Številka: 03/2021
Datum: 27. 9. 2021

KONČNO POROČILO O IZVAJANJU OPERACIJE RI-SI-CLARIN

Naziv operacije / akronim: Razvoj raziskovalne infrastrukture za mednarodno konkurenčnost slovenskega RRI prostora / RI-SI-CLARIN
Šifra operacije (IS e-MA): OP20.04765
Številka pogodbe: C3330-19-952059
Naziv upravičenca: Institut »Jožef Stefan«

Kazalo

Povzetek ... 1
Predstavitev projekta ... 2
Doseženi cilji ... 3
Doseganje zastavljenih ciljev, kot so bili predvideni v okviru sklenjene konzorcijske pogodbe ... 7
Skladnost s predmetom, cilji in nameni NPO ... 9
Realizacija doseganja kazalnikov ... 10
Način dostopanja in opis zasedenosti opreme ... 12
Institut »Jožef Stefan« ... 13
Univerza v Ljubljani ... 13
Univerza v Mariboru ... 13
Načrti za prihodnost ... 14
Dodatek 1. Seznam vnosov v repozitorij CLARIN.SI 1.1.2020–1.9.2021 ... 16

Povzetek

CLARIN.SI je slovenska nacionalna infrastruktura za jezikovne vire in tehnologije, in je članica evropske raziskovalne infrastrukture CLARIN ERIC, v kateri je združenih 20 držav. CLARIN.SI domuje na Institutu »Jožef Stefan« in je organiziran kot konzorcij 12 partnerjev, ki mdr. vključuje vse slovenske univerze. Namen CLARIN(.SI) je, da raziskovalcem na področju humanistike, družboslovja in drugih, z jezikom povezanih ved zagotavlja dostop do in varno deponiranje jezikovnih virov in tehnologij ter strokovno podporo in prenos znanja. CLARIN.SI za obratovanje potrebuje sodobno in visokozmogljivo računalniško opremo, saj na svojih strežnikih hrani velike jezikovne vire in ponuja raznovrstna spletna orodja za dostop in analizo obsežnih in bogato označenih jezikovnih korpusov in slovarjev.

Namen operacije je bil posodobiti in nagraditi strojno opremo treh članic konzorcija CLARIN.SI, in sicer Instituta »Jožef Stefan« (IJS), Univerze v Ljubljani (UL) in Univerze v Mariboru (UM). IJS je v teku operacije obnovil ter nadgradili svojo strojno opremo z dvema gručama visokozmogljivih računalnikov s pripadajočo opremo, s čimer je omogočil hitrejše in proti okvaram odpornejše delovanje spletnih storitev CLARIN.SI, predvsem repozitorijske platforme, spletnih konkordančnikov in storitev za avtomatsko jezikoslovno označevanje besedil. UL je pridobila visokozmogljivi strežnik za hranjenje in dostop do jezikovnih virov, katerih skrbnik je infrastrukturni Center za jezikovne vire in tehnologije Univerze v Ljubljani. UM je teku operacije pridobila gručo GPU strežnikov, ki služi za raziskave globokega strojnega učenja obdelave jezikovnih podatkov s strani članov konzorcija CLARIN.SI, ter visokozmogljive strežnike za obdelavo velikih jezikovnih podatkov, kot tudi financiranje tehničnega sodelavca zadolženega za vzdrževanje te opreme.

S temi nadgraditvami je lahko CLARIN.SI slovenski raziskovalni skupnosti zagotovil odlično raziskovalno infrastrukturo, ki mdr. pripomore k privlačnosti slovenskih partnerjev v mednarodnih raziskovalnih in inovacijskih projektih, kot tudi motivira študente k nadaljevanju kariere raziskovalca na področjih raziskovanja in obdelave slovenskega jezika.

Predstavitev projekta

Ministrstvo za izobraževanje, znanost in šport je, kot sofinancer in posredniški organ, dne 24.10.2018 posredovalo Poziv za oddajo vloge za projekt »Razvoj raziskovalne infrastrukture za mednarodno konkurenčnost slovenskega RRI prostora – RI-SI«, št. poziva 5442-200/2016/42, vsem, ki so skladno z metodologijo European Strategy Forum on Research Infrastructures (ESFRI) ob oddaji vloge na evropski ravni dosegli status »implementirani projekt. Tovrstni projekti so usmerjeni v raziskave in razvoj ter sledijo doseganju ciljev prednostne osi 1 »Mednarodna konkurenčnost raziskav, inovacij in tehnološkega razvoja v skladu s pametno specializacijo za večjo konkurenčnost in ozelenitev gospodarstva« iz Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014 – 2020, kjer so vlaganja v raziskovalno infrastrukturo ukrep, ki je namenjen krepitvi infrastrukture za raziskave, inovacije ter zmogljivost za razvoj odličnosti na tem področju in spodbujanju pristojnih centrov, zlasti takšnih, ki so evropskega pomena. Med tovrstne projekte tako spada tudi Evropska raziskovalna infrastruktura CLARIN, ki je namenjena zagotavljanju enostavnega, prostega in odprtega dostopa do obsežnih jezikovnih virov, orodij in storitev za raziskovanje in izobraževanje v humanistiki in družboslovju, kot tudi raziskavam in razvoju jezikovnih tehnologij. Zajema jezike držav članic EU in jezike, ki jih v državah članicah poučujejo ali so pomembni zaradi migracijskih tokov.

Eden izmed temeljnih ciljev CLARIN-a je delovanje platforme za stabilen dostop do repozitorija jezikovnih virov, ki mora ustrezati zahtevnim merilom, ki jih postavlja evropski CLARIN ERIC. Repozitorij CLARIN.SI te pogoje izpolnjuje in je že ob začetku operacije imel status centra CLARIN B (nacionalnega centra) in vseboval prek 100 odprtodostopnih jezikovnih virov. CLARIN.SI je ponujal tudi storitve za korpusno jezikoslovje in jezikovne tehnologije, na prvem mestu spletni konkordančniki, v 2019 z bazo prek 40 označenih korpusov. Podobno tudi drugi konzorcijski partnerji CLARIN.SI ponujajo prostodostopne vire slovenskega jezika na svojih strežnikih, predvsem Center za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT UL).

CLARIN je (in še vedno) uspešno sodeluje z drugima dvema raziskovalnima infrastrukturama ESFRI s področja humanistike in družboslovja v Sloveniji, tj. DARIAH (oz. SI-DIH) in CESSDA (oz. ADP). CLARIN.SI je bil tudi ustanovni član slovenske iniciative SLING, Slovenskega

nacionalnega superračunalniškega omrežja, z namenom, da bi lahko dolgoročno čimbolj učinkovito in fleksibilno reševali dodatne potrebe po procesnih in spominskih kapacitetah. Skozi članico konzorcija Univerzo v Mariboru je sodeloval CLARIN.SI tudi v projektu HPC RIVR, v okviru katerega sta bila vzpostavljena dva superračunalnika, ki se uporabljata tudi za obdelavo velikih količin besedil in izvajanje računsko zahtevnih metod strojnega učenja.

Nova raziskovalna oprema, pridobitev katere je bil namen operacije, naj bi omogočila članom konzorcija izvajanje najsodobnejših in računsko zahtevnih metod strojnega učenja in obdelavo velikih količin besedil. S tem naj bi CLARIN.SI slovenski raziskovalni skupnosti zagotovil odlično raziskovalno infrastrukturo, ki bo pripomogla k privlačnosti slovenskih partnerjev v mednarodnih raziskovalnih in inovacijskih projektih, kot tudi motivirala študente k nadaljevanju kariere raziskovalca.

Celotno trajanje investicijskega projekta oz. izvajanje projektnih aktivnosti in obdobje upravičenosti nastanka stroškov sta bila po Pozivu opredeljena na obdobje od 1.6.2018 do 31.8.2021. Obdobje upravičenosti izdatkov za nastale stroške (datum plačila računov oz. verodostojnih knjigovodskih listin) pa je opredeljeno od 1.6.2018 do 30. 9. 2021.

Splošni namen operacije je torej krepitev nacionalnih nosilnih znanstvenih partnerskih institucij v posameznem infrastrukturnem projektu z doseženim statusom »implementirani projekt«. Slednji so vezani tako na prednostna področja NRRI kot prednostna področja Slovenske Strategije Pametne Specializacije (v nadaljevanju: S4) in s tem prispevajo k uresničevanju ciljev Resolucije o raziskovalni in inovacijski strategiji Slovenije 2011-2020 (v nadaljevanju: RISS 2011-2020), ki predvideva vključitev v mednarodne projekte na področju raziskovalne infrastrukture. Podpora oz. sofinanciranje tem infrastrukturnim projektom bo tako omogočila bistveno hitrejše, intenzivnejše in visoko kakovostno vključevanje v velike mednarodne projekte in s tem h krepitvi mreže raziskovalne infrastrukture, človeških virov v znanosti, omogočanju prostega pretoka ljudi, zamisli in znanja v evropskem raziskovalnem prostoru.

Investicijski projekt je imel dva specifična namena:

1. Obnoviti in nagraditi strojno opremo, ki je nujna za delovanje spletnih storitev infrastrukture, ob tem pa je bila ob začetku operacije bodisi že zastarana, bodisi bi bila zastarana do leta 2021. Nova strojna oprema naj bi imela tudi večje kapacitete kot tista z začetka operacije, saj se število in velikost jezikovnih virov neprestano veča. S tem bo CLARIN.SI omogočili optimizacijo uporabe raziskovalne infrastrukture in še naprej omogočal njeno dostopnost zunanjim uporabnikom.
2. Opremiti CLARIN.SI z delovnimi postajami, ki bodo omogočile članom konzorcija izvajanje najsodobnejših in računsko zahtevnih metod strojnega učenja in obdelavo velikih količin besedil. S tem bi CLARIN.SI slovenski raziskovalni skupnosti zagotovil odlično raziskovalno infrastrukturo, ki bo pripomogla k privlačnosti slovenskih partnerjev v mednarodnih raziskovalnih in inovacijskih projektih, kot tudi motivirala študente k nadaljevanju kariere raziskovalca.

Doseženi cilji

Projekt je imel pet splošnih oz. razvojnih ciljev:

Prvič, **slovenski raziskovalni skupnosti omogočiti dostop do najnovejše raziskovalne infrastrukture**, ki podpira visoko kakovostne raziskave in inovacije pri reševanju velikih sodobnih družbenih izzivov. Na področju računalniškega jezikoslovja in digitalne humanistike

namreč v zadnjem desetletju prihaja do potenciranega razvoja področja ter vzpostavljanja povezav z vedno več drugimi raziskovalnimi in družbenimi področji; orodja in algoritmi ter njihova uporaba v vsakdanjih aplikacijah ter pri obdelavi velepodatkov, pri umetni inteligenci in novih komunikacijskih omrežjih pa povezujejo področje z vlaganjem v velike mednarodne infrastrukture ter ga vključujejo v pomembne družbene spremembe in reševanje velikih izzivov, ki so povezani z digitalizacijo jezika. Vložek v raziskovalno opremo naj bi omogočil slovenskim raziskovalcem enakovredno vključevanje v te tokove. Ta cilj je bil dosežen, saj je dobra opremljenost raziskovalne infrastrukture CLARIN.SI npr. omogočila, da je CLARIN.SI postal zahtevani repozitorij za veliki kohezijski projekt Ministrstva za kulturo RS Slovenije »Razvoj slovenščine v digitalnem okolju« (RSDO), ki je oz. bo izdelal odprto dostopne velike in bogato označene vire slovenskega jezika in tehnologije za njegovo obdelavo, v projekt pa so vključeni tudi vsi partnerji operacije. Oprema je tudi omogočila deponiranje prek 100 novih jezikovnih virov, ki jih navajamo v Prilogi 1.

V okviru operacije pridobljena GPU gruča na UM omogočila računsko izredno zahtevno strojno učenje globokih modelov slovenskega jezika za namene več projektov, mdr. projekta RSDO.

Drugi cilj je bil **okrepiti in koordinirati nacionalno raziskovalno infrastrukturo** na področjih, na katerih ima Slovenija odlične aktivne raziskovalne skupine. Slovenija ima dolgo in plodno tradicijo na področju računalniškega jezikoslovja ter povezanih disciplin, npr. tehnologijah znanja, strojnega učenja, zajemanja in upravljanja s podatki in znanjem ipd. V preteklosti tega potenciala pogosto niso mogli ustrezno realizirati v kontekstu mednarodno konkurenčnih objav in projektov zaradi pomanjkanja infrastrukture in vpetosti v mednarodna sodelovanja. Tudi ta cilj je bil izpolnjen. Uporaba GPU gruče na UM, je že omogočila izdelavo globokih modelov za obdelavo slovenskega jezika, s katerimi slovenski raziskovalci lahko enakovredno sodelujejo pri mednarodnih projektih, ki vključujejo obdelavo naravnega jezika. Tako npr. ZRC SAZU sodeluje v projektu EU INTAVIA »In/Tangible European Heritage Visual Analysis, Curation & Communication« v sklopu katerega se jezikoslovno označuje Slovenski biografski leksikon z modeli razvitimi na GPU UM. Kot drugi primer izpostavimo projekt EU MACOCU »Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages« na katerem sodeluje IJS, in v sklopu katerega se ena od gruč računalnikov, pridobljenih v okviru operacije na IJS, uporablja za zajem in obdelavo velepodatkov zajetih s spleta.

Tretji cilj je bil omogočiti **optimizacijo uporabe raziskovalne infrastrukture in jo naredili bolje dostopno zunanjim uporabnikom**. Za zagotavljanje dostopnosti, uporabniške prijaznosti in podpore je potrebno dodatno vlaganje sredstev in raziskovalnega časa, ki v visoko konkurenčnih razmerah raziskovalnega dela pogosto ni opravičljivo, saj bi zmanjšalo raziskovalne rezultate v okviru obstoječih raziskovalnih ur ter infrastrukturnih kapacitet. S tem, da v projektu zagotavljamo ustrezne kapacitete, bodo omogočili optimizacijo in večjo dostopnost, kar ne pomeni le boljšega izkoristka raziskovalnih infrastruktur, ampak tudi večjo vidnost ter uporabnost za več institucij, vključno z visokim in potencialno srednjim šolstvom ter širšo strokovno javnostjo. Tretji cilj je bil, koliko je to že mogoče ugotoviti, tudi dosežen, saj so storitve infrastrukture CLARIN.SI sedaj bolje dostopne, po eni strani zato, ker je čas, ki mine od poizvedbe po velikih korpusih prek spletnih konkordančnikov infrastrukture, bistveno manjši, s čimer so storitve bolj odzivne in s tem prijaznejše za uporabo, po drugi pa so storitve bolj odporne na izpade, saj smo z novo strojno opremo dobili možnost redundantnih storitev, s čimer je manjša možnost dolgotrajnih izpadov storitev. Dodatno smo v trajanju operacije

bistveno nadgradili spletno dokumentacijo repozitorija CLARIN.SI, vključno s podrobnim opisom zahtevanih metapodatkov in formata podatkov, s čimer smo raziskovalcem olajšali deponiranje njihovih virov in jih s tem naredili bolj dostopne in obvarovali pred izgubo in tehnološkim zastaranjem. Vsa navodila so tudi na voljo tako v angleškem jeziku, s čimer so storitve razumljive tudi mednarodni skupnosti, kot v slovenščini, za slovenske raziskovalce in študente. Storitve CLARIN.SI se tudi že uporabljajo pri visokošolskem izobraževanju, tako npr. študenti višjih letnikov dodiplomskega ali podiplomskega študija na Fakulteti za računalništvo in informatikov UL že redno deponirajo jezikovne vire, ki so jih izdelali v okviru diplomskih ali magistrskih nalog v repozitorij CLARIN.SI, študenti na jezikovnih smereh Filozofske fakultete UL in Filozofske fakultete UM pa uporabljajo konkordančnike CLARIN.SI (vključno s konkordančniki CJVT UL, ki je bil partner v projektu) pri svojem učnem procesu.

Četrti cilj je bil **slovenskim raziskovalnim skupnostim zagotoviti odlično raziskovalno infrastrukturo, ki bo pripomogla k privlačnosti slovenskih partnerjev v mednarodnih raziskovalnih in inovacijskih projektih, zagotovila zaposlovanje vrhunskih raziskovalcev in motivirala študente k nadaljevanju kariere raziskovalca**. Zaradi zmanjšanja investicij ter poslabšanja javne podobe raziskovalne dejavnosti v zadnjem desetletju grozi raziskovalnemu sektorju pomanjkanje vrhunskih kadrov, ki jih hkrati z boljšimi pogoji zaposlovanja ter infrastrukture privabljajo tudi tuje razvojno-raziskovalne organizacije in podjetja. Z ustrezno infrastrukturo in vpetostjo v mednarodne raziskovalne in inovacijske projekte želijo ta trend ustaviti ali celo obrniti ter zagotoviti, da se odlični centri znanja, ki smo jih ohranili, še razvijejo in oplajajo z mednarodnim sodelovanjem. Tudi ta cilj se je začel izvajati že v teku operacije, predvsem skozi inštrument mladih raziskovalcev, projekta RSDO in mednarodnih (EU) projektov, skozi katere je partnerjem v operaciji uspelo pridobiti več odličnih mladih raziskovalcev oz. raziskovalcev, ki so po eni strani vpeti v zanimive raziskovalno-razvojne projekte, po drugi pa so vpisani na doktorski študij, vsi pa uporabljajo raziskovalno infrastrukturo CLARIN.SI, bodisi, da so vpeti v njeno vzdrževanje, vanjo deponirajo izdelane jezikovne vire, ali pa izkoriščajo UM GPU za raziskave globokega strojnega učenja.

Peti cilj je bil **zagotoviti podporo slovenskemu sodelovanju v prihajajočih mednarodnih raziskovalnih infrastrukturah**. Za tvorno in uspešno sodelovanje v projektnih prijavah, raziskovalno-razvojnih tekmovanjih ter skupnih evropskih in mednarodnih raziskovalnih ter infrastrukturnih projektih bodo v okviru operacije zagotovili ustrezne kapacitete, standarde ter protokole, da bodo lahko slovenske organizacije in raziskovalci nastopali enakovredno ali celo s pozicije tehnološke in infrastrukturne odličnosti, hkrati pa bodo podatkovni viri, orodja ter objavljene publikacije v okviru operacije in povezanih infrastruktur predstavljali pomembno osnovo in okolje za konkurenčno delo in vključevanje v mednarodne raziskovalne infrastrukture. Na tem, dolgoročnem cilju zaenkrat še ne moremo poročati o konkretnih uspehih, saj od začetka operacije do sedaj še ni bilo razpisov za nove mednarodne raziskovalne infrastrukture s področja jezikovnih virov in tehnologij, v katere bi se CLARIN.SI lahko vključil. Vendar pa dva partnerja operacije (UL in IJS) sodelujeta (IJS kot vodja projekta) pri EU projektu ELEXIS »European Lexicographic Infrastructure«, namen katerega je vzpostaviti pogoje za infrastrukturo namenjeno slovaropisju. V sklopu tega projekta so ključne nove kapacitete CLARIN.SI, predvsem na CJVT UL. Poleg tega je CLARIN.SI 2019 pridobil, skupaj z bolgarskim CLARIN, status CLARIN K-centre, torej centra znanja z imenom CLASSLA (CLARIN Knowledge Centre for South Slavic Languages), s čimer je postal referenčna točka za računalniško obdelavo južnoslovanskih jezikov, s čimer si je bistveno povečal mednarodno prepoznavnost, saj s tem pokriva tudi jezike bivše Jugoslavije in bolgarski jezik.

Operacija je imela tudi pet specifičnih ciljev:

Prvi cilj je bil **zagotoviti nadaljnje delovanje tehničnih storitev infrastrukture s posodobitvijo njene strojne opreme**. Ta cilj je bil v celoti dosežen, saj tehnične storitve CLARIN.SI brezhibno delujejo, kar je mdr. omogočilo repozitoriju CLARIN.SI pridobitev certifikata CTS »Core Trust Seal«, s čimer je postal drugi repozitorij v Sloveniji (prvi je ADP »Arhiv družboslovnih podatkov« na FDV UL, ki je del infrastrukture CESSDA), ki je pridobil certifikat zaupanja vrednega repozitorija.

Drugi specifični cilj je bil **omogočiti hranjenje velikih multimodalnih jezikovnih podatkov v infrastrukturi** z nadgradnjo strojne opreme z novim strežnikom, ki vsebuje hitre in velike diskovne kapacitete, ter z omogočanjem izdelave rednih varnostnih kopij vseh deponiranih podatkov. Tudi ta cilj je bil v celoti dosežen, saj ima repozitorij, kot tudi druge spletne storitve CLARIN.SI dovolj kapacitet ne samo za hranjenje že deponiranih podatkov (število teh se je v trajanju operacije podvojilo) temveč tudi zelo velikih multimodalnih podatkov, ki bodo deponirani v repozitorij CLARIN.SI v okviru projekta RSDO v obdobju 2021-2023.

Tretji cilj je bil omogočiti, da **CLARIN.SI sledi paradigmi »velepodatkov« (big data)** z nakupom zmogljivega strežnika in velikega diskovnega polja, ki bo omogočal razvoj delotokov za obdelavo jezikovnih podatkovnih zbirk na način velepodatkov ter bo integriran v nacionalno superračunalniško infrastrukturo. Ta cilj je bil v celoti dosežen, saj nam strežnik služi za shranjevanje velikih količin jezikovno označenih besedil, njihovih n-gramov in odprto dostopnih pomenskih zbirk. Strežniška infrastruktura je povezana s superračunalnikom Maister, ki se uporablja za izvajanje raziskav na področju ugotavljanja pomena in ekstrakcije terminologije ter znanja iz besedil v slovenskem in drugih jezikih. Možna pa bo uporaba velepodatkovnih jezikovnih virov tudi na superračunalniku Vega, kar nam bo zelo skrajšalo statistične obdelave velike količine besedil. Četrti cilj je bil omogočiti, da **CLARIN.SI ponuja javno dostopne spletne storitve obdelav velikih količin slovenskih besedil** s pomočjo nakupa ustreznih strežnikov. Tudi ta cilj je bil dosežen, saj so spletne storitve CLARIN.SI (konkordančniki, spletno označevanje besedil) sedaj bistveno odzivnejše, možno pa je tudi ponujati in obdelovati bistveno večje količine podatkov. Tako je npr. v okviru projekta RSDO predvidena izdelava in spletna dostopnost prek konkordančnikov korpusa, ki bo združeval vse obstoječe večje korpuse slovenskega jezika, z velikostjo prek treh milijard besed, pri čemer so testi pokazali, da bo sedaj možna obdelava tako velike količine besedil.

Peti cilj je bil **vzpostaviti namensko gručo GPU računalnikov**. Ta cilj je bil dosežen s nakupom strežnika DGX-1 proizvajalca NVIDIA ter z nakupom strežnika Supermicro. Dostop do vzpostavljene GPU gruče je omogočen osebam, ki se uvrstijo na seznam uporabnikov. Na ta seznam se lahko uvrstijo zaposleni pri projektnih partnerjih (UM, IJS in UL), njihovi magistrski in doktorski študenti ter gostujoči raziskovalci, ki bodo opremo uporabljali v namene, skladne s projektom RI-SI CLARIN. Uporabniki, ki jim je odobrena uporaba sistema, imajo omogočen dostop do GPU gruče prek sistema SLURM, ki omogoča upravljanje z viri in razvrščanje nalog v vrsto glede na parametre, ki jih določi uporabnik. Gre za sistem, pri katerem uporabnik kvantitativno določi vire, ki jih za izvajanje neke naloge potrebuje. Naloge se uvrstijo v vrsto. Ko so na voljo specificirani viri, se naloga izvede. GPU gruča ima ob zaključku projekta 21 aktivnih uporabnikov, raziskovalcev in mladih raziskovalcev iz vseh treh projektnih partneric, ki s pomočjo GPU gruče izvajajo projekte na najrazličnejših področjih jezikovnih tehnologij, od označevanja večjih korpusov slovenskega jezika prek raziskav na področju jezikovnih in govornih tehnologij, strojnega prevajanja, procesiranja emocionalnega

govora do izgradnje vektorskih vložitev in nevronskih modelov strojnega učenja, razvoja strojnih modelov za postavljanje ločil ipd.

Doseganje zastavljenih ciljev, predvidenih v okviru konzorcijske pogodbe

Konzorcijska pogodba je v 2. členu (odgovornost konzorcijskih partnerjev) pooblastila IJS, da kot upravičenec in poslovodeči konzorcijski partner ter vodja konzorcija v imenu konzorcija z ministrstvom sklene pogodbo o sofinanciranju operacije, da konzorcij zastopa pri vseh opravilih z ministrstvom, ki so povezana z izvedbo operacije, da odgovarja za vnos podatkov vseh svojih konzorcijskih partnerjev v aplikacije, predpisane s strani ministrstva, ter da odgovarja za pravilnost vnesenih podatkov, kar je Institut »Jožef Stefan« v teku operacije tudi storil. Konzorcijski partnerji so pri izvajanju pogodbenih obveznosti tudi upoštevali predpise in navodila, ki so bila navedena v pozivu za oddajo vloge za projekt.

Konzorcijska pogodba je v 3. členu (delovanje konzorcija in naloge) definirala konzorcij (IJS, UM, UL) in definirala osnovne naloge projekta razdeljene po partnerjih, vendar z namenom vzpostavljanja skupnih virov, ki bodo v skladu z merili evropskega CLARIN ERIC na voljo vsej slovenski (in širši) raziskovalni skupnosti:

- IJS: nadgradnja in vzdrževanje repozitorija jezikovih virov, ki ustreza merilom evropskega CLARIN ERIC in odprtega dostopa do znanstvenih virov, ter storitve za korpusno jezikoslovje in jezikovne tehnologije;
- UM: nadaljevanje razvoja javno dostopnih virov za repozitorij slovenskih govornih virov in zagotavljanje namenske izrabe svoje infrastrukture ter dostopnost in stalno razpoložljivost računskih kapacitet s pomočjo splošno sprejetih metod dostopa na ustrezen način;
- UL: omogočanje dostopa do referenčnih korpusov slovenskega jezika ter dostop do slovarskih baz in obstoječih ter novih pripadajočih orodij.

Zgornji cilji so bili v celoti izvedeni.

Nadalje je konzorcijska pogodba v tem členu določila, da bi imel vsak konzorcijski partner lastninsko pravico na strojni opremi, pridobljeni v okviru te operacije, ki je postavljena v prostorih ali na nepremičninah, ki so v lasti posameznega konzorcijskega partnerja ter, da bo za vzdržnost in delovanje vzpostavljene infrastrukture med trajanjem operacije in tudi po njenem zaključku je odgovoren imetnik lastninske pravice, pri čemer so bili posamezni partnerji dolžni vse člane konzorcija operacije obveščati o morebitnih spremembah ter o aktivnostih te operacije, ki bi lahko vplivale na delo ostalih partnerjev. Tudi te postavke so konzorcijski partnerji v celoti upoštevali.

Delitev del je bila dogovorjena po naslednjem načrtu:

- IJS: izdelava in potrditev investicijske dokumentacije; izdelava razpisa, nakup, dobava in montaža 1. gruče za spletne storitve (2019), 2. strežnika repozitorija (2019), 3. diskovnega polja za varnostne kopije (2019), 4. stikala za optični kanal (2019), 5. obnove opreme www.slovenscina.eu (2020), 6. obnove obstoječe opreme (2020).
- UM: izdelava in potrditev investicijske dokumentacije; izdelava razpisa, nakup, dobava in montaža 1. gruče GPU strežnikov (2019), 2. opreme UPS (2019), 3. strežnikov za obdelavo velikih jezikovnih podatkov (2019), 4. diskovnega polja za hranjenje velikih

podatkov (2019), 5. obnovo GPU gruče (2021), 6. vzdrževanje raziskovalne opreme (2019-2021);
- UL: izdelava in potrditev investicijske dokumentacije; izdelava razpisa, nakup, dobava in montaža 1. strežnika »viri.cjvt.si« (2019).

Načrt je bil v celot izveden, z izjemo točke 5. pri IJS, kjer je bila iz upravičenih razlogov in ob strinjanju MIZŠ obnova opreme www.slovenscina.eu spremenjena v nakup dodatnih strežnikov za izvajanje nalog spletnih storitev CLARIN.SI na IJS.

V točki 3. je bilo definirano tudi vključevanje konzorcijskih partnerjev v upravljavsko strukturo, kjer je bila predvidena ustanovitev dveh organov konzorcija, in sicer Upravnega odbora konzorcija ter Strokovnega sveta konzorcija.

Upravni odbor konzorcija je najvišji organ upravljanja konzorcija. Vsaka pogodbena stranka v Upravni odbor konzorcija imenuje enega člana in enega nadomestnega člana. Upravni odbor konzorcija vodi predsednik upravnega odbora, ki ga imenuje poslovodeči konzorcijski partner. V upravni odbor so bili imenovani naslednji člani:

- IJS
  - Predsednik, član: Tomaž Erjavec
  - Namestnik: Nikola Ljubešić
- UM
  - Član: Darinka Verdonik
  - Namestnik: Matej Rojc
- UL
  - Član: Simon Krek
  - Namestnik: Nataša Logar

Upravni odbor se je dvakrat sestal fizično, nato pa so zaradi pandemije COVID-19 seje potekale dopisno. Ker bistvenih sprememb v izvajanju operacije glede na načrtovano ni bilo, upravnemu odboru tudi ni bilo potrebno sprejemati za operacijo bistvenih odločitev. Najbolj kompleksno je bilo definiranje režima in spremljanja uporabe GPU gruče UM, saj je bila ta na voljo vsem partnerjem operacije. Partnerji so napisali in podpisali sporazum o uporabi gruče GPU, s katerim so definirali, kdo in kako ima pravico uporabe gruče, kako se prijavijo novi uporabniki itd. in se tega sporazuma potem tudi držali v teku izvajanja operacije. Sporov pri uporabi GPU ali sicer ni bilo.

Strokovni odbor konzorcija so sestavljali člani konzorcija CLARIN.SI in je bil zadolžen za pripravo znanstveno-strokovnih izhodišč in priporočil s področja delovanja raziskovalne infrastrukture CLARIN. Strokovni odbor konzorcija je sestavljen iz vsaj petih članov, ki so jih predlagali člani in članice upravnega odbora konzorcija CLARIN.SI. Strokovni odbor konzorcija vodi predsednik odbora, ki ga izmed sebe izvolijo člani Strokovnega odbora, imenuje pa ga Upravni odbor konzorcija. Zasedba strokovnega odbora je bila sledeča:

- Predsednik: Jan Jona Javoršek, CMI IJS
- Člani:
  - Darja Fišer, FF UL
  - Bojan Klemenc, FRI UL
  - Iztok Kosem, Trojina
  - Izidor Mlakar, FERI UM
  - Milan Ojsteršek, FERI UM

- Marko Robnik Šikonja, FRI UL

Za upravljanje z raziskovalno infrastrukturo znotraj konzorcija in z vsemi potencialnimi uporabniki je bilo določeno, da bodo spletne storitve in dostop do jezikovnih virov, ki jih ta raziskovalna oprema omogoča, lahko brezplačno in v skladu s sprejetimi pravili dobre prakse in priporočili organov konzorcija uporabljajo vsi zainteresirani uporabniki oz. (odvisno od dostopnosti posameznega vira znotraj storitve) raziskovalci, ki se prijavijo v spletno storitev prek svojega EduGain uporabniškega imena in se strinjajo s pogoji uporabe vira, kar se je striktno izvajalo.

Za raziskovalno opremo, ki je neposredno namenjena raziskavam in razvoju (delovne postaje) bo imel izključno pravico upravljanja v primerih, kjer je potreben fizični dostop do opreme imetnik lastninske pravice opreme, za upravljanje, kjer zadostuje oddaljeni dostop pa vsi trije partnerji konzorcija. V skladu z zmožnostmi lahko to opremo brezplačno in v skladu s sprejetimi pravili dobre prakse in priporočili organov konzorcija uporabljajo, vendar ne upravljajo, tudi ostali člani konzorcija CLARIN.SI. Tudi ta določila so se v celoti izvajala.

V primeru sporov ali nejasnosti pri upravljanju ali uporabi raziskovalne opreme naj bi zavezujoče mnenje podal Upravni odbor konzorcija, vendar do takšnih primerov ni prišlo.

Pogodbene stranke so se v konzorcijski pogodbi tudi dogovorile o delitvi del in izvedbi aktivnosti operacije, rokih za izvedbo posameznih aktivnosti in delitvi sredstev, ki jih bo sofinanciralo ministrstvo, in se dogovorjene delitve, izvedbe, rokov in delitve sredstev tudi držale. Ravno tako so bile prek konzorcijske pogodbe seznanjene s svojimi obveznostmi, in so te obveznosti upoštevale.

Skladnost s predmetom, cilji in nameni NPO

Predmet neposrednih potrditev operacij je bilo sofinanciranje raziskovalne infrastruktur, ki je potrebna za izvajanje večletnih, iz Načrta razvoja raziskovalne infrastrukture 2011-2020 ter Načrta razvoja raziskovalne infrastrukture 2011-2020 (NRRI) Revizija 2016 izhajajočih prednostnih mednarodnih projektov, ki so skladno z metodologijo ESFRI (European Strategy Forum on Research Infrastructures) ob oddaji vloge na evropski ravni dosegli status »implementirani projekt« oz. angleško »landmark«. Ti projekti so usmerjeni v raziskave in razvoj ter sledijo doseganju ciljev prednostne osi 1 »Mednarodna konkurenčnost raziskav, inovacij in tehnološkega razvoja v skladu s pametno specializacijo za večjo konkurenčnost in ozelenitev gospodarstva« iz Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020, kjer so vlaganja v raziskovalno infrastrukturo ukrep, ki je namenjen krepitvi infrastrukture za raziskave, inovacije ter zmogljivosti za razvoj odličnosti na tem področju in spodbujanju pristojnih centrov, zlasti takšnih, ki so evropskega pomena.

CLARIN je bil eden od mednarodnih projektov, ki so dosegli status »implementirani projekt«, zato je operacija skladna s predmetom NPO.

Namen sofinanciranja operacij je bila krepitev nacionalnih nosilnih znanstvenih partnerskih institucij v posameznem infrastrukturnem projektu z doseženim statusom »implementirani projekt«. Slednji so vezani tako na prednostna področja NRRI kot prednostna področja Slovenske Strategije Pametne Specializacije in s tem prispevajo k uresničevanju ciljev Resolucije o raziskovalni in inovacijski strategiji Slovenije 2011-2020, ki predvideva vključitev v mednarodne projekte na področju raziskovalne infrastrukture. Podpora tem infrastrukturnim projektom naj bi tako omogočila bistveno hitrejše, intenzivnejše in visoko kakovostno

vključevanje v velike mednarodne projekte in s tem h krepitvi mreže raziskovalne infrastrukture, človeških virov v znanosti, omogočanju prostega pretoka ljudi, zamisli in znanja v evropskem raziskovalnem prostoru.

Operacija je te namene že dosegla v obdobju svojega trajanja, saj je bistveno izboljšala opremljenost CLARIN.SI z raziskovalno opremo, na osnovi česar je CLARIN.SI omogočila boljše delovanje infrastrukture, kot tudi izvajanje zahtevnih in obsežnih raziskav globokega učenja jezikovnih modelov, s čimer je omogočila oz. olajšala vključevanje partnerjev v mednarodne projekte.

CLARIN.SI je v obdobju trajanja operacije pridobil mednarodni certifikat zaupanja vrednega repozitorija »Core Trust Seal«, in bil recertificiran kot center tipa B CLARIN ERIC. Uspešni smo tudi bili v povezovanju CLARIN.SI in sestrskih slovenskih infrastruktur DARIAH-SI oz. CESSDA-SI (ADP), saj smo z DARIAH-SI (oz. INZ, ki je nosilec te infrastrukture v Sloveniji) sodelovali v projektu CLARIN ERIC ParlaMint »Towards Comparable Parliamentary Corpora«, ki bi ga bilo težko izvesti brez nabavljene raziskovane opreme. S CESDDA-SI (oz. FDV, ki je nosilec te infrastrukture v Sloveniji) je CLARIN.SI uspešno sodeloval v njihovem projektu »RDA Node Slovenia«, ki je vzpostavil slovensko vozlišče mednarodne pobude »Research Data Alliance«.

Kot že omenjeno, ima CLARIN.SI tudi ključno vlogo pri projektu RSDO, kot zahtevani repozitorij za vse jezikovne vire in orodja projekta, pri čemer bodo korpusi projekta RSDO tudi dostopni na konkordančnikih CLARIN.SI.

Repozitorij CLARIN.SI se sedaj uporablja tudi v vedno več nacionalnih raziskovalnih projektih in programih ARRS, in sicer »Tehnologije znanja«, »Jezikovni viri in tehnologije za slovenščino«, »Slovenski jezik - bazične, kontrastivne in aplikativne raziskave«, »Jezikovna krajina sovražnega govora na družbenih omrežjih, »Nova slovnica sodobne slovenščine«, in »Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki«. Repozitorij je postal privlačen tudi za tuje raziskovalce, saj je npr. pridobil več vnosov raziskovalcev in Nizozemske, Švice, Srbije itd. Tudi gruča GPU na UM je bila uporabljena pri več projektih ARRS, npr. “Napredne metode interakcij v telekomunikacijah” in “Fuzija verbalnih in neverbalnih signalov za naslednjo generacijo inteligentnih komunikacijskih vmesnikov – HUMANIPA”.

CLARIN.SI nudi podporo več evropskim projektom, v katerih sodelujejo slovenski raziskovalci, npr. ELEXIS »European Lexicographic Infrastructure« in EMBEDDIA »Cross-Lingual Embeddings for Less-Represented Languages in European News Media«, ki ju vodimo na IJS, že omenjena projekta EU INTAVIA »In/Tangible European Heritage Visual Analysis, Curation & Communication« in MACOCU »Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages«, kot tudi IMSyPP »Innovative Monitoring Systems and Prevention Policies of Online Hate Speech« (Inovativno spremljanje sovražnega govora na spletu in smernice za njegovo preprečevanje), “Patients-centered SurvivorShip care plan after Cancer treatments based on Big Data and Artificial Intelligence technologies” ter projekti CEF (Connecting Europe Facility): Multilingual Resources for CEF.AT in the legal domain (MARCELL), eTranslation TermBank (eTTB), European Language Grid (ELG), Federated TermBank (FedTerm).

Realizacija doseganja kazalnikov

Del operacije je bilo tudi spremljanje kazalnikov učinka, ki so vezani na krovni programski dokument »Operativni program za izvajanje evropske kohezijske politike za obdobje 2014 –

2020«, in sicer Kazalnik CO 25. Kazalnik CO 25 meri število obstoječih delovnih mest v objektih z raziskovalnimi zmogljivostmi, ki so neposredno vezana na aktivnosti raziskav in razvoja in jih neposredno zadeva izvedena operacija. Le ta morajo seveda biti zasedena in prav tako se ne upoštevajo podporna delovna mesta.

V predmetnem projektu so bili načrtovani sledeči kazalniki:

ID CO 25						
Partner	**2019**	**2020**	**2021**	**2022**	**2023**	**Skupaj**
IJS	0,50	1,00	1,00	1,00	1,00	4,50
CVJT UL	0,50	1,00	1,00	1,00	1,00	4,50
FERI UM	1,00	1,25	1,25	1,25	1,25	6,00
VZHODNA SLOVENIJA	**1,00**	**1,25**	**1,25**	**1,25**	**1,25**	**6,00**
ZAHODNA SLOVENIJA	**1,00**	**2,00**	**2,00**	**2,00**	**2,00**	**9,00**
SKUPAJ PROJEKT	**2,00**	**3,25**	**3,25**	**3,25**	**3,25**	**15,00**

Kazalniki naj bi se nanašali na uporabo raziskovalne opreme s strani raziskovalcev konzorcijskih partnerjev, bodisi uporabo spletnih storitev CLARIN.SI, ki za svoje delovanje uporablja raziskovalno opremo, bodisi z neposredno uporabo raziskovalne opreme, tj. novih delovnih postaj. Od teh zaposlenih bodo aktivnosti, ki so neposredno vezane na aktivnosti raziskav in razvoja in jih neposredno zadeva izvedena operacija, aktivno izvajalo več oseb z različno povprečno letno obremenitvijo. Konkretno je ocena kazalnikov po partnerjih temeljila na sledečih predpostavkah:

- **IJS –** na oddelkih, kjer bo locirana nova oprema, deluje prek 25 FTE raziskovalcev. Od teh bo na novi opremi delovala ena oseba, ki bo obremenjena s povprečno 35 % letne delovne obveznosti, 2 osebi s povprečno 25 % in še ena oseba s povprečno 15 % letne delovne obveznosti. IJS bo na tak način dosegel kazalnik učinka ID CO 25 v višini 1,00 FTE letno v osrednjih letih izvajanja operacije, oziroma v letih od 2020 do 2023 v višini 4,50 FTE.
- **CJVT UL** – Center za jezikovne vire in tehnologije Univerze v Ljubljani združuje več fakultet, kjer skupno deluje prek 100 FTE raziskovalcev. Od teh bodo na novi opremi delovali predvsem zaposleni na Fakulteti za računalništvo in informatiko, na Filozofski fakulteti in na Fakulteti za družbene vede. Tu bosta z delom na raziskovalni opremi dve osebi obremenjeni s povprečno 20 % letne delovne obveznosti, 4 osebe pa s povprečno 15% letne delovne obveznosti. CJVT UL bo na tak način dosegel kazalnik učinka ID CO 25 v višini 1,00 FTE letno v osrednjih letih izvajanja operacije, oziroma v letih od 2020 do 2023 v višini 4,50 FTE.
- **FERI UM** – Fakulteto za elektrotehniko, računalništvo in informatiko UM sestavlja 14 raziskovalnih skupin, kjer deluje prek 30 FTE raziskovalcev. Od teh bodo na novi opremi delovali predvsem zaposleni na Inštitutu za elektroniko in telekomunikacije in Inštitutu za računalništvo (Laboratorij za heterogene računalniške sisteme). Tu bosta dve osebi obremenjeni s povprečno 50 % letne delovne obveznosti, ena oseba pa s povprečno 25% letne delovne obveznosti. FERI UM bo na tak način dosegel kazalnik

učinka ID CO 25 v višini 1,25 FTE letno v osrednjih letih izvajanja operacije, oziroma v letih od 2020 do 2023 v višini 6,00 FTE.

Višino kazalnikov smo v sklopu operacije spremljali v odvisnosti od opreme in s tem partnerja:

- **IJS** – na IJS vsak zaposleni dela na enem ali več stroškovnih mestih (raziskovalnih projektih oz. programskih skupinah), in za vsako stroškovno mesto se mesečno obračuna število ur, ki jih je raziskovalec na njem opravil. Za izračun CO25 smo sešteli ure raziskovalcev po projektih, ki so uporabljali v sklopu operacije nabavljeno opremo.
- **UL CJVT** – na UL CJVT vsak zaposleni dela na enem ali več stroškovnih mestih (raziskovalnih projektih oz. programskih skupinah) z različnimi delovnimi obveznostmi po posameznih projektih. Število ur, ki jih je raziskovalec na posameznem projektu opravil, se obračuna mesečno, upoštevajoč njegov delež zaposlitve na tem projektu. Za izračun CO25 smo sešteli ure raziskovalcev po projektih, ki so uporabljali opremo, ki je bila nabavljena v sklopu operacije.
- **FERI UM** – na UM FERI vsak zaposleni dela na enem ali več stroškovnih mestih (raziskovalnih projektih oz. programskih skupinah), in za vsako stroškovno mesto se mesečno obračuna število ur, ki jih je raziskovalec na njem opravil. Za izračun CO25 smo sešteli ure raziskovalcev po projektih, ki so uporabljali v sklopu operacije nabavljeno opremo.

V skladu z izmerjenimi kazalniki, smo v teku operacije dosegli naslednje vrednosti CO25:

ID CO 25				
Partner	**2019***	**2020**	**2021 (do 31.8.2021)**	**Skupaj**
IJS	0,00	1,01	1,03	2,04
CVJT UL	0,00	1,00	0,70	1,70
FERI UM	0,00	2,06	2,86	4,92
SKUPAJ PROJEKT	**0,00**	**4,07**	**4,59**	**8,66**

***Pojasnilo: Pogodba je bila podpisana šele konec julija 2019, nato pa je bilo potrebno pripraviti razpise za raziskovalno opremo predvideno za nakup v 2019 na IJS, UL in UM, izbrati najboljše ponudnike, opremo prevzeti, jo fizično namestiti, preizkusiti njeno delovaje in nanjo namestiti potrebno programsko opremo. To je bilo sicer storjeno do konca 2019, vendar pa raziskovalci še niso mogli začeti uporabljati nove raziskovalne opreme. Oprema je dejansko prišla v uporabo v 2020.

S kazalniki CO25 je tako operacija dosegla predvideno vrednost v času trajanja operacije.

Način dostopanja in opis zasedenosti opreme

Oprema je locirana glede na partnerje, obenem pa ima, glede na vrsto opreme po partnerju tudi precej drugačno namembnost, zato opis zasedenosti in dostopanja navajamo po partnerjih.

Institut »Jožef Stefan«

Oprema IJS je služila oz. služi hitrejšemu in proti napakam bolj odpornemu delovanju osnovnih spletnih storitev CLARIN.SI, predvsem repozitorija, spletnih konkordančnikov in drugih spletnih storitev CLARIN.SI. Način dostopa je prek spleta, pri čemer je večina storitev povsem odprtih, in za njih torej ni potrebna nikakršna prijava, omejen nabor storitev (predvsem dostop do delno zaprtih virov repozitorija) pa zahteva prijavo prek AAI, torej prijavo prek akademskega gesla EduGain. Ker je zasedenost opreme nepredvidljiva, kapacitete pa naravnane, da lahko pokrivajo tudi maksimalne obremenitve (kot npr. cel letnik študentov pri vajah uporablja konkordančnike) to pomeni, da večino časa oprema ni polno zasedena. Zato eno od dveh gruč računalnikov uporabljamo z manjšo prioriteto tudi za dolgotrajne naloge, kot je npr. zajem in obdelava spletnih podatkov. Zaradi teh dejstev je tudi procesorsko zasedenost nabora vseh računalnikov v gručah težko meriti, oz. te meritve ne bi imele posebnega smisla, vendar lahko okvirno rečemo, da je trenutna procesorska obremenitev opreme nabavljene v okviru operacija približno 30 %, kar pomeni, da bo predvidoma, glede na trend povečevanja uporabe in ponudbe v zadnjih leti trenutna strojna oprema zadoščala še za nadaljnjih 3 do 5 let. Zasedenost lahko bolj enostavno merimo tudi v diskovnih kapacitetah, torej, koliko prostora na trenutno instaliranih diskovnih poljih je še na voljo. Tu je zasedenost trenutno 25 %, kar, spet glede na trend objave novih jezikovnih virov, in posebej z ozirom na predvidene nove vire, ki bodo deponirani v okviru projekta RSDO tudi pomeni, da diskovne kapacitete zadoščajo še za 3 do 5 let.

Univerza v Ljubljani

Oprema Centra za jezikovne vire in tehnologije (CJVT) UL je namenjena delovanju spletnih storitev CJVT, predvsem spletnih konkordančnikov in vmesnikov za slovarske in leksikonske podatkovne zbirke: Slovar sopomenk sodobne slovenščine, Slovenski oblikoslovni leksikon, Gigafida 2.0 - Korpus pisne standardne slovenščine, Kolokacijski slovar sodobne slovenščine, Veliki slovensko-madžarski slovar. Poleg tega tudi spletna orodja: Orodje za strojno postavljanje vejic, Orodje za strojno ocenjevanje berljivosti besedil. Način dostopa je prek spleta, pri čemer so storitve povsem odprte, za dostop prijava ni potrebna. Narava storitev je takšna, da je oprema v času večjih obremenitev (v času velikega števila uporabnikov in poizvedb) polno izkoriščena iz vidika procesorskih kapacitet, v času manjših obremenitev pa manj, vendar dnevno povprečje okrog 25 %. Zaradi prostorsko obsežnih zbirk, ki jih storitve ponujajo, je poraba pomnilnika visoka (80 %), da se zagotavlja odzivnost; prav tako so podatki na lokalnem disku NVMe, kjer je zasedenost okrog 50 %.

Univerza v Mariboru

Na Fakulteti za elektrotehniko, računalništvo in informatiko UM (FERI) je dostop do GPU gruče omogočen uporabnikom iz vseh članic projektne skupine, v kolikor ti uporabniki opremo uporabljajo v namene, skladne s cilji in projektom RI-SI CLARIN. Uporabniki so tako dolžni slediti navodilom za uporabo, objavljenim na http://wiki.ietk.um.si/mediawiki, izvajati naloge na GPU gruči skladno s tem, kar so napovedali, ter ne smejo uporabljati sistema za hranjenje podatkov, temveč prenesejo rezultate na lastne nosilce podatkov. UM FERI omogoča dostop do GPU gruče prek sistema SLURM. Pri tem sistemu uporabniki napovejo število GPU-jev, na katerih naj se naloga izvaja (omejitev na največ 4 od 8), številu CPU-jev (omejitev največ 35 od 80), skupno količina RAM-a (največ 250 GB), diskovni prostor

(zaenkrat ne omejujemo), čas trajanja naloge (najdaljši predviden čas je do 7 dni, več po dogovoru s tehničnim administratorjem). Če viri niso na voljo, se uporabniki razvrščajo glede na to, koliko virov zahtevajo. Tisti, ki zahtevajo veliko virov, so uvrščeni na nižje mesto. UM FERI vodi dnevnik uporabe GPU gruče. V nadaljevanju je prikazan povzetek dnevnika uporabe:

Leto	Mesec	Minute procesiranja	Ure procesiranja
2020	junij	42893	714.8833333
	julij	10201	170.0166667
	avgust	43205	720.0833333
	september	129423	2157.05
	oktober	131225	2187.083333
	november	57633	960.55
	december	71265	1187.75
2021	januar	1662299	27704.98333
	februar	1413585	23559.75
	marec	1179875	19664.58333
	april	946695	15778.25
	maj	709840	11830.66667
	junij	429746	7162.433333
	julij	265442	4424.033333
	avgust	136461	2274.35

UM FERI je skozi operacijo pridobil tudi strežniško infrastrukturo, se uporablja za izvajanje raziskav na področju ugotavljanja pomena in ekstrakcije terminologije ter znanja iz besedil v slovenskem in drugih jezikih. Z vzpostavljenim diskovnim poljem lahko hranimo precej več jezikovnih vhodnih virov, iz katerih pridobivamo kvalitetne podatke in na njih izvajamo analize ter ekstrakcijo znanja. Z dodatnimi rezervnimi diski zagotavljamo neprekinjeno delovanje strežniške in podatkovne infrastrukture v primeru okvar podatkovnih nosilcev na diskovnih poljih. Trenutna sistemska obremenitev (CPU, pomnilnik) se giblje okrog 40 %, medtem ko je zasedenost prostorskih kapacitet diskovnega polja okrog 30 %. Glede na trend uporabe in polnjenja diskovnih kapacitet, ocenjujemo, da bo trenutna strojna oprema zadoščala za nadaljnja 4 leta.

Načrti za prihodnost

Konzorcijski partnerji se bodo prizadevali za dolgoročno vzdržnost in delovanje infrastrukture po zaključku projekta.

Infrastrukturo bodo vzdrževali, kar se tiče programske opreme repozitorija, konkordančnikov in ostalih spletnih storitev, kot tudi uvajali nove spletne storitve. Še naprej bo imela prednost kvalitetna in ažurna dokumentacija spletnih storitev, tako v slovenščini kot v angleščini.

Infrastrukturo, storitve in repozitorij bomo še naprej vključevali v nacionalne in mednarodne infrastrukture in projekte, v tekoče znanstveno-raziskovalno in akademsko delo ter v učno-izobraževalne procese. Kot omenjeno, smo uspešno sodelovali v projektu CLARIN ERIC ParlaMint »Towards Comparable Parliamentary Corpora«, kjer smo že dobili odobreno nadaljevanje projekta (začetek 1.10.2021), in v katerem bomo prevzeli več računalniško zelo

zahtevnih nalog (kot je npr. strojno prevajanje vse korpusov parlamentarnih razprav projekta v predvideni velikosti ene milijarde besed v angleški jezik) in katerih izvajanje bi bilo nemogoče brez pridobljene raziskovalne opreme.

Partnerji bodo tudi delovali na skupnem razvoju v okviru, interdisciplinarnega dela ter integracij z nacionalnimi in internacionalnimi infrastrukturami za znanost, kakršne so CLARIN.SI in CLARIN ERIC za področje dela, ter SLING, Arnes, nacionalna infrastruktura odprtega dostopa, EGI, PRACE, OpenAire, EOSC in EUDAT za širše področje znanstvene infrastrukture. Člani konzorcija naj bi vzpostavljeno infrastrukturo in dejavnosti v okviru možnosti vključili tudi v svoje razvoje projekte ter infrastrukturne centre, kar bo omogočalo ustrezno sofinanciranje povezanih dejavnosti preko virov raziskovalnih in infrastrukturnih skupin.

Pripravil: izr. prof. dr. Tomaž Erjavec vodja projekta	Žig	Odgovorna oseba upravičenca: prof. dr. Boštjan Zalar direktor IJS
Podpis: _____		Podpis: _____

Dodatek 1. Seznam vnosov v repozitorij CLARIN.SI 1.1.2020–1.9.2021

Repozitorij CLARIN.SI je glavna storitev infrastrukture. Navajamo vnose v repozitorij CLARIN.SI od datuma instalacije prvega paketa raziskovalne opreme na IJS od zaključka operacije:

1. Erjavec, Tomaž; Ljubešić, Nikola; Fišer, Darja, 2020-05-05, *English-Slovene term candidates KAS-biterm 1.0*, https://hdl.handle.net/11356/1263.
   133,710 entries
2. Dobrovoljc, Kaja; Roblek, Rebeka; Vianello, Chiara; Diaci, Ajda; Vuga, Zala, 2020-01-06, *List of formulaic sequences in spoken Slovenian*, https://hdl.handle.net/11356/1279.
   2,374 expressions
3. Dobrovoljc, Kaja; Roblek, Rebeka; Vianello, Chiara; Diaci, Ajda; Vuga, Zala, 2020-01-06, *List of formulaic sequences in standard written Slovenian*, https://hdl.handle.net/11356/1280.
   1,891 expressions
4. Vuković, Teodora, 2020-09-01, *Spoken Torlak dialect corpus 1.0 (transcription)*, https://hdl.handle.net/11356/1281.
   498,021 tokens, 92,232 utterances, 96 texts, 86 hours
5. Ljubešić, Nikola, 2020-01-07, *The CLASSLA-StanfordNLP model for lemmatisation of standard Slovenian 1.1*, https://hdl.handle.net/11356/1286.
6. Čibej, Jaka; Arhar Holdt, Špela; Dobrovoljc, Kaja; Krek, Simon, 2020-02-13, *Consonant-vowel structures in the Gigafida 2.0 corpus*, https://hdl.handle.net/11356/1289.
7. Ljubešić, Nikola, 2020-01-07, *The CLASSLA-StanfordNLP model for lemmatisation of standard Croatian 1.1*, https://hdl.handle.net/11356/1287.
8. Verdonik, Darinka, 2020-01-23, *Dialogue act annotated spoken corpus GORDAN 1.0 (transcription)*, https://hdl.handle.net/11356/1291.
   1 hours
9. Ljubešić, Nikola, 2020-01-07, *The CLASSLA-StanfordNLP model for lemmatisation of standard Serbian 1.1*, https://hdl.handle.net/11356/1288.
10. Čibej, Jaka; Arhar Holdt, Špela; Dobrovoljc, Kaja; Krek, Simon, 2020-02-13, *Consonant-vowel structures in the GOS 1.0 corpus*, https://hdl.handle.net/11356/1290.
11. Zwitter Vitez, Ana; Zemljarič Miklavčič, Jana; Krek, Simon; Stabej, Marko; Erjavec, Tomaž; Verdonik, Darinka; Krajnc Ivič, Mira; Antloga, Špela; Majhenič, Simona, 2020-01-23, *Dialogue act annotated spoken corpus GORDAN 1.0 (audio/video)*, https://hdl.handle.net/11356/1292.
    1 hours
12. Pančur, Andrej; Erjavec, Tomaž; Ojsteršek, Mihael; Šorn, Mojca; Blaj Hribar, Neja, 2020-04-13, *Slovenian parliamentary corpus (1990-2018) siParl 2.0*, https://hdl.handle.net/11356/1300.
    11,967 texts, 1,134,540 utterances, 11,537,491 sentences, 239,749,733 tokens, 106 gb
13. Antloga, Špela, 2020-02-04, *Metaphor corpus KOMET 1.0*, https://hdl.handle.net/11356/1293.
    218,730 words, 259,881 tokens
14. Pollak, Senja; Vulić, Ivan; Pelicon, Andraž; Repar, Andraž; Armendariz, Carlos; Matthew, Purver; Ljubešić, Nikola, 2020-05-15, *SimLex-999 Slovenian translation SimLex-999-sl 1.0*, https://hdl.handle.net/11356/1309.
    999 entries
15. Ljubešić, Nikola, 2020-04-29, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Slovenian 1.1*, https://hdl.handle.net/11356/1312.
16. Mlakar, Izidor; Majhenič, Simona; Rojc, Matej; Verdonik, Darinka, 2020-04-22, *Multimodal corpus EVA 1.0*, https://hdl.handle.net/11356/1311.
    57 minutes
17. Ulčar, Matej; Robnik-Šikonja, Marko, 2020-06-16, *CroSloEngual BERT*, https://hdl.handle.net/11356/1317.
18. Žejn, Andrejka; Erjavec, Tomaž, 2021-03-18, *The corpus of older Slovenian narrative prose PriLit 1.0*, https://hdl.handle.net/11356/1319.
    43 texts, 1,275,209 tokens

19. Daelemans, Walter; Fišer, Darja; Franza, Jasmin; Kranjčić, Denis; Lemmens, Jens; Ljubešić, Nikola; Markov, Ilia; Popič, Damjan, 2020-06-04, *The LiLaH Emotion Lexicon of Croatian, Dutch and Slovene*, https://hdl.handle.net/11356/1318.
    14,182 entries
20. Ljubešić, Nikola, 2020-06-19, *The CLASSLA-StanfordNLP model for named entity recognition of standard Croatian 1.0*, https://hdl.handle.net/11356/1322.
21. Ljubešić, Nikola, 2020-06-19, *The CLASSLA-StanfordNLP model for named entity recognition of standard Slovenian 1.0*, https://hdl.handle.net/11356/1321.
22. Ljubešić, Nikola, 2020-06-19, *The CLASSLA-StanfordNLP model for named entity recognition of standard Serbian 1.0*, https://hdl.handle.net/11356/1323.
23. Ljubešić, Nikola, 2020-06-24, *The CLASSLA-StanfordNLP model for JOS dependency parsing of standard Slovenian 1.0*, https://hdl.handle.net/11356/1325.
24. Ljubešić, Nikola; Osenova, Petya; Simov, Kiril, 2020-06-24, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Bulgarian 1.0*, https://hdl.handle.net/11356/1326.
25. Ljubešić, Nikola; Osenova, Petya; Simov, Kiril, 2020-06-24, *The CLASSLA-StanfordNLP model for lemmatisation of standard Bulgarian 1.0*, https://hdl.handle.net/11356/1327.
26. Ljubešić, Nikola; Osenova, Petya; Simov, Kiril, 2020-06-24, *The CLASSLA-StanfordNLP model for UD dependency parsing of standard Bulgarian 1.0*, https://hdl.handle.net/11356/1328.
27. Ljubešić, Nikola; Osenova, Petya; Simov, Kiril, 2020-07-07, *The CLASSLA-StanfordNLP model for named entity recognition of standard Bulgarian 1.0*, https://hdl.handle.net/11356/1329.
28. Ljubešić, Nikola; Štefanec, Vanja, 2020-07-17, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of non-standard Serbian 1.0*, https://hdl.handle.net/11356/1332.
29. Ljubešić, Nikola; Štefanec, Vanja, 2020-07-17, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of non-standard Croatian 1.0*, https://hdl.handle.net/11356/1331.
30. Ulčar, Matej; Robnik-Šikonja, Marko, 2020-07-09, *CroSloEngual BERT 1.1*, https://hdl.handle.net/11356/1330.
31. Ljubešić, Nikola; Štefanec, Vanja, 2020-07-17, *The CLASSLA-StanfordNLP model for lemmatisation of non-standard Serbian 1.0*, https://hdl.handle.net/11356/1334.
32. Ljubešić, Nikola; Štefanec, Vanja, 2020-07-17, *The CLASSLA-StanfordNLP model for lemmatisation of non-standard Croatian 1.0*, https://hdl.handle.net/11356/1333.
33. Škvorc, Tadej; Gantar, Polona; Robnik-Šikonja, Marko, 2020-07-27, *Dataset of Slovene idiomatic expressions SloIE*, https://hdl.handle.net/11356/1335.
    29,400 sentences, 695,636 tokens
34. Ljubešić, Nikola, 2020-08-06, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of non-standard Slovenian 1.0*, https://hdl.handle.net/11356/1337.
35. Ljubešić, Nikola, 2020-08-06, *The CLASSLA-StanfordNLP model for lemmatisation of non-standard Slovenian 1.0*, https://hdl.handle.net/11356/1338.
36. Ljubešić, Nikola, 2020-08-07, *The CLASSLA-StanfordNLP model for named entity recognition of non-standard Slovenian 1.0*, https://hdl.handle.net/11356/1339.
37. Pelicon, Andraž; Pranjić, Marko; Miljković, Dragana; Škrlj, Blaž; Pollak, Senja, 2020-09-15, *Sentiment Annotated Dataset of Croatian News*, https://hdl.handle.net/11356/1342.
    2,025 entries
38. Ljubešić, Nikola, 2020-08-07, *The CLASSLA-StanfordNLP model for named entity recognition of non-standard Croatian 1.0*, https://hdl.handle.net/11356/1340.
39. Ljubešić, Nikola, 2020-08-07, *The CLASSLA-StanfordNLP model for named entity recognition of non-standard Serbian 1.0*, https://hdl.handle.net/11356/1341.
40. Erjavec, Tomaž; Grigorova, Vladislava; Ljubešić, Nikola; Ogrodniczuk, Maciej; Osenova, Petya; Pančur, Andrej; Rudolf, Michał; Simov, Kiril, 2020-10-15, *Multilingual comparable corpora of parliamentary debates ParlaMint 1.0*, https://hdl.handle.net/11356/1345.
    678,094 utterances, 88,307,799 words
41. Pollak, Senja; Arhar Holdt, Špela; Krek, Simon; Robnik-Šikonja, Marko, 2020-09-10, *Reference List of Slovene Frequent Common Words*, https://hdl.handle.net/11356/1346.
    4,768 entries

42. Ljubešić, Nikola, 2020-09-11, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Croatian 1.1*, https://hdl.handle.net/11356/1348.
43. Supej, Anka; Ulčar, Matej; Robnik-Šikonja, Marko; Pollak, Senja, 2020-09-24, *List of single-word male and female occupations in Slovenian*, https://hdl.handle.net/11356/1347.
    234 entries
44. Ljubešić, Nikola, 2020-09-11, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Serbian 1.1*, https://hdl.handle.net/11356/1349.
45. Ljubešić, Nikola, 2020-09-15, *The CLASSLA-StanfordNLP model for lemmatisation of non-standard Slovenian 1.1*, https://hdl.handle.net/11356/1350.
46. Ljubešić, Nikola; Štefanec, Vanja, 2020-07-17, *The CLASSLA-StanfordNLP model for lemmatisation of non-standard Croatian 1.1*, https://hdl.handle.net/11356/1352.
47. Ljubešić, Nikola; Štefanec, Vanja, 2020-09-15, *The CLASSLA-StanfordNLP model for lemmatisation of non-standard Serbian 1.1*, https://hdl.handle.net/11356/1351.
48. Ljubešić, Nikola; Osenova, Petya; Simov, Kiril, 2020-06-24, *The CLASSLA-StanfordNLP model for lemmatisation of standard Bulgarian 1.1*, https://hdl.handle.net/11356/1353.
49. Ljubešić, Nikola, 2020-09-15, *The CLASSLA-StanfordNLP model for lemmatisation of standard Slovenian 1.2*, https://hdl.handle.net/11356/1354.
50. Ljubešić, Nikola, 2020-09-15, *The CLASSLA-StanfordNLP model for lemmatisation of standard Serbian 1.2*, https://hdl.handle.net/11356/1355.
51. Ljubešić, Nikola, 2020-09-15, *The CLASSLA-StanfordNLP model for lemmatisation of standard Croatian 1.2*, https://hdl.handle.net/11356/1357.
52. Ljubešić, Nikola, 2020-10-13, *Word embeddings CLARIN.SI-embed.mk 0.1*, https://hdl.handle.net/11356/1359.
53. Bogetić, Ksenija; Batanović, Vuk, 2020-10-30, *Annotated corpus of Slovenian language-related news articles MetaLangNEWS-Sl*, https://hdl.handle.net/11356/1360.
    555 articles, 588,854 tokens
54. Bogetić, Ksenija; Batanović, Vuk, 2020-10-30, *Annotated corpus of Slovenian language-related news comments MetaLangNEWS-COMMENTS-Sl*, https://hdl.handle.net/11356/1362.
    555 articles, 1,384 texts, 44,176 tokens
55. Čibej, Jaka; Arhar Holdt, Špela; Dobrovoljc, Kaja; Krek, Simon, 2020-10-28, *Frequency lists of character-level n-grams from the GOS 1.0 corpus 1.1*, https://hdl.handle.net/11356/1363.
    15 files
56. Čibej, Jaka; Arhar Holdt, Špela; Dobrovoljc, Kaja; Krek, Simon, 2020-10-28, *Frequency lists of words from the GOS 1.0 corpus 1.1*, https://hdl.handle.net/11356/1364.
57. Čibej, Jaka; Arhar Holdt, Špela; Dobrovoljc, Kaja; Krek, Simon, 2020-10-28, *Frequency lists of word-level n-grams from the GOS 1.0 corpus 1.1*, https://hdl.handle.net/11356/1365.
    23 files
58. Čibej, Jaka; Arhar Holdt, Špela; Dobrovoljc, Kaja; Krek, Simon, 2020-10-28, *Frequency lists of word parts from the GOS 1.0 corpus 1.1*, https://hdl.handle.net/11356/1366.
    390 files
59. Čibej, Jaka; Arhar Holdt, Špela; Dobrovoljc, Kaja; Krek, Simon, 2020-10-28, *Consonant-vowel structures in the GOS 1.0 corpus 1.1*, https://hdl.handle.net/11356/1367.
60. Šimko, Ivan, 2020-11-03, *Annotated Corpus of Pre-Standardized Balkan Slavic Literature*, https://hdl.handle.net/11356/1368.
    15 texts, 32,729 tokens
61. Bogetić, Ksenija; Batanović, Vuk, 2020-10-30, *Annotated corpus of Croatian language-related news articles MetaLangNEWS-Hr*, https://hdl.handle.net/11356/1369.
    738 articles, 555,890 tokens
62. Bogetić, Ksenija; Batanović, Vuk, 2020-10-30, *Annotated corpus of Croatian language-related news comments MetaLangNEWS-COMMENTS-Hr*, https://hdl.handle.net/11356/1370.
    738 articles, 21,533 texts, 823,459 tokens
63. Bogetić, Ksenija; Batanović, Vuk, 2020-10-30, *Annotated corpus of Serbian language-related news articles MetaLangNEWS-Sr*, https://hdl.handle.net/11356/1371.
    1,088 articles, 659,084 tokens

64. Bogetić, Ksenija; Batanović, Vuk, 2020-10-30, *Annotated corpus of Serbian language-related news comments MetaLangNEWS-COMMENTS-Sr*, https://hdl.handle.net/11356/1372.
    1,088 articles, 14,791 texts, 878,482 tokens
65. Ljubešić, Nikola; Zdravkova, Katerina; Stojanoska, Sanja; Erjavec, Tomaž, 2020-11-05, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Macedonian 1.0*, https://hdl.handle.net/11356/1373.
66. Ljubešić, Nikola; Zdravkova, Katerina; Erjavec, Tomaž, 2020-11-05, *The CLASSLA-StanfordNLP model for lemmatisation of standard Macedonian 1.0*, https://hdl.handle.net/11356/1374.
67. Pobežin, Gregor, 2020-12-06, *Epigraphic corpus of Medieval and Early Modern inscriptions in Slovenia MEMIS 1.0*, https://hdl.handle.net/11356/1376.
    51 texts
68. Žagar, Aleš; Robnik-Šikonja, Marko; Goli, Teja; Arhar Holdt, Špela, 2020-11-13, *Slovene translation of SuperGLUE*, https://hdl.handle.net/11356/1380.
69. Vasić, Daniel; Žitko, Branko; Gašpar, Angelina; Ljubešić, Nikola; Štrkalj Despot, Kristina; Merkler, Danijela, 2020-11-20, *Semantic hypergraph corpus SemCRO 1.0*, https://hdl.handle.net/11356/1377.
    184 sentences, 176 semanticUnits
70. Pirman, Alenka; Doma, Mitja, 2020-12-29, *Dictionary of living Slovenian "Razvezani Jezik" (The Unleashed Tongue)*, https://hdl.handle.net/11356/1385.
    5,800 entries
71. Čibej, Jaka; Arhar Holdt, Špela; Krek, Simon, 2020-12-22, *List of word relations from the Sloleks 2.0 lexicon 1.0*, https://hdl.handle.net/11356/1386.
    66,347 entries
72. Pirman, Alenka, 2020-11-15, *The "Arcticae horulae" dictionary of German borrowings in Slovenian*, https://hdl.handle.net/11356/1379.
    2,080 entries
73. Erjavec, Tomaž; Ogrodniczuk, Maciej; Osenova, Petya; Ljubešić, Nikola; Simov, Kiril; Grigorova, Vladislava; Rudolf, Michał; Pančur, Andrej; Kopp, Matyáš; Barkarson, Starkaður; Steingrímsson, Steinþór; van der Pol, Henk; Depoorter, Griet; de Does, Jesse; Jongejan, Bart; Haltrup Hansen, Dorte; Navarretta, Costanza; Calzada Pérez, María; de Macedo, Luciana D.; van Heusden, Ruben; Marx, Maarten; Çöltekin, Çağrı; Coole, Matthew; Agnoloni, Tommaso; Frontini, Francesca; Montemagni, Simonetta; Quochi, Valeria; Venturi, Giulia; Ruisi, Manuela; Marchetti, Carlo; Battistoni, Roberto; Sebők, Miklós; Ring, Orsolya; Darģis, Roberts; Utka, Andrius; Petkevičius, Mindaugas; Briedienė, Monika; Krilavičius, Tomas; Morkevičius, Vaidas, 2021-05-10, *Multilingual comparable corpora of parliamentary debates ParlaMint 2.0*, https://hdl.handle.net/11356/1388.
    462,217,184 words, 3,307,079 utterances
74. Ulčar, Matej; Robnik-Šikonja, Marko, 2020-12-29, *Slovenian RoBERTa contextual embeddings model: SloBERTa 1.0*, https://hdl.handle.net/11356/1387.
75. Ljubešić, Nikola; Krsnik, Luka, 2021-02-02, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Slovenian 1.2*, https://hdl.handle.net/11356/1391.
76. Ljubešić, Nikola; Krsnik, Luka, 2021-02-02, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Serbian 1.2*, https://hdl.handle.net/11356/1392.
77. Ljubešić, Nikola; Krsnik, Luka, 2021-02-02, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Croatian 1.2*, https://hdl.handle.net/11356/1393.
78. Ljubešić, Nikola; Osenova, Petya; Simov, Kiril; Krsnik, Luka, 2021-02-02, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Bulgarian 1.1*, https://hdl.handle.net/11356/1394.
79. Shekhar, Ravi; Pranjic, Marko; Pollak, Senja; Pelicon, Andraž; Purver, Matthew, 2021-04-19, *24sata news comment dataset 1.0*, https://hdl.handle.net/11356/1399.
    21,548,192 texts
80. Ljubešić, Nikola; Zdravkova, Katerina; Stojanoska, Sanja; Erjavec, Tomaž; Krsnik, Luka, 2021-02-02, *The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Macedonian 1.1*, https://hdl.handle.net/11356/1395.
81. Ulčar, Matej; Robnik-Šikonja, Marko, 2021-01-17, *Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0*, https://hdl.handle.net/11356/1397.

82. Kralj Novak, Petra; Mozetič, Igor; Ljubešić, Nikola, 2021-02-17, *Slovenian Twitter hate speech dataset IMSyPP-sl*, https://hdl.handle.net/11356/1398.
120,000 items
83. Shekhar, Ravi; Pollak, Senja; Pelicon, Andraž; Matthew, Purver; Krustok, Ivar, 2021-04-19, *Ekspress user comment dataset 1.0*, https://hdl.handle.net/11356/1401.
31,473,732 texts
84. Jemec Tomazin, Mateja; Trojar, Mitja; Žagar, Mojca; Atelšek, Simon; Fajfar, Tatjana; Erjavec, Tomaž, 2021-03-15, *Corpus of term-annotated texts RSDO5 1.0*, https://hdl.handle.net/11356/1400.
12 texts, 38,043 terms, 257,029 words, 310,588 tokens
85. Koloski, Boshko; Pollak, Senja; Škrlj, Blaž; Martinc, Matej, 2021-06-04, *Keyword extraction datasets for Croatian, Estonian, Latvian and Russian 1.0*, https://hdl.handle.net/11356/1403.
2,000,000 bytes
86. Ljubešić, Nikola, 2021-02-24, *Choice of plausible alternatives dataset in Croatian COPA-HR*, https://hdl.handle.net/11356/1404.
1,000 items
87. Shekhar, Ravi; Purver, Matthew; Pollak, Senja; Pelicon, Andraž; Krustok, Ivar, 2021-04-19, *Latvian user comment dataset 1.0*, https://hdl.handle.net/11356/1407.
12,412,463 texts
88. Erjavec, Tomaž; Ogrodniczuk, Maciej; Osenova, Petya; Ljubešić, Nikola; Simov, Kiril; Grigorova, Vladislava; Rudolf, Michał; Pančur, Andrej; Kopp, Matyáš; Barkarson, Starkaður; Steingrímsson, Steinþór; van der Pol, Henk; Depoorter, Griet; de Does, Jesse; Jongejan, Bart; Haltrup Hansen, Dorte; Navarretta, Costanza; Calzada Pérez, María; de Macedo, Luciana D.; van Heusden, Ruben; Marx, Maarten; Çöltekin, Çağrı; Coole, Matthew; Agnoloni, Tommaso; Frontini, Francesca; Montemagni, Simonetta; Quochi, Valeria; Venturi, Giulia; Ruisi, Manuela; Marchetti, Carlo; Battistoni, Roberto; Sebők, Miklós; Ring, Orsolya; Darģis, Roberts; Utka, Andrius; Petkevičius, Mindaugas; Briedienė, Monika; Krilavičius, Tomas; Morkevičius, Vaidas; Bartolini, Roberto; Cimino, Andrea, 2021-05-12, *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.0*, https://hdl.handle.net/11356/1405.
462,217,184 words, 3,307,079 utterances
89. Purver, Matthew; Pollak, Senja; Freienthal, Linda; Kuulmets, Hele-Andra; Krustok, Ivar; Shekhar, Ravi, 2021-04-19, *Ekspress news article archive (in Estonian and Russian) 1.0*, https://hdl.handle.net/11356/1408.
1,441,112 texts
90. Arhar Holdt, Špela; Čibej, Jaka; Laskowski, Cyprian; Krek, Simon, 2020-12-12, *Morphological patterns from the Sloleks 2.0 lexicon 1.0*, https://hdl.handle.net/11356/1411.
96,290 units
91. Purver, Matthew; Shekhar, Ravi; Pranjić, Marko; Pollak, Senja; Martinc, Matej, 2021-04-19, *24sata news article archive 1.0*, https://hdl.handle.net/11356/1410.
657,806 texts
92. Pollak, Senja; Purver, Matthew; Shekhar, Ravi; Freienthal, Linda; Kuulmets, Hele-Andra; Krustok, Ivar, 2021-04-19, *Latvian Delfi article archive (in Latvian and Russian) 1.0*, https://hdl.handle.net/11356/1409.
180,401 texts
93. Ahačič, Kozma; Atelšek, Simon; Erjavec, Tomaž; Holozan, Peter; Jakop, Nataša; Jemec Tomazin, Mateja; Ježovnik, Janoš; Ledinek, Nina; Perdih, Andrej; Romih, Miro; Trojar, Mitja, 2021-04-01, *Corpus of Slovenian school texts*, https://hdl.handle.net/11356/1413.
96,257 tokens, 7,161 sentences
94. Ljubešić, Nikola; Krsnik, Luka, 2021-03-02, *The CLASSLA-StanfordNLP model for lemmatisation of standard Slovenian 1.3*, https://hdl.handle.net/11356/1412.
95. Krek, Simon; Gantar, Polona; Kosem, Iztok; Dobrovoljc, Kaja; Arhar Holdt, Špela; Čibej, Jaka; Laskowski, Cyprian; Klemenc, Bojan; Krsnik, Luka, 2021-03-09, *Frequency lists of collocations from the Gigafida 2.1 corpus*, https://hdl.handle.net/11356/1415.
82 files, 4,002,918 collocations

96. Bogunović, Irena; Kučić, Mario; Ljubešić, Nikola; Erjavec, Tomaž, 2021-03-14, *Corpus of Croatian news portals ENGRI (2014-2018)*, https://hdl.handle.net/11356/1416.
694,799,268 tokens, 1,756,735 texts
97. Erjavec, Tomaž; Fišer, Darja; Ljubešić, Nikola; Ferme, Marko; Borovič, Mladen; Boškovič, Borko; Ojsteršek, Milan; Hrovat, Goran, 2021-03-31, *Abstracts from the KAS corpus KAS-Abs 1.0*, https://hdl.handle.net/11356/1420.
108,254 texts, 30,728,838 words
98. Krek, Simon; Gantar, Polona; Krsnik, Luka; Laskowski, Cyprian; Dobrovoljc, Kaja; Arhar Holdt, Špela; Čibej, Jaka; Kosem, Iztok; Klemenc, Bojan; Robnik-Šikonja, Marko; Gorjanc, Vojko, 2021-03-16, *Valency lexicon extracted from the Gigafida 2.1 corpus*, https://hdl.handle.net/11356/1418.
14,595 entries
99. Krek, Simon; Gantar, Apolonija; Laskowski, Cyprian; Krsnik, Luka; Kosem, Iztok; Brank, Janez; Dobrovoljc, Kaja; Arhar Holdt, Špela; Čibej, Jaka; Robnik-Šikonja, Marko; Klemenc, Bojan; Gorjanc, Vojko, 2021-03-25, *Multiword Expressions lexicon extracted from the Gigafida 2.1 corpus*, https://hdl.handle.net/11356/1421.
5,242 entries
100. Evkoski, Bojan; Pelicon, Andraž; Mozetič, Igor; Ljubešić, Nikola; Kralj Novak, Petra, 2021-07-20, *Slovenian Twitter dataset 2018-2020 1.0*, https://hdl.handle.net/11356/1423.
12,961,136 texts
101. Lemmenmeier-Batinić, Dolores; Ljubešić, Nikola; Samardžić, Tanja, 2021-04-06, *Corpus of Serbian Forms of Address 1.0*, https://hdl.handle.net/11356/1422.
19 files, 171,552 tokens, 16,040 turns
102. Kosem, Iztok; Čibej, Jaka; Ljubešić, Nikola; Krek, Simon; Gantar, Polona; Arhar Holdt, Špela; Logar, Nataša; Laskowski, Cyprian; Klemenc, Bojan; Dobrovoljc, Kaja; Gorjanc, Vojko; Pori, Eva, 2020-10-26, *The Orange workflow for observing collocation clusters ColEmbed 1.0*, https://hdl.handle.net/11356/1425.
103. Kosem, Iztok; Krek, Simon; Čibej, Jaka; Gantar, Polona; Arhar Holdt, Špela; Logar, Nataša; Laskowski, Cyprian; Klemenc, Bojan; Ljubešić, Nikola; Dobrovoljc, Kaja; Gorjanc, Vojko; Pori, Eva, 2020-10-26, *The Orange workflow for observing collocation trends ColTrend 1.0*, https://hdl.handle.net/11356/1424.
104. Ljubešić, Nikola, 2021-05-05, *Text collection for training the BERTić transformer model BERTić-data*, https://hdl.handle.net/11356/1426.
8,387,681,518 words
105. Ljubešić, Nikola; Markoski, Filip; Markoska, Elena; Erjavec, Tomaž, 2021-05-05, *Comparable corpora of South-Slavic Wikipedias CLASSLA-Wikipedia 1.0*, https://hdl.handle.net/11356/1427.
1,928,450 articles, 37,677,016 sentences, 486,258,862 tokens
106. Kosem, Iztok; Pori, Eva; Gantar, Polona; Logar, Nataša; Krek, Simon; Laskowski, Cyprian; Arhar Holdt, Špela; Čibej, Jaka; Dobrovoljc, Kaja; Gorjanc, Vojko; Klemenc, Bojan; Ljubešić, Nikola, 2020-10-26, *Slovene ontology of semantic types for nouns SLONEST-noun 1.0*, https://hdl.handle.net/11356/1428.
271 entries
107. Ljubešić, Nikola; Erjavec, Tomaž, 2021-05-13, *Montenegrin web corpus meWaC 1.0*, https://hdl.handle.net/11356/1429.
321,573 texts, 3,654,071 sentences, 90,871,077 tokens
108. Juvan, Marko; Žejn, Andrejka; Šorli, Mojca; Mandić, Lucija; Tomažin, Andrej; Jež, Andraž; Balžalorsky Antić, Varja; Erjavec, Tomaž, 2021-05-31, *Corpus of 1968 Slovenian literature Maj68 1.0*, https://hdl.handle.net/11356/1430.
874 texts, 646,970 words, 794,382 tokens
109. Erjavec, Tomaž; Ogrodniczuk, Maciej; Osenova, Petya; Ljubešić, Nikola; Simov, Kiril; Grigorova, Vladislava; Rudolf, Michał; Pančur, Andrej; Kopp, Matyáš; Barkarson, Starkaður; Steingrímsson, Steinþór; van der Pol, Henk; Depoorter, Griet; de Does, Jesse; Jongejan, Bart; Haltrup Hansen, Dorte; Navarretta, Costanza; Calzada Pérez, María; de Macedo, Luciana D.; van Heusden, Ruben; Marx, Maarten; Çöltekin, Çağrı; Coole, Matthew; Agnoloni, Tommaso; Frontini, Francesca; Montemagni, Simonetta; Quochi, Valeria; Venturi, Giulia; Ruisi, Manuela; Marchetti, Carlo; Battistoni, Roberto; Sebők, Miklós; Ring, Orsolya; Darģis, Roberts; Utka, Andrius; Petkevičius, Mindaugas; Briedienė, Monika; Krilavičius, Tomas; Morkevičius, Vaidas; Bartolini, Roberto; Cimino, Andrea; Diwersy, Sascha;

Luxardo, Giancarlo; Rayson, Paul, 2021-06-18, *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1*, https://hdl.handle.net/11356/1431.
3,774,204 utterances, 494,949,904 words
110. Ljubešić, Nikola; Fišer, Darja; Erjavec, Tomaž, 2021-05-28, *Offensive language dataset of Croatian, English and Slovenian comments FRENK 1.0*, https://hdl.handle.net/11356/1433.
32,795 texts
111. Erjavec, Tomaž; Ogrodniczuk, Maciej; Osenova, Petya; Ljubešić, Nikola; Simov, Kiril; Grigorova, Vladislava; Rudolf, Michał; Pančur, Andrej; Kopp, Matyáš; Barkarson, Starkaður; Steingrímsson, Steinþór; van der Pol, Henk; Depoorter, Griet; de Does, Jesse; Jongejan, Bart; Haltrup Hansen, Dorte; Navarretta, Costanza; Calzada Pérez, María; de Macedo, Luciana D.; van Heusden, Ruben; Marx, Maarten; Çöltekin, Çağrı; Coole, Matthew; Agnoloni, Tommaso; Frontini, Francesca; Montemagni, Simonetta; Quochi, Valeria; Venturi, Giulia; Ruisi, Manuela; Marchetti, Carlo; Battistoni, Roberto; Sebők, Miklós; Ring, Orsolya; Darģis, Roberts; Utka, Andrius; Petkevičius, Mindaugas; Briedienė, Monika; Krilavičius, Tomas; Morkevičius, Vaidas; Diwersy, Sascha; Luxardo, Giancarlo; Rayson, Paul, 2021-06-18, *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1*, https://hdl.handle.net/11356/1432.
3,774,204 utterances, 494,949,904 words
112. Krek, Simon; Dobrovoljc, Kaja; Erjavec, Tomaž; Može, Sara; Ledinek, Nina; Holz, Nanika; Zupan, Katja; Gantar, Polona; Kuzman, Taja; Čibej, Jaka; Arhar Holdt, Špela; Kavčič, Teja; Škrjanec, Iza; Marko, Dafne; Jezeršek, Lucija; Zajc, Anja, 2021-07-07, *Training corpus ssj500k 2.3*, https://hdl.handle.net/11356/1434.
586,248 tokens, 27,829 sentences, 500,295 words
113. Kuvač Kraljević, Jelena; Hržica, Gordana; Štefanec, Vanja; Kolografić Belić, Lana; Ljubešić, Nikola, 2021-06-15, *Croatian corpus of non-professional written language by typical speakers and speakers with language disorders RAPUT 1.0*, https://hdl.handle.net/11356/1435.
6,760 texts, 34,469 sentences, 426,187 tokens
114. Verdonik, Darinka; Potočnik, Tomaž; Sepesy Maučec, Mirjam; Erjavec, Tomaž; Majhenič, Simona; Žgank, Andrej, 2021-06-21, *Spoken corpus Gos VideoLectures 4.1 (transcription)*, https://hdl.handle.net/11356/1439.
55 texts, 1,872 utterances, 11,260 sentences, 170,830 words
115. Šimko, Ivan, 2021-07-02, *Annotated Corpus of Pre-Standardized Balkan Slavic Literature 1.1*, https://hdl.handle.net/11356/1441.
23 texts, 53,257 tokens
116. Brank, Janez, 2021-07-14, *Q-CAT Corpus Annotation Tool 1.2*, https://hdl.handle.net/11356/1442.