REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI SKLAD ZA REGIONALNI RAZVOJ
NALOŽBA V VAŠO PRIHODNOST

Referenčni korpus pisne standardne slovenščine Gigafida

Poročilo projekta Razvoj slovenščine v digitalnem okolju

Aktivnost DS1.4

Avtorji: Špela Arhar Holdt[○,□], David Bordon[□], Rok Chitrakar[□], Helena Dobrovoljc[△], Nataša Gliha Komac[△], Vojko Gorjanc[□], Janoš Ježovnik[△], Simon Krek[⬡], Nataša Logar[◇], Duša Race[△], Andreja Žele[△]

[○] Fakulteta za računalništvo in informatiko Univerze v Ljubljani

[□] Filozofska fakulteta Univerze v Ljubljani

[△] ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša

[⬡] Institut "Jožef Stefan", Ljubljana

[◇] Fakulteta za družbene vede Univerze v Ljubljani

Ljubljana: Center za jezikovne vire in tehnologije, Univerza v Ljubljani, 2023

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Vsebina

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

1 Vojko Gorjanc in Špela Arhar Holdt: Projektni okvir

Poročilo oz. kazalnik *Referenčni korpus pisne standardne slovenščine Gigafida* je nastalo pod okriljem projekta Razvoj slovenščine v digitalnem okolju, ki sta ga med leti 2020 in 2023 sofinancirali Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj (Operacija se je izvajala v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020). Cilj projekta je bil zadovoljiti potrebe po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik za raziskovalne organizacije, podjetja in širšo javnost. Spletna stran projekta z dostopom do rezultatov: https://slovenscina.eu/.

Kazalnik se umešča v prvi projektni delovni sklop z naslovom *Jezikovni viri*. Namen delovnega sklopa je bil nadgraditi slovenske besedilne korpuse in leksikon besednih oblik. Prenovili smo učne množice in postopke za strojno označevanje sodobne slovenščine. Rezultat so osveženi in povečani jezikovni viri, ki so na voljo tako uporabniški skupnosti kot za strojno rabo. Z razvitimi postopki in orodji bo posodabljanje slovenskih korpusov v prihodnosti hitrejše in preprostejše.

Cilj aktivnosti 1.4 je bil pripraviti načrt za dolgoročno in kontinuirano nadgradnjo korpusov Gigafida in Janes ter vzpostaviti delotoke za redno zbiranje besedil. Aktivnost se je posvečala dvema temeljnima besedilnima korpusoma pisne slovenščine, in sicer:

- referenčnemu korpusu pisne slovenščine Gigafida in
- korpusu Janes.

Korpusa predstavljata osnovo za jezikovni opis, predpis, priročnike, jezikovne tehnologije in postopke vseh vrst. Namen aktivnosti je bil osredinjen na to, kako poskrbeti za dolgoročno osveževanje obeh korpusov, pri čemer je treba v načrtih upoštevati izkušnje deležnikov, ki korpusa uporabljajo za razvoj izdelkov. Temeljni premisleki so bili potrebni na ravni gradivne zastopanosti, opredeljevanja standardnosti besedil in s tem povezane vključitve slovenskega zamejstva. Ob tem je bilo potrebno načrtovati in pripraviti infrastrukturo za kontinuirano nadgrajevanje korpusov, to pomeni urejanje pravnih vprašanj glede pridobivanja besedil, vzpostavitev spletne platforme z informacijami za besedilodajalce in analizo mreže besedilodajalcev.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Projektno poročilo v nadaljevanju govori o analizah in pripravljenih rešitvah za referenčni korpus pisne slovenščine Gigafida.

2 Vojko Gorjanc: Opis projektnih aktivnosti

Kazalnik predstavi opravljene analize in na njihovi osnovi pripravljene rešitve za referenčni korpus pisne slovenščine Gigafida.

- Analiza potreb in pričakovanj uporabnikov in različnih deležnikov, ki pri svojem delu uporabljajo korpusa Gigafida in Janes, je izhodišče za načrtovanje nadgradnje korpusov, zato je bila načrtovana in opravljena uporabniška študija pri različnih deležnikih in v različnih kontekstih.
- Za aktualno varianto korpusa je bil prvič uporabljen avtomatski postopek pripisovanja standardnosti; ta je bil evalviran, evalvacija pa umeščena v kontekst koncepta standardnega jezika z določitvijo okvira standardnosti besedil.
- Slovenski referenčni korpusi do sedaj niso sistematično vključevali zamejske besedilne produkcije, v zadnji različici pa so bila tovrstna besedila načrtno izključena, saj je bil potreben celovit premislek o tem, na kakšen način naj bodo vključena ter kako zagotoviti reprezentativen vzorec zamejskih besedil. Na primeru slovenskega zamejstva v Republiki Italiji je bil pripravljen predloga za vključevanje besedil zamejskih besedil v (pod)korpus(e) standardne slovenščine.
- Rešitve za dolgoročno zbiranje gradiva vključujejo pripravo pravnih rešitev za zbiranje besedil neodvisno od posameznega projektnega dela, izdelavo spletnega portala za oddajo besedil z informacijami za besedilodajalce in celostno analizo mreže potencialnih besedilodajalcev z vsemi kontaktnimi informacijami ter izbor besedilodajalcev za prioritetno urejanje pravnih razmerij.
- Izdelava specifikacij nadgradnje korpusa Gigafida v kontekstu razvoja referenčnih korpusov slovenskega jezika od korpusa Fida in FidaPLUS do različnih verzij korpusa Gigafida. Predstavljen je načrt uravnoteženosti korpusa za doseganje njegove reprezentativnosti z vidika do sedaj podreprezentiranih besedil.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

REZULTATI AKTIVNOSTI

- Spletno mesto z informacijami za besedilodajalce, pravnimi rešitvami za dolgoročno zbiranje gradiva: https://zbiranje.cjvt.si/gigafida/.
- Tabela s celotnim seznamom potencialnih besedilodajalcev z vsemi kontaktnimi informacijami: https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/03/RSDO_Kazalnik_Gigafida_Priloga2.pdf in tabela z izborom besedilodajalcev za prioritetno urejanje pravnih razmerij: https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/03/RSDO_Kazalnik_Gigafida_Priloga3.pdf.
- Analiza in predstavitev uporabniških potreb in pričakovanj deležnikov, ki uporabljajo referenčni korpus Gigafida: https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/03/RSDO_Kazalnik_Gigafida_Priloga1.pdf.
- Na evalvacijah temelječe pričujoče specifikacije za nadgradnjo korpusa Gigafida z opredelitvijo vprašanja ne/standardnosti in vključitvijo zamejskega dela.

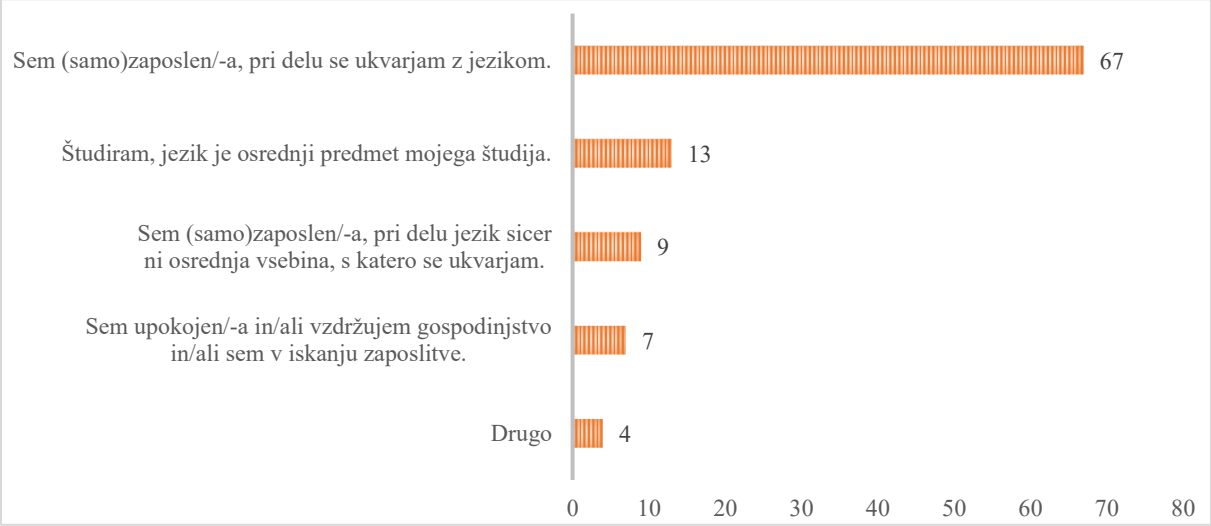
3 Nataša Logar, Vojko Gorjanc in Špela Arhar Holdt: Uporabniške potrebe in pričakovanja

