliko večino slovenskih deležnikov, ki bodisi proizvajajo ali uporabljajo jezikovne vire in tehnologije, ne pa vseh. V naslednjem obdobju se bo CLARIN.SI trudil razširiti svoj konzorcij, s čimer bo pokril tudi skupnosti potencialnih uporabnikov infrastrukture, ki do sedaj še niso bili zajeti v njeno delovanje. CLARIN.SI prav tako načrtuje študijo potreb posameznih skupnosti (raziskovalci in predavatelji s področja humanistike in s področja računalniške lingvistike, slovaropisci, prevajalci, osebe s posebnimi potrebami) in izboljšanje svoje ponudbe v skladu z ugotovitvami.

Kažipot med drugim poudarja pomen izobraževanja, šolanja in podpore pri uporabi infrastruktur za obstoječe in bodoče uporabnike. V prvem obdobju obstoja je bil CLARIN.SI izrazito kadrovsko podhranjen, a smo kljub temu izvedli vrsto dogodkov na delavnicah po Sloveniji in v tujini, predvsem na različnih fakultetah, kjer smo infrastrukturo predstavili študentom. V naslednjem obdobju se bomo načrtno lotili teh aktivnosti z bolj proaktivnim pristopom k izvedbi predavanj in delavnic tako za študente kot tudi za raziskovalce in predavatelje ter razvoju in promociji izobraževalnih materialov.

Nenazadnje je za prihodnost CLARIN.SI pomemben tudi pred kratkim sprejet Načrt razvoja raziskovalne infrastrukture 2030 (NRRI 2030)[33] v Sloveniji, ki ima »v načrtu nadaljevanje in krepitev dejavnosti še v okviru mednarodnih projektov CLARIN« (str. 60), priznava dosedanje sodelovanje z RI DARIAH-SI in CESSDA/ADP, ob tem pa predvideva tudi povezovanje z novima RI, in sicer OPERAS (Odprta znanstvena komunikacija v evropskem raziskovalnem prostoru za družboslovne in humanistične vede)[34], ki jo v Sloveniji vodi ZRC SAZU, in PRACE (Partnerstvo za napredno računalništvo v Evropi)[35], ki jo vodi ARNES.

Zahvala

Delo predstavljeno v prispevku so podprli ARRS v okviru financiranja raziskovalnih infrastruktur ESFRI, Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvo v okviru projektov C3330-19-952059 »Razvoj raziskovalne infrastrukture za mednarodno konkurenčnost slovenskega RRI prostora / RI-SI-CLARIN« in OP20.06780 »Razvoj slovenščine v digitalnem okolju« ter projekti CLARIN ERIC.

Zahvaljujemo se tudi sodelavcem CLARIAH-CZ za pomoč pri nadgradnjah in vzdrževanju platforme repozitorija, sodelavcem Češkega nacionalnega korpusa, predvsem Tomášu Macháleku, za pomoč pri instalaciji konkordančnika KonText in sodelavcem podjetja Lexical Computing, predvsem Janu Bušti in Tomášu Svobodi, za pomoč pri instalaciji konkordančnika Sketch Engine Crystal.

[33] https://www.gov.si/assets/ministrstva/MIZS/Dokumenti/ZNANOST/Novice/NRRI-2030/NRRI-2030_SLO.pdf

[34] https://www.operas-eu.org

[35] https://www.prace-ri.eu

8. Literatura

Jacob Devlin, Ming-Wei Chang, Kenton Lee in Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805. https://doi.org/10.48550/arXiv.1810.04805.

Tomaž Erjavec, Jan Jona Javoršek in Simon Krek. 2014. Raziskovalna infrastruktura CLARIN.SI. V: *Zbornik Devete konference JEZIKOVNE TEHNOLOGIJE*, Ljubljana, 9. - 10. oktober 2014. Slovensko društvo za jezikovne tehnologije. https://nl.ijs.si/isjt14/proceedings/isjt2014_03.pdf.

Tomaž Erjavec, Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Darja Fišer, Cyprian Laskowski in Katja Zupan. 2016. Annotating CLARIN.SI TEI corpora with WebAnno. V: *Proceedings of the CLARIN annual conference*. https://www.clarin.eu/sites/default/files/erjavec-etal-CLARIN2016_paper_17.pdf.

Tomaž Erjavec, Darja Fišer in Nikola Ljubešić. 2021. The KAS corpus of Slovenian academic writing. *Language Resources and Evaluation*, 55(2):551–583. https://rdcu.be/b7GrB.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx in Darja Fišer. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09574-0.

Tomaž Erjavec in Andrej Pančur. V tisku. The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings. *Journal of the Text Encoding Initiative*. https://journals.openedition.org/jtei/index.html.

Darja Fišer in Kristina Pahor de Maiti. 2021. "Prvič, sem političarka in ne politik, drugič pa...". *Contributions to Contemporary History*, 61(1). https://doi.org/10.51663/pnz.61.1.07.

Darja Fišer, Jakob Lenardič, Ilze Auziņa, Nan Bernstein Ratner, Koenraad De Smedt, Kaja Dobrovoljc, Réka Dodé, Rickard Domeij, Helge Dyvik, Tomaž Erjavec, Olga Gerassimenko, Jan Hajič, Michal Křen, Nikola Ljubešić, Brian MacWhinney, Monica Monachini, Beatrice Nava, Costanza Navarreta, Aneta Nedyalkova, Klaus Nielsen, Marin Noémi VadászLaak, Susanne Nylund Skog, Lene Offersgaard, Petya Osenova, Valeria Quochi, Sanita Reinsone, Inguna Skadiņa, Kiril Simov, Ondřej Tichý, Noémi Vadász, Tamás Váradi in Kadri Vider. 2019. *Tour de CLARIN Volume Two*. Zenodo. https://doi.org/10.5281/zenodo.3754164.

Darja Fišer, Nikola Ljubešić in Tomaž Erjavec. 2020.

The Janes project: Language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, 54:223–246. https://rdcu.be/7RX4.

Franciska De Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer in Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. V: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7–12, 2018. European Language Resources Association (ELRA). https://aclanthology.org/L18-1515.

Franciska De Jong, Bente Maegaard, Darja Fišer, Dieter Van Uytvanck in Andreas Witt. 2020. Interoperability in an infrastructure enabling multidisciplinary research: The case of CLARIN. V: *Proceedings of the 12th Language Resources and Evaluation Conference*, str. 3406–3413. European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.417/.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý in Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography*, 1:7–36.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban in Igor Mozetič. 2015. *Emoji Sentiment Ranking 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1048.

Simon Krek. 2022. Delivrable D1.31: Report on the Slovenian Language. Tehnično poročilo, European Language Equality Project. https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_31__Language_Report_Slovenian_.pdf.

Luka Krsnik, Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Aleksander Ključevšek, Simon Krek in Marko Robnik-Šikonja. 2019. Corpus extraction tool LIST 1.2. http://hdl.handle.net/11356/1276.

Jakob Lenardič in Darja Fišer. 2022. CLARIN Depositing Guidelines: State of Affairs and Proposals for Improvement. V: *Proceedings of the CLARIN Annual Conference*, Prague, Czech Republic, October 10–12, 2022. https://www.clarin.eu/event/2022/clarin-annual-conference-2022.

Nikola Ljubešić, Filip Markoski, Elena Markoska in Tomaž Erjavec. 2021. *Comparable corpora of South-Slavic Wikipedias CLASSLA-Wikipedia 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1427.

Tomáš Machálek. 2020. KonText: Advanced and Flexible Corpus Query Interface. V: *Proceedings of the 12th Language Resources and Evaluation Conference*, str. 7003–7008, Marseille, France, May. European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.865.

Katja Meden in Tomaž Erjavec. 2021. Pregled slovenskih repozitorijev raziskovalnih podatkov. Tehnično poročilo, Jožef Stefan Institute. https://www.clarin.si/info/services/projects/#RDA_Node_Slovenia.

Andrej Pančur, Tomaž Erjavec, Mihael Ojsteršek, Mojca Šorn in Neja Blaj Hribar. 2020. Slovenian parliamentary corpus (1990-2018) siParl 2.0. http://hdl.handle.net/11356/1300.

Pavel Rychlý. 2007. Manatee/Bonito - A Modular Corpus Manager. V: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, str. 65–70, Brno. Masarykova univerzita.

Matej Ulčar in Marko Robnik-Šikonja. 2021. *Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1397.

Tamás Váradi, Steven Krauwer, Peter Wittenburg, Martin Wynne in Kimmo Koskenniemi. 2008. CLARIN: Common language resources and technology infrastructure. V: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/317_paper.pdf.

Darinka Verdonik, Tomaž Potočnik, Mirjam Sepesy Maučec, Tomaž Erjavec, Simona Majhenič in Andrej Žgank. 2019. Spoken corpus Gos VideoLectures 4.0 (transcription). http://hdl.handle.net/11356/1223.

Špela Vintar in Matej Martinc. 2022. Framing karstology: From definitions to knowledge structures and automatic frame population. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 28(1):129–156.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho in Chris Biemann. 2013. WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. V: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, str. 1–6, Sofia, Bulgaria, August. Association for Computational Linguistics. https://aclanthology.org/P13-4001.

Aleš Žagar, Matic Kavaš, Marko Robnik-Šikonja, Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Marko Ferme, Mladen Borovič, Borko Bošković, Milan Ojsteršek in Goran Hrovat. 2022. Corpus of academic Slovene KAS 2.0. http://hdl.handle.net/11356/1448.