3.1 Uvod

Za analizo uporabniških potreb in pričakovanj je bila pripravljena in izvedena spletna anketa v orodju 1KA, sestavljena iz 18 vprašanj, od katerih je bilo 13 namenjenih korpusu Gigafidi 2.0. Anketa je bila odprta od 16. junija 2021 do 15. oktobra 2021. Izpolnila sta jo 302 anketirana, od tega je bilo 234 anket izpolnjenih v celoti, in sicer v spodnjih deležih glede na dejavnost anketiranih (Graf 1).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

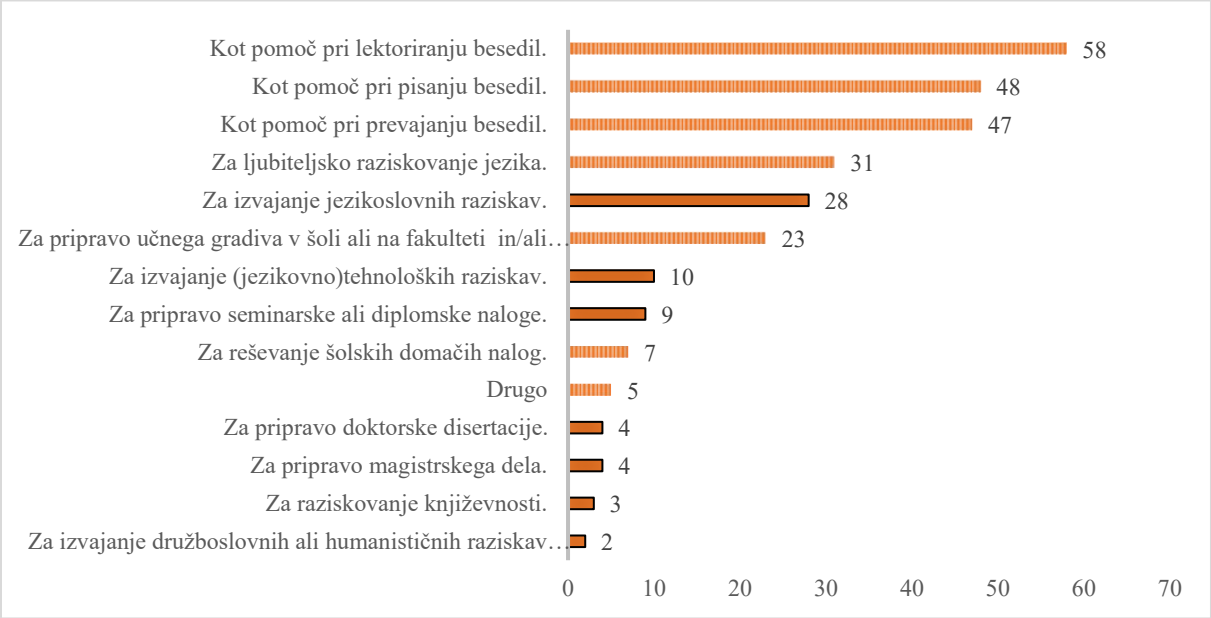
Graf 1: Anketirani po dejavnosti, deleži v %



3.2 Analiza ankete

Celotna anketa skupaj z rezultati je na voljo kot Priloga 1, njena natančnejša analiza je pripravljena v obliki znanstvenega članka, oddanega v recenzijski postopek (Logar et al. 2023). Tu izpostavljamo samo rezultate osrednjih vprašanj v anketi.

Graf 2: Namen uporabe korpusa Gigafida 2.0 v zadnjem letu anketirancev



Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Anketiranci uporabljajo korpus največ za lektoriranje, sledita pisanje besedil in prevajanje. Manj, a še vedno nad 30 % pa je t. i. ljubiteljske rabe, sledita še približno četrtinska obsega uporabe za raziskovanje jezika in za pripravo učnih gradiv. Preostali nameni so bili izbrani pri manj kot 10-odstotnem deležu anketiranih.

Tabela 1: Povprečna ocena posameznih lastnosti korpusa Gigafida 2.0 in število anketiranih, ki so podali oceno; ocenjevalna lestvica: 1 = 'niti slučajno'; 5 = 'izredno, popolnoma'.

Trditev, s katero se *niti slučajno ne strinjam, /.../, se izredno, popolnoma strinjam*	**Povprečna ocena**	**N**
Korpus je uporaben za moje delo.	**4,28**	249
Prikaz rezultatov iskanj je pregleden.	**4,15**	247
Prikaz rezultatov iskanj je hiter.	**4,01**	248
Na voljo imam vse možnosti iskanj, ki jih potrebujem za svoje delo.	**3,58**	235
Delo s korpusom je uporabniku prijazno.	**4,06**	245
Informacije o korpusu, ki so na voljo v vmesniku (»O viru«), so razumljive.	**4,36**	185
Korpus (natančneje: vmesnik, v katerem je korpus na voljo) mi je vizualno všeč.	**3,95**	242

Najbolje so bile sicer ocenjene informacije, dostopne v zavihku »O viru« (kratek opis vsebine korpusa in seznam glavnih publikacij o njem), najslabšo povprečno oceno pa so anketirani dali iskalnim možnostim, o čemer so na koncu ankete podali tudi največ komentarjev. Povprečna ocena konkordančnika je bila sorazmerno visoka – 4.

Odgovore na vprašanja, in sicer tako tista, ki smo jih prikazali zgoraj, kot tudi druga v uvodnem delu ankete, lahko strnemo v naslednje:

- korpus Gigafida 2.0 se največ uporablja za jezikoslovno in drugo raziskovanje jezika, lektoriranje, pisanje ter prevajanje;
- uporabniki so posamezne lastnosti korpusa ocenili z visoko skupno oceno 4;
- uporabnikom je najbolj všeč preglednost vmesnika, sprememba, ki si jo najbolj želijo, pa je redno posodabljanje besedil;
- v primerjavi s šestimi drugimi jezikovnimi viri za slovenščino se korpus Gigafida 2.0 uporablja najpogosteje;

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- dodajanje na novo razvitih virov ni prineslo niti povečanja niti zmanjšanja uporabe korpusa Gigafida 2.0.

Vprašanja odprtega tipa prinašajo dragocene informacije za nadaljnji razvoj korpusa in korpusnega vmesnika, izpostavljamo pa dve ključni nalogi za nadaljnji razvoj:

- Pri razvoju korpusa je treba več pozornosti posvetiti iskalnim možnostim, pri čemer se trenutno kažeta dve težavi:
  - začetna zahtevnost za nevešče uporabnike: rešitev bi bila že poljudna pomoč, kakršna je bila pred leti dostopna pri Gigafidi;
  - potrebe specializiranih uporabnikov: za posamezne profile specializiranih uporabnikov bi bilo treba pridobiti dodatne informacije, izkušnje v tujini pa kažejo, da tovrstni uporabniki ne posegajo po korpusnih virih s klasičnimi konkordančniki, ampak zahtevnejšimi orodji ali z analizo korpusa kot podatkovne baze.
- Glede na komentarje uporabnikov se zdi najbolj ključno v resnici redno posodabljanje korpusa z novimi besedili.

4 Rok Chitrakar, Vojko Gorjanc in David Bordon: Evalvacija avtomatskega pripisovanja standardnosti besedil

4.1 Uvod

Namen korpusa Gigafida je predstaviti slovenski standardni jezik, kar pomeni, da se vanj vključujejo besedila, ki jih prepoznavamo kot standardna. Za aktualno varianto korpusa je bil prvič uporabljen avtomatski postopek pripisovanja standardnosti. Z evalvacijo izločenih besedil iz baze korpusa smo želeli ugotoviti, ali je postopek ustrezen, prav tako pa tudi, katera besedila z avtomatskim označevanjem ostajajo zunaj baze korpusa Gigafida.

4.2 Okvir standardnega jezika

Korpus standardnega jezika Gigafida temelji na konceptu standardnosti, kot je bil opredeljen v Gorjanc et al. (2015: 47–48). Zanj je značilno, da se zavestno omejuje na tiste dele jezikovne stvarnosti, ki jezikovni skupnosti služijo kot skupno mesto sporazumevanja v okviru formalnih sistemov, kot so

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

izobraževalni sistem, sistem javne oz. državne uprave, vključno z umeščenostjo v širšo skupnost EU, z izhodiščem v funkcionalni dogovornosti, ki izhaja iz ustrezno interpretirane jezikovne rabe. Ideološko se standardni jezik umešča v nadaljevanje standardnojezikovne kulture, ki je značilna za slovensko jezikovno skupnost, pri čemer za del jezika, ki ga razume in definira kot standardnega, vir dogovora išče v spremljanju jezikovne rabe s sprejemanjem konsenzualnih odločitev na podlagi maksime o prilagajanju standarda na delih, ki največ govorcem povzročajo največ težav. Izvedbeno je pojem standardnega jezika usmerjen v zagotavljanje možnosti, da vsak govorec slovenščine vsak trenutek lahko sam interpretira vse okoliščine rabe kateregakoli izraza v odnosu do njegove standardnosti ali nestandardnosti, kar mu omogočajo ne le klasični jezikovni opisi, temveč tudi drugi jezikovni viri in orodja, med katerimi je osrednji prav referenčni korpus pisne slovenščine Gigafida. Osnovno izhodišče za vključevanje besedil s stališča standardnosti je bilo predvidevanje, da so njihovi avtorji imeli intenco, ustvariti besedilo v standardni slovenščini (Krek at al. 2019: 12).

4.3 Avtomatsko pripisovanje standardnosti: korpus Gigafida 2.0

Za korpus Gigafida 2.0 je bila določitev nestandardnih besedil opravljena avtomatsko, in sicer z orodjem za avtomatsko določanje stopnje nestandardnosti besedil (Ljubešić et al. 2015), ki je bilo razvito v sklopu projekta JANES. Orodje besedilu pripiše raven nestandardnosti na dveh nivojih. Tehnična nestandardnost upošteva predvsem oblikovno ustreznost besedila (npr. ustrezna raba presledkov, ločil in velikih začetnic), jezikovna nestandardnost pa je določena glede na besedišče in zapis (Krek et al. 2019: 13).

4.4 Evalvacija avtomatskega pripisovanja standardnosti

Da bi ugotovili, kako uspešno je orodje oz. ali je dovolj robustno pri določanju nestandardnosti, smo kvalitativno evalvirali vzorec izločenih besedil. Trije evalvatorji so neodvisno pregledali 500 besedil, ki so bila izločena za korpus Gigafida 2.0. Besedila so bila opredeljena kot standardna ali nestandardna v dveh korakih. V prvem je vsak od evalvatorjev neodvisno označil vsa besedila, po diskusiji v krogu skupine RSDO D1.4 glede različnih interpretacij nestandardnosti pa je bila pripisana dokončna odločitev evalvatorjev. Ti so kot nedvoumno nestandardna besedila označili sorazmerno majhen odstotek izločenih besedil, kot kaže Diagram 1.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

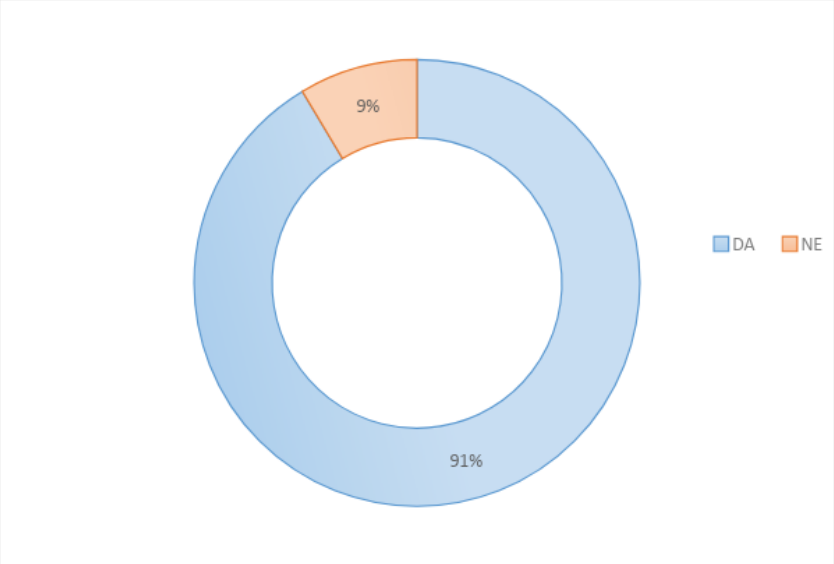
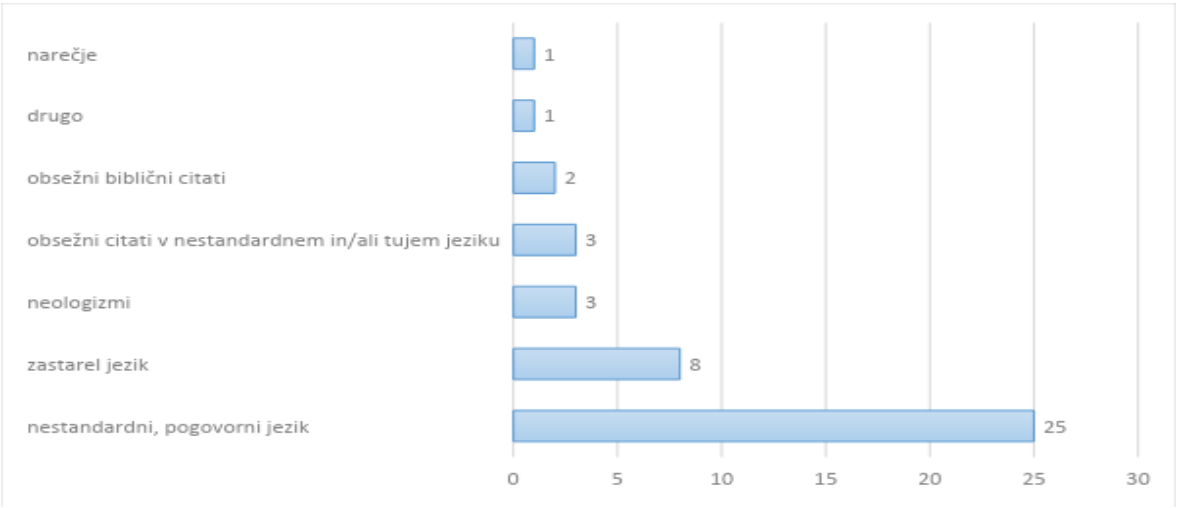


Diagram 1: Kvalitativna ocean (ne)standardnosti besedil

Vsem nestandardnim besedilom so evalvatorji pripisali tudi razlog, zakaj po njihovem mnenju besedilo ni standardno. Ko smo opisne oznake združili, smo dobili razporeditev razlogov nestandardnosti besedil, predstavljeno v Grafu 3.

Graf 3: Razlog za oceno nestandardnosti



Največ besedil je bilo izločenih, ker so zapisana v pogovornem jeziku, vsi drugi razlogi so veliko redkejši. Prav besedila v celoti v pogovornem ali zastarelem jeziku so označevali za nestandardna najbolj nedvoumno.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Pri označevanju je bilo največ dilem povezanih z naslednjimi vprašanji:

- Kako obravnavam primere, kjer so dialogi večinoma v pogovornem ali narečnem jeziku?
- Ali so standardna besedila z velikim številom neologizmov oz. kreativnega besedotvorja, npr. znanstvena fantastika?
- Kako je z besedili, v katerih je veliko besedila v tujem jeziku (npr. celi dialogi)?

Nedvoumno so se označevalci odločali glede otroške literature, pisane fonetično v pogovornem jeziku, v pregledanih primerih predvsem v ljubljanščini, na drugi strani pa so nedvoumno označili za standardna besedila literaturo za otroke, ki je pri avtomatskem pripisovanju standardnosti izločena iz standardnih besedil. V avtomatsko izločenih besedilih je tudi sicer veliko leposlovnih besedil, pisanih v nestandardnem jeziku ali z veliko kreativnega besedišča in neologizmov. Vsekakor gre za dragocena besedila, za katera je treba premisliti, kje in kako jih predstaviti. Glede vključevanja otroške literature, pisane v standardnem jeziku, pa je tudi glede na izhodiščno opredelitev standardnega jezika potreben celovit konceptualni premislek.

5 Janoš Ježovnik, Helena Dobrovoljc, Nataša Gliha Komac, Duša Race in Andreja Žele: Vključevanje zamejskih besedil

5.1 Uvod

Slovenska jezikovna skupnost je zunaj Republike Slovenije prisotna na območjih Avtonomne zvezne dežele Furlanije - Julijske krajine v Republiki Italiji (in sicer na območjih nekdanjih Goriške, Tržaške in Videmske pokrajine), na južnih delih zveznih dežel Koroška in Štajerska v Republiki Avstriji, na območju Porabja v Železni županiji v Republiki Madžarski ter v obmejnih delih Republike Hrvaške.

Čeprav zlasti na avstrijskem Koroškem ter na Tržaškem in Goriškem v Italiji tamkajšnje slovenske skupnosti proizvajajo znaten delež besedil v knjižni slovenščini in ta dosegajo in vplivajo na razmeroma velik krog naslovnikov, slovenski referenčni korpusi do sedaj zamejske besedilne produkcije niso sistematično vključevali. V zadnji različici je bil delež tovrstnih besedil načrtno izključen (Krek et al. 2020: 3341) zaradi potrebe po celovitem premisleku o tem, na kakšen način naj bodo vključena ter kako zagotoviti reprezentativen vzorec zamejskih besedil.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

5.2 Slovenska manjšina v Republiki Italiji

Slovenska jezikovna skupnost je ena od ustavno priznanih jezikovnih manjšin v Republiki Italiji. Zaščito ji na državni ravni zagotavljata *Zakon o zaščiti zgodovinskih jezikovnih manjšin* (zakon št. 482/1999) in *Zakon o zaščiti slovenske jezikovne manjšine v Italiji* (zakon št. 38/2001). Območje prisotnosti slovenske skupnosti v obmejnem pasu Avtonomne zvezne dežele Furlanije - Julijske krajine je smiselno zaradi naravnogeografskih, zgodovinskih, družbeno-kulturnih, ekonomskih in upravnopolitičnih razlik obravnavati v okviru šestih bolj ali manj zaokroženih enot: Kanalska dolina, Rezija, Terske doline, Nadiške ali Nediške doline (vse del nekdanje Videmske pokrajine), Goriško in Tržaško (Jagodic 2017: 42; za natančnejši oris družbenih in zgodovinskih okoliščin prim. tudi Stranj 1999. Zaradi naštetih različnih okoliščin obstajajo med posameznimi območji velike razlike v prisotnosti in statusu slovenskega knjižnega oziroma standardnega jezika; območje nekdanje Videmske pokrajine se tako bistveno razlikuje od Goriškega in Tržaškega (Jagodic, Kaučič-Baša in Dapit 2017) ter terja posebno obravnavo.

5.3 Korpus zamejskih besedil

5.3.1 Utemeljitev potrebe

Grgič (2017: 91) opozarja na neraziskanost »številni[h] ključni[h] vidik[ov] jezikovne rabe« zamejske slovenščine in na pomanjkanje sodobnih jezikovnih priročnikov. Zaradi stika z romanskimi jezikovnimi različicami, zaradi političnozgodovinskih okoliščin in zaradi različnih poimenovalnih potreb je slovenski jezik v Italiji – tudi v svoji knjižni oziroma standardni različici – razvil določene posebnosti, po katerih se razlikuje od osrednjeslovenske različice (Jagodic, Kaučič-Baša in Dapit 2017: 68); živi in se razvija predvsem v leposlovnih besedilih, v medijih in v nekaterih strokovnih besedilih, med drugim v učbeniških besedilih (ibid.). Prav tako je zaradi različnih okoliščin mogoče, da se od ostalih razlikuje tudi regionalna različica knjižnega oziroma standardnega jezika, ki se rabi na območju Videmske pokrajine. Ustrezno zasnovan korpus zamejske knjižne produkcije je nujna gradivska osnova tako za raziskave regionalnih knjižnih oziroma standardnih jezikovnih različic z vidika leksikologije, terminologije,

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

pragmatike, sociolingvistike ipd. (prim. Franza 2018: 17–20) kot za pripravo sodobnih slovarjev in drugih jezikovnih priročnikov.[1]

5.3.2 Izhodišča za pripravo korpusa

Besedilno produkcijo slovenske jezikovne skupnosti v Republiki Italiji v slovenskem knjižnem oz. standardnem jeziku je smiselno vključiti v obstoječi referenčni korpus slovenskega jezika Gigafida 2.0. Besedila, primerna za vključitev, so tista, ki nastajajo v slovenskem zamejskem okolju in so namenjena zlasti za to okolje (gl. seznam potencialnih besedilodajalcev oz. publikacij v Tabeli 1 nižje).

Zaradi skladnosti z referenčnim korpusom je smiselno slediti načelom izgradnje tega korpusa – specifikacije izdelave aktualne različice korpusa Gigafida 2.0 so strnjene v Krek et al. 2019. V nadaljevanju so povzete smernice za izgradnjo korpusa, ki zagotavljajo skladnost s trenutno aktualno različico Gigafida 2.0; v primeru posodobitve tega korpusa z drugimi metodami je seveda treba uporabiti te.

Za tokenizacijo in stavčno segmentacijo Gigafida 2.0 je bil uporabljen označevalnik Obeliks (Grčar et al. 2012), za oblikoskladenjsko označevanje in lematizacijo pa metaoznačevalnik, ki združuje izhode omenjenega označevalnika in označevalnika ReLDI (Ljubešić in Erjavec 2016). Ker odstopanj, ki bi lahko bistveno vplivala na uspešnost označevanja, ne pričakujemo, je mogoče za označevanje zamejskega dela korpusa uporabiti enake postopke.

Korpus Gigafida 2.0 obsega besedila, objavljena med letoma 1990 in 2018, načrtovane so sprotne posodobitve z novim aktualnim gradivom. Zaradi primerljivosti bi bilo idealno v korpus vključiti zamejska besedila, objavljena od leta 1990 dalje, ob nadaljnjih posodobitvah pa sproti vključevati tudi besedila iz zamejske produkcije. Za potrebe razlikovanja in primerjave s preostalimi deli korpusa bi bilo treba knjižna oziroma standardna besedila zamejske produkcije kot taka označiti.

V korpus Gigafida 2.0 oziroma njegove predhodnike Gigafida, FidaPLUS in FIDA so bila le priložnostno že vključena nekatera besedila zamejske produkcije (npr. vsebine časopisa *Novi glas* 2004–2010 v

[1] Načrt za korpusnojezikoslovni opis slovenščine v Italiji je skupaj s testnimi slovarskimi sestavki na podlagi korpusne analize testnega korpusa pripravila Franza (2018).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

skupnem obsegu približno 7,8 milijona besed; *Novi Matajur* 2006–2009 v skupnem obsegu približno 1,5 milijona besed – izločeno ob posodobitvi na različico 2.0). Ker predstavljajo le majhen delež zamejske besedilne produkcije v knjižni oz. standardni slovenščini, je treba pretežni del (pod)korpusa šele izgraditi. S tem v mislih je smiselno gradivo za korpus zbirati širše in bolj splošno, obenem pa imeti v mislih urejanje pravnih razmerij z besedilodajalci, saj to omogoča javni dostop do vsebine in čim bolj prosto nadaljnjo distribucijo korpusa. Pri razreševanju avtorskopravne problematike je treba upoštevati specifike italijanske zakonodajne ureditve področja ter zato za pripravo dogovorov z besedilodajalci pritegniti ustrezno usposobljene pravne strokovnjake.

5.3.3 Posebni tipi besedil

Slovenski knjižni oz. standardni jezik je na območju Videmske pokrajine rabljen le redko in v zelo omejenih krogih ter zato razmeroma malo vpliven. V pisni obliki se pojavlja v publikacijah slovenskih kulturnih in založniških ustanov, kjer se rabi vzporedno z zlasti italijanskim jezikom ter z lokalnimi narečji. Edini redni periodični publikaciji sta tednik *Novi Matajur* in štirinajstdnevnik *Dom*. Za razliko od publikacij s Tržaškega in Goriškega, kjer lahko utemeljeno pričakujemo prisotnost knjižnih oz. standardnih besedil, je pri publikacijah z območja nekdanje Videmske pokrajine potrebna predhodna selekcija, v primeru obeh časnikov tudi na ravni posameznih prispevkov. V Tabeli 2 so te publikacije označene z rumeno barvo. Obenem je treba širše presoditi, ali so besedila, namenjena otrokom, primerna za vključitev v referenčni korpus knjižne oz. standardne slovenščine. V Tabeli 2 so takšne publikacije iz zamejstva označene z modro barvo.[2]

[2] Revijo *Galeb* izdaja isti založnik kot tednik *Novi Matajur*, vendar je revija pisana v knjižni oz. standardni slovenščini in praviloma ne vsebuje narečnih besedil.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Tabela 2: Potencialni besedilodajalci oz. publikacije za vključitev v korpus: z zeleno barvo so označene publikacije, ki so primerne za vključitev; z modro besedila za otroke, pri čemer bi bilo treba presoditi, ali so primerna za vključitev ali ne; z rumeno pa publikacije iz nekdanje Videmske pokrajine, katerih ustreznost za vključitev bi morali individualno presoditi, v primeru časnikov tudi po prispevkih.

Založnik	Publikacija	Besedilne zvrsti[3]
Družba za založniške pobude	Primorski dnevnik (www.primorski.eu)	časopisi
Zadruga Goriška Mohorjeva	Novi glas (www.noviglas.eu)	časopisi
	Pastirček	revije
	knjižne izdaje	leposlovje, stvarna besedila
Založba Mladika	Mladika (www.mladika.com/revija-mladika/)	revije
	Rast [mladinska priloga revije Mladika]	revije
	knjižne izdaje	leposlovje, stvarna besedila
Zadružna kraška banka	Skupaj	revije
	Isonzo – Soča (www.isonzo-soca.it)	revije
Založništvo tržaškega tiska	knjižne izdaje	leposlovje, stvarna besedila
Slovenski raziskovalni inštitut	knjižne izdaje	stvarna besedila
Deželni šolski urad	didaktična, učbeniška besedila	stvarna besedila
Centralni urad za slovenski jezik		stvarna besedila
občinski prevajalski uradi		stvarna besedila
SloSport	spletne vsebine v slovenščini (www.slosport.org)	internet
RAI	spletne vsebine v slovenščini (http://www.sedezfjk.rai.it)	internet
Zadruga Novi Matajur	Novi Matajur (novimatajur.it)	časopisi
	Galeb (www.galeb.it)	revije
Zadruga Most	Dom, kulturno-verski list (www.dom.it)	časopisi
Inštitut za slovensko kulturo	knjižne izdaje	stvarna besedila
Kulturno društvo Ivan Trinko	knjižne izdaje	stvarna besedila
Študijski center Nediža	knjižne izdaje	stvarna besedila
Zadruga don Evgen Blankin	knjižne izdaje	stvarna besedila, leposlovje

[3] Glede na oznake besedilnih zvrsti, uporabljene v korpusu Gigafida 2.0 (Krek et al. 2019).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

6 Vojko Gorjanc, Simon Krek in David Bordon: Rešitve za dolgoročno zbiranje gradiva

6.1 Uvod

Namen tega razdelka je predstaviti rešitve za dolgoročno zbiranje gradiva, ki so vključevale:

- pripravo pravnih rešitev za zbiranje besedil;
- izdelavo spletnega portala za oddajo besedil in
- celostno analizo mreže potencialnih besedilodajalcev.

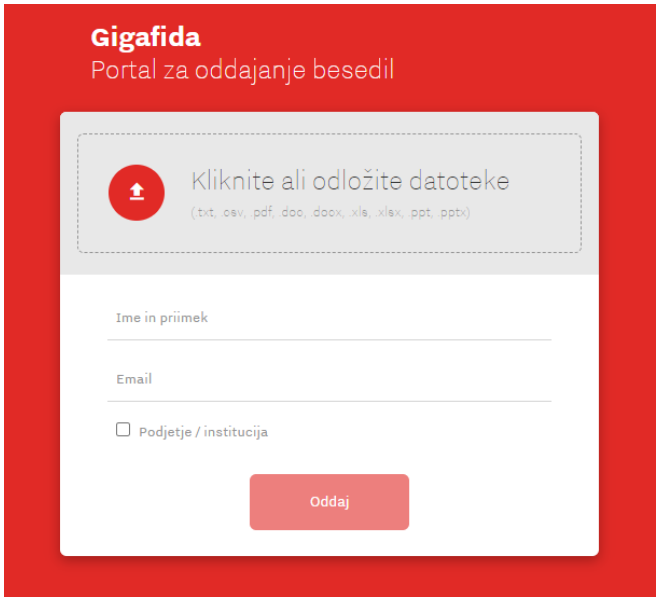
6.2 Pravne rešitve

V projektu je bila za zbiranje besedil, ki bodo vključena v korpus Gigafida 2.0, v sodelovanju s pravnimi strokovnjaki, pripravljena pogodba, ki se bo podpisovala z vsemi nosilci avtorskih pravic besedil, vključenih v korpus. Dosedanje rešitve so bile oblikovane za posamezne projekte, v okviru projekta RSDO pa je pogodba pripravljena tako, da je neodvisna od posameznega projektnega dela in bo na voljo za vse nadaljnje korpusne projekte (ob predpostavki o nespremenljivosti zakonodaje o avtorskih in sorodnih pravicah v segmentih, ki urejajo razmerja do avtorjev za potrebe vključevanja besedil v korpusno podatkovno bazo). Pogodba je v Prilogi 4.

6.3 Spletni portal

Med posameznimi aktivnostmi projekta RSDO je bil usklajeno pripravljen portal za oddajo besedil. Portal za zbiranje besedil za korpus Gigafida je na naslovu: https://zbiranje.cjvt.si/gigafida/.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.



Slika 1: Portal za zbiranje besedil za korpus Gigafida

6.4 Mreža besedilodajalcev

Pripravljena je tabela (Priloga 2) s seznamom celotne mreže potencialnih besedilodajalcev z vsemi kontaktnimi informacijami ter izbor besedilodajalcev za prioritetno urejanje pravnih razmerij (Priloga 3), saj gre za besedilodajalce, pri katerih je že vzpostavljen sistem za redno pridobivanje besedil. Na voljo so vsi podatki za kontaktiranje besedilodajalcev in urejanje pravnih razmerij z njimi za nadaljnje nadgrajevanje korpusa.

7 Vojko Gorjanc in David Bordon: Nadgradnja korpusa

7.1 Uvod

Razmislek o nadgradnji korpusa Gigafida umestimo v razvoj referenčnega korpusa od prvega korpusa Fida do aktualnega Gigafida 2.0. V tem kontekstu pokažemo na razmerje med besedili z večjim dosegom, za katerega so že pripravljene tehnične rešitve za stalno zbiranje, in regionalno razpršenostjo velikega števila besedil, ki imajo sicer regionalni doseg, a znotraj regionalnosti dosegajo praktično vse prebivalce regije, hkrati pa zanje velja, da je proces zbiranja teh besedil kompleksnejši in zamudnejši.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Kazalnik je povezan zgolj z razmislekom o zbiranju in vključevanju besedil v korpus, ne pa o tehničnih rešitvah njihove vključitve, torej procesiranju in označevanju.

7.2 Od Fide k FidiPLUS in Gigafidi

Korpus FIDA se je pri zajemanju skušal držati vrste mednarodnih priporočil in izkušenj dotedanje gradnje referenčnih korpusov. Glede na podatke Mediane o slovenskih tiskanih medijih in njihovi branosti pa je upošteval posebnost slovenskega prostora. Za vključevanje besedil je bila oblikovana mreža kriterijev, npr. glede na zvrst besedil, prenosnik, predmetno področje itd., pri čemer je skušal zajemanje besedil uravnotežiti tudi regionalno (Gorjanc 2005). Ker je korpus FidaPLUS pri nadgradnji korpusa glede vključevanja besedil sledil oblikovani mreži kriterijev za zajem raznoterih besedil glede na vrsto predvsem besediloslovnih in sociolingvističnih parametrov, se je nadgrajeval z istimi postopki uravnoteževanja korpusa kot korpus Fida (Arhar in Gorjanc 2007: 98). V razvojni fazi referenčnega korpusa načelom uravnoteževanja sledi tudi korpus Gigafida, pri katerem se glede meril za zajem korpusov izpostavlja besedilne zvrsti in vrste, značilnosti tvorca ter naslovnika, (ne)fikcijskost vsebine ipd., pa tudi prenosnik in besedilna tema oz. predmetno področje (Logar Berginc in Ljubešić 2013: 79).

7.3 Od Gigafide do Gigafide 2.0

Pri pripravi korpusa Gigafida 2.0 se prekine z dotedanjo tradicijo zbiranja besedil in uravnoteževanja korpusa, saj so se gradiva zbirala selektivno in ciljno, in sicer glede na ugotovljene pomanjkljivosti obstoječih korpusov, ne pa na splošno zbiranje vseh gradiv po načelih, ki so bila uporabljena pri gradnji korpusa Gigafida in njegovih predhodnikov, korpusov FIDA in FidaPLUS (Krek et al. 2016: 200). Za nadgradnjo korpusa Gigafida se je ciljno zbiralno besedila, za katera je bilo glede na tip, zvrst ali druge kriterije po analizi korpusov Gigafida in Kres ugotovljeno, da so v korpusu podreprezentirana, ter besedila izbranih spletnih besedilodajalcev z veliko besedilno produkcijo, kot so besedila novičarskih portalov, dnevno časopisje ipd. Pri prvi kategoriji je šlo na eni strani za šolska gradiva (učbenike, delovne zvezki ter druga sorodna pedagoška gradiva), na drugi pa za leposlovna besedila, zlasti novejše izdaje ter tista besedila, ki so imela visoko branost ali visoko besedilno produkcijo (Krek et al. 2019: 5–6).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

7.4 Gigafida 2.0 in njena nadgradnja

Pri korpusu Gigafida 2.0 se je veliko pozornost posvetilo med drugim tudi ciljnemu zbiranju besedil izbranih spletnih besedilodajalcev z veliko besedilno produkcijo (Krek et al. 2019: 5). Čeprav je jasno, da so spletna besedila danes tista, ki prevzemajo ključno mesto tudi pri dosegu potencialnih bralcev, pa analiza celostne podobe mreže besedilodajalcev kaže, da je prevladujoči način razširjanja še vedno tiskani (Diagram 2), kar pomeni, da je vsaj v določenem segmentu korpusne nadgradnje treba posvetiti pozornost tudi razmisleku o zajemu medijev z vidika načina razširjanja, torej glede na prenosnik, kar je bila ena od osrednjih tem razmisleka pri korpusu Gigafida, takrat v veliki meri zaradi vse večje razširjenosti interneta in prenosa medijskih vsebin na splet (Logar Berginc in Ljubešić 2013: 81–85). Razmislek je povezan tudi z razmislekom o zastopanosti besedil glede na regionalno pripadnost založnikov in regionalni/lokalni doseg medijev.

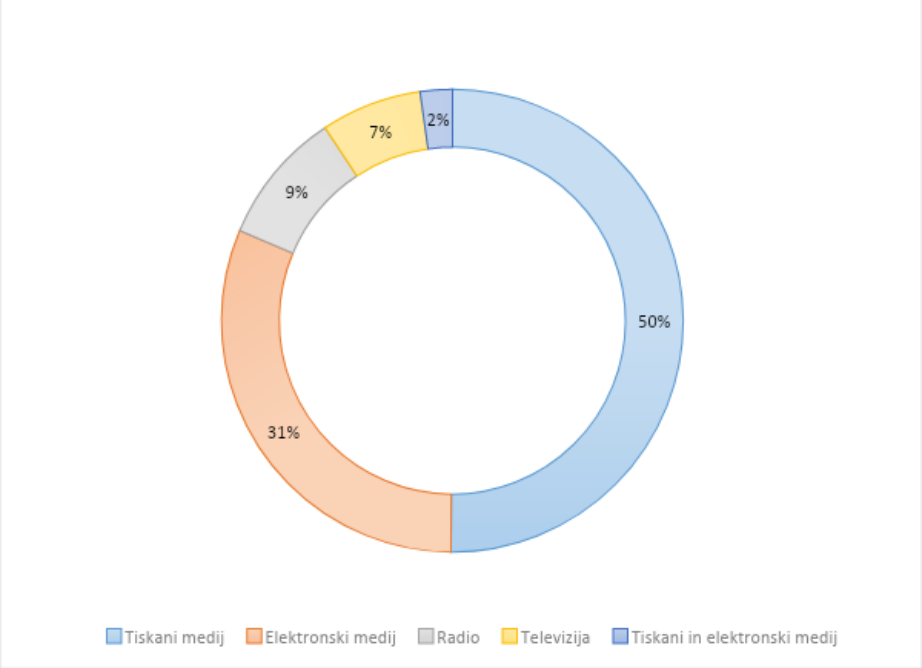
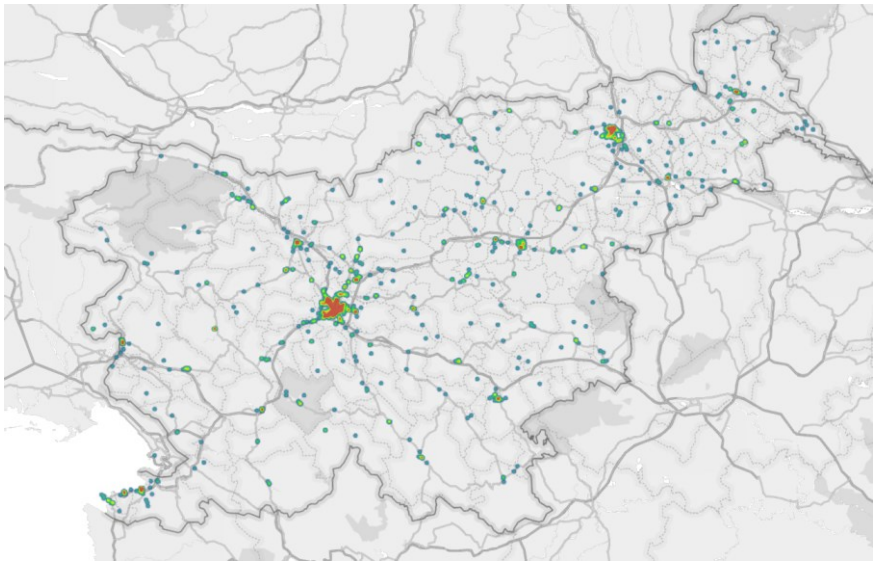


Diagram 2: Tip medija glede na način razširjanja

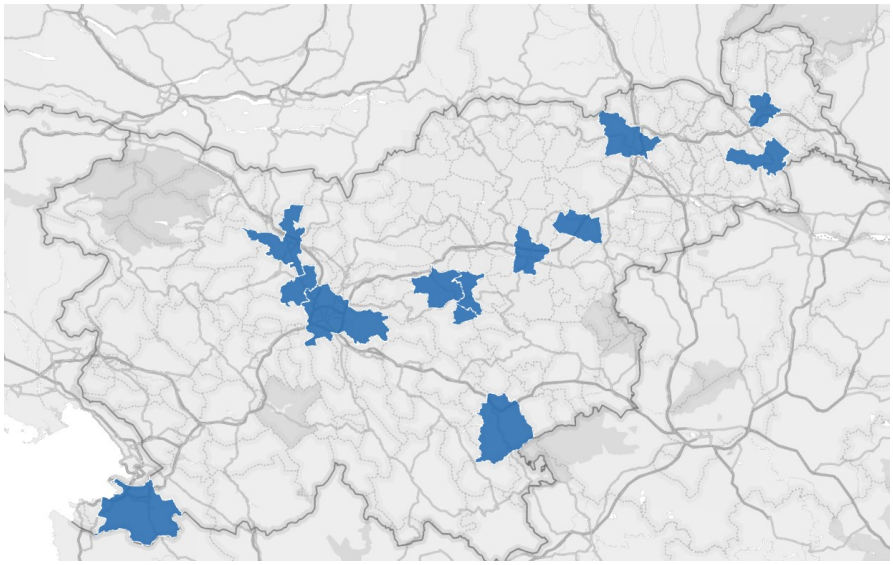
Pregled celotne mreže potencialnih besedilodajalcev kaže izjemno geografsko razpršenost z urbanimi centri, v katerih je logično koncentrirano večje število njihovih uredništev (Slika 2).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.



Slika 2: Geografska umestitev uredništev besedilodajalcev

Koncentracija uredništev na zgornji Sliki 2 je prekrivna z geografsko umeščenostjo tistih besedilodajalcev, pri katerih je tehnično že vzpostavljen stalni dotok besedil, po večini pa gre prav za spletne medije z veliko besedilno produkcijo (Slika 3).

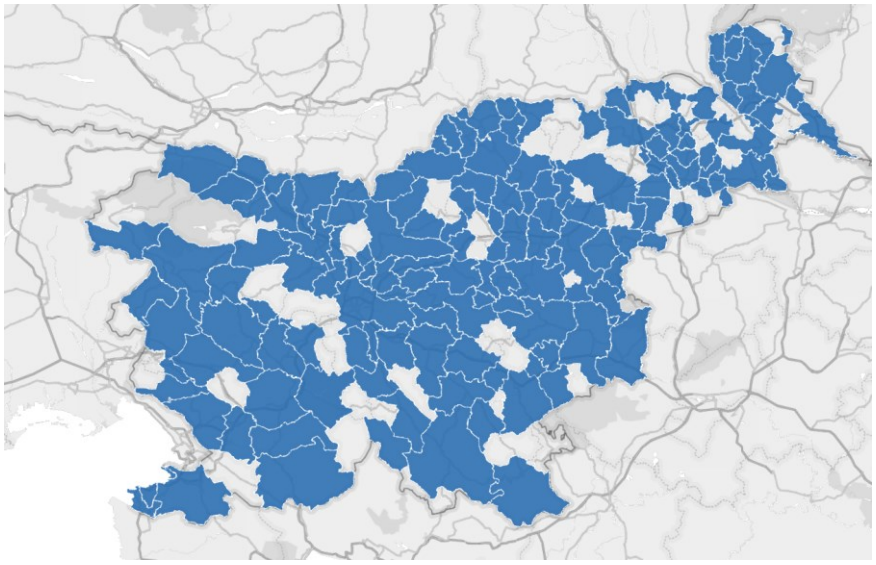


Slika 3: Geografska umeščenost uredništev besedilodajalcev s tehnično pripravljenim stalnim dotokom besedil

Čeprav zaradi narave spleta to pomeni tudi njihov potencialno velik doseg, pa je pri zajemu besedil v zadnjih različicah korpusa popolnoma zanemarjena regionalna razpršenost medijev. Čeprav ti nimajo

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

nujno velikega dosega, pa imajo lahko na lokalni ravni praktično največji možni doseg, kar velja za občinska glasila, pri katerih je prevladujoči medij tiskani, dosežejo pa vsako gospodinjstvo. Če umestimo sedeže uredništev geografsko v sliko slovenskih občin (Slika 4), vidimo, da so le redke brez medijske prisotnosti, v veliki večini občin pa je prisoten medij, ki ima glede na svojo intenco po dosegu (občinske meje) izjemen doseg.



Slika 4: Geografska umeščenost uredništev besedilodajalcev v občinske meje

Trenutna različica korpusa dejansko predstavlja centralno centralni standard. Glede na odločitev pri zadnji nadgradnji korpusa za vključevanje podreprezentoranih besedil bi naslednja varianta korpusa morala posebno pozornost posvetiti trenutno podreprezentiranim besedilom, to pa pomeni zajem besedil glede na regionalno razpršenost besedilodajalcev. Taka odločitev je skladna tudi s sociolingvistično sliko slovenskega prostora, ki izkazuje veliko regionalno pestrost, za katero pa pri standardni varianti nimamo ustreznih podatkov.

Glede na dosedanjo prakso je nadaljevanje gradnje glede vključevanja besedil in doseganje uravnoteženosti korpusa možno v dve smeri:

- dodajanje podreprezentiranih besedil novim različicam korpusa, kar vedno lahko vodi v novo podreprezentacijo drugih besedil;
- sistematičen pristop uravnoteževanja, značilen za prve variante korpusa.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Odločitev je povezana tako s finančnim kot uredniškim pristopom, sistematičen pristop uravnoteževanja je namreč mogoč le ob stabilnem finančnem in uredniškem pristopu k nadgrajevanju korpusa.

Glede na trenutno stanje je smiselno nadaljnje delo pri nadgradnji usmeriti v naslednje:

- urediti pravna razmerja z besedilodajalci, pri katerih je tehnično urejena možnost nenehnega pridobivanja besedil;
- zagotoviti regionalno uravnoteženost korpusa tako zaradi podreprezentiranosti teh besedil kot glede na slovensko jezikovno situacijo in sociolingvistično sliko slovenskega diskurznega prostora;
- sprejeti načelno odločitev o načinu uravnoteževanja korpusa pri njegovem nadgrajevanju.

Brez tovrstne odločitve je podrobnejši načrt nesmiseln. Glede na neurejeno in nestabilno financiranje pa je mogoč tudi načrt z alternativnimi variantami.

8 Vojko Gorjanc: Ocena uspešnosti in strategija nadaljnjega razvoja

Zastavljena naloga je bila glede na zgornje poročilo opravljena skladno z načrti. Čeprav se v nekaterih segmentih časovno ni v celoti pokrivala z načrtovanim delom, so si posamezne naloge še vedno sledile dovolj časovno usklajeno, da je bil dosežen želeni končni rezultat.

V nadaljevanju navajam nekaj razmislekov glede projektnih spoznanj in nekatere ideje za nadaljnje raziskovalno delo v tem tematskem okviru.

Rezultati opravljene analize uporabniških potreb in pričakovanj so zelo jasno izpostavili segmente, katerim je treba v nadaljnjem razvoju korpusa posvetiti posebno pozornost, hkrati pa so odlično izhodišče za nadaljnje uporabniške študije tudi z drugimi analitičnimi tehnikami in metodami, npr. s polstrukturiranimi intervjuji ali sledenjem zaslona uporabnikom iz različnih uporabniških skupin.

Pri kazalniku o pripisovanju standardnosti besedil se zdi, da so bolj kot sami podatki o ustreznosti avtomatičnega pripisovanja standardnosti zanimivi podatki o razlogih za nestandardnost, ki odpirajo nekatera konceptualna vprašanja definiranja standardnega jezika. Čeprav se ta zdi na načelni ravni

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

dovolj jasno konceptualiziran, pri konkretnem opazovanju besedil trči na nekatere težave, predvsem ko gre za ocenjevanje celotnega besedila v razmerju do delov besedila, še posebej pa odpira vprašanje, kako v tem kontekstu opredeljujemo kreativno rabo jezika.

Za slovensko jezikovno skupnost je razmislek o zamejskem korpusu izjemnega pomena. Čeprav že več desetletij govorimo o skupnem slovenskem kulturnem prostoru, se hkrati zavedamo njegove jezikovne raznolikosti tudi na ravni jezikovnega standarda. Dosedanji referenčni korpusi so sistematično predstavljali osrednji jezikovni standard, tj. standardni jezik v Republiki Sloveniji. Za celovito sliko vpogleda v stanje slovenskega jezikovnega standarda pa je nujno potrebno to sliko dopolniti. Glede na pripravljen predlog zajema besedil v zamejstvu v Republiki Italiji bi bil smiseln prvi korak realizacija tega korpusa, z izkušnjami njegove gradnje pa nadaljevanje z dodajanjem še drugih zamejskih korpusov.

Projekt RSDO je pripravil odlične rešitve za dolgoročno zbiranje besedil, ki omogočajo stalno nadgradnjo korpusa, a zavedati se moramo, da so še tako odlične rešitve zgolj izhodišče za dejansko nadgradnjo korpusa. Poleg tega bo nujno sprejeti načelno odločitev o načinu uravnoteževanja korpusa pri njegovem nadgrajevanju. Brez tovrstne odločitve je podrobnejši načrt nesmiseln. Glede na neurejeno in nestabilno financiranje je mogoč tudi načrt z alternativnimi variantami. Ključna za nadaljnje delo je torej rešitev vprašanje stalnega infrastrukturnega financiranja njegove konstantne rasti, ki bo neodvisna od projektnega financiranja. Šele stabilno infrastrukturno financiranje bo namreč omogočilo tudi stabilnejši uredniški pristop, torej tudi vsebinsko stabilnejše stalno ukvarjanje z nadgrajevanjem korpusa, hkrati pa bo naslovilo temeljno pričakovanje uporabnikov, izraženo v analizi njihovih potreb in pričakovanj: redno posodabljanje korpusa z novimi besedili.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

9 Priloge

Priloga 1: Uporabniška anketa in rezultati o Gigafidi 2.0 in Janesu[4]

Priloga 2: Tabela besedilodajalcev s kontaktnimi podatki[5]

Priloga 3: Tabela besedilodajalcev za prioritetno urejanje pravnih razmerij[6]

Priloga 4: Pogodba za besedilodajalce Gigafida

10 Literatura

Arhar in Gorjanc 2007 = Arhar, Špela, Gorjanc, Vojko. Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2 (2007). 95–110.

Bogatec 2020 = Bogatec, Norina (ur.): *Slovene. The Slovene language education in Italy*. Regional Dossiers series. Ljouwert/Leeuwarden: Mercator European Research Centre on Multilingualism and Language Learning. 2020.

Franza 2018 = Franza, Jasmin: *Načrt za korpusnojezikoslovni opis slovenščine v Italiji*. Magistrsko delo. Univerza v Ljubljani, Filozofska fakulteta. 2018.

Gorjanc 2005 = Vojko Gorjanc: *Uvod v korpusno jezikoslovje*. Domžale: Izolit. 2015.

Gorjanc et al. 2015 = Gorjanc, Vojko, Krek, Simon, Popič, Damjan. Med ideologijo knjižnega in standardnega jezika. Gorjanc; Vojko et al. (ur). *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete. 2015. 32–48.

Grčar et al. 2012 = Grčar, Miha, Krek, Simon, Dobrovoljc, Kaja. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: ERJAVEC, Tomaž (ur.), ŽGANEC GROS, Jerneja (ur). *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 2012. 89–94. http://nl.ijs.si/isjt12/proceedings/isjt2012_17.pdf.

[4] https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/03/RSDO_Kazalnik_Gigafida_Priloga1.pdf
[5] https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/03/RSDO_Kazalnik_Gigafida_Priloga2.pdf
[6] https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/03/RSDO_Kazalnik_Gigafida_Priloga3.pdf

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Grgič 2017 = Grgič, Matejka: Italijansko-slovenski jezikovni stik med ideologijo in pragmatiko. Jezik in slovstvo 62/1 (2017). 89–98.

Jagodic 2017 = Jagodic, Devan. Slovenci v Italiji: poselitveni prostor in demografsko gibanje. V: Bogatec, Norina (ur.), Vidau, Zaira (ur.). *Skupnost v središču Evrope. Slovenci v Italiji od padca Berlinskega zidu do izzivov tretjega tisočletja.* Trst: Slovenski raziskovalni inštitut, Založništvo tržaškega tiska. 2017. 40–49.

Jagodic, Kaučič-Baša in Dapit 2017 = Jagodic, Devan, Kaučič-Baša, Majda, Dapit, Roberto. Jezikovni položaj Slovencev v Italiji. V: Bogatec, Norina (ur.), Vidau, Zaira (ur.). *Skupnost v središču Evrope. Slovenci v Italiji od padca Berlinskega zidu do izzivov tretjega tisočletja*. Trst: Slovenski raziskovalni inštitut, Založništvo tržaškega tiska. 2017. 67–88.

Krek et al. 2016 = Krek, Simon, Gantar, Polona, Arhar Holdt, Špela, Gorjanc, Vojko. Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres. V: Erajvec, Tomaž (ur.), Fišer, Darja (ur). *Zbornik conference Jezikovne tehnologije in digitalna humanistika, 29. September do 1. oktober 2016, Ljubljana, Slovenija.* Ljubljana: Znanstvena založba Filozofske fakultete. 2016. 200–203. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Krek-et-al_Nadgradnja-korpusov-Gigafida-Kres-ccGigafida-ccKres.pdf.

Krek et al. 2019 = Krek, Simon, Arhar Holdt, Špela, Čibej, Jaka, Repar, Andraž, Ljubešić, Nikola. *Specifikacije izdelave korpusa Gigafida 2.0 // Gigafida 2.0 Corpus Compilation: Specifications*. Center za jezikovne vire in tehnologije Univerze v Ljubljani. 2019. https://www.cjvt.si/gigafida/wp-content/uploads/sites/10/2019/06/Gigafida2.0_specifikacije.pdf.

Krek et al. 2020 = Krek, Simon, Arhar Holdt, Špela, Erjavec, Tomaž, Čibej, Jaka, Repar, Andraž, Gantar, Polona, Ljubešić, Nikola, Kosem, Iztok, Dobrovoljc, Kaja. Gigafida 2.0: the reference corpus of written standard Slovene. V: Calzolari, Nicoletta (ur.). *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11-16, 2020, Marseille, France.* Pariz: ELRA – European Language Resources Association. 2020. 3340–3345. http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Ljubešić et al. 2015 = Ljubešić, Nikola, Fišer, Darja, Erjavec, Tomaž, Čibej, Jaka, Marko, Dafne, Pollak, Senja, Škrjanec, Iza. Predicting the level of text standardness in user-generated content. International Conference on Recent Advances in Natural Language Processing: *Proceedings of RANLP 2015 Conference*, 7-9 September 2015, Hissar, Bulgaria. 2015. 371–378.

Ljubešić in Erjavec 2016 = Ljubešić, Nikola, Erjavec, Tomaž. Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. V: Calzolari, Nicoletta (ur.). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Pariz: ELRA – European Language Resources Association. 2016. 1527–1531. http://www.lrec-conf.org/proceedings/lrec2016/pdf/811_Paper.pdf.

Logar Berginc in Ljubešić 2013 = Nataša Logar Berginc, Nikola Ljubešić. Gigafida in slWaC: tematska primerjava. *Slovenščina 2.0*, 1/1 (2013). 78–110.

Logar et al. 2023 = Logar, Nataša, Arhar Holdt, Špela, Gorjanc, Vojko. Korpus Gigafida 2.0: mnenja uporabnikov. *Jezik in slovstvo*. Oddano v recenzijski postopek.

Stranj 1999 = P. Stranj: *Slovensko prebivalstvo Furlanije - Julijske krajine v družbeni in zgodovinski perspektivi*. Trst: Slovenski raziskovalni inštitut, Narodna in študijska knjižnica, Znanstveni inštitut Filozofske fakultete. 1999.

Tuta 2017 = Tuta, Igor. Sredstva javnega obveščanja v slovenskem jeziku v Italiji. V: Bogatec, Norina (ur.), Vidau, Zaira (ur.). *Skupnost v središču Evrope. Slovenci v Italiji od padca Berlinskega zidu do izzivov tretjega tisočletja*. Trst: Slovenski raziskovalni inštitut, Založništvo tržaškega tiska. 2017. 157–166.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Priloga 4: Pogodba za besedilodajalce Gigafida

Univerza v Ljubljani, Kongresni trg 12, 1000 Ljubljana, ki jo zastopa **[naziv, ime, priimek]**; matična številka: 5085063000, davčna številka: 54162513 (v nadaljnjem besedilu: **naročnik**)

in

avtor oz. imetnik pravic na avtorskem delu _____ , ki ga zastopa

_____ (v nadaljnjem besedilu: **imetnik pravic**)

sklepata naslednjo

P O G O D B O

o zbiranju in uporabi besedilnega korpusa v okviru infrastrukturnih nalog Centra za jezikovne vire in tehnologije Univerze v Ljubljani

1. Pogodbeni stranki uvodoma ugotavljata:

a) da naročnik pripravlja **korpus pisne standardne slovenščine Gigafida** (v nadaljnjem besedilu: **korpus Gigafida**), kar zajema zbiranje besedil različnih vrst za namene elektronske analize, obdelave, označevanja, reproduciranja in druge uporabe njihovih besed, besednih zvez ali povedi;
b) da tehnične in infrastrukturne naloge v zvezi s korpusom Gigafida izvaja **Center za jezikovne vire in tehnologije** (v nadaljnjem besedilu: **CJVT UL**), ki deluje v okviru Univerze v Ljubljani;
c) da se raziskovalne in vsebinske naloge v zvezi s korpusom Gigafida izvajajo v okviru **raziskovalnih programov Slovenski jezik – bazične, kontrastivne in aplikativne raziskave (P6-0215)** in **Jezikovni viri in tehnologije za slovenski jezik (P6- 0411)** na Filozofski fakulteti Univerze v Ljubljani;
d) da imetnik pravic razpolaga z avtorskimi pravicami, avtorski sorodnimi pravicami in drugimi pravicami avtorja v skladu z Zakonom o avtorski in sorodnih pravicah (Uradni list RS, št. 16/07 – uradno prečiščeno besedilo, 68/08, 110/13, 56/15, 63/16 – ZKUASP in 59/19; v nadaljevanju: ZASP) na avtorskih delih, ki so predmet te pogodbe (v nadaljnjem besedilu: **delo**) in so navedena v prilogi k tej pogodbi.

2. Imetnik pravic na naročnika prenaša naslednje pravice na delih, ki so predmet te pogodbe: pravico reproduciranja, vključno s pravico shranitve v elektronski obliki iz 23. člena ZASP in pravico predelave teh del iz 33. člena ZASP. Ti pravice se prenašajo na naročnika **neizključno** in **neodplačno**; prenos je brez časovnih omejitev ter velja **za namene korpusa Gigafida**, v obsegu, ki bo naročniku omogočil pripravo korpusov oziroma izvedbo tega projekta. Imetnik pravic bo naročniku omogočil dostop do del v digitalni obliki preko nosilcev (DVD, trdi disk ipd.), preko spleta ali na drug primeren način.

3. Imetnik pravic dovoli naročniku, da se za namen nadgradnje korpusa ***"ccGigafida"*** do 10% vsakega dela uporabi in vključi v navedeni korpus. Za ta namen imetnik pravic na naročnika neizključno, neodplačno in brez časovnih ter teritorialnih omejitev prenaša zlasti pravico reproduciranja (23. člen ZASP), distribuiranja (24. člena ZASP), dajanja v najem (25. člen ZASP), priobčitve javnosti in predelave avtorskih del (33. člen ZASP), ki so predmet te pogodbe, in na predelavah teh del. Imetnik pravic soglaša z nadaljnjim prenosom materialnih pravic, navedenih v prejšnjem stavku tega člena, na tretje osebe. Imetnik pravic soglaša, da se avtorska dela, ki so predmet te pogodbe, dajo na voljo javnosti pod pogoji licence Creative Commons 4.0 »priznanje avtorstva«. Imetnik pravic je seznanjen, da bo korpus ***"ccGigafida*** na voljo javnosti pod pogoji licence Creative Commons 4.0 "priznanje avtorstva". Ta licenca omogoča uporabnikom avtorsko delo in njegove predelave reproducirati, distribuirati, dajati v najem, priobčiti javnosti in predelovati, vendar samo pod pogojem, da navedejo avtorja tega dela.

4. Imetnik pravic jamči, da razpolaga z avtorskimi oziroma sorodnimi pravicami na delu, da na njem ne obstajajo pravice tretjih oseb, ki bi bile v nasprotju s to pogodbo, in da ne obstajajo katere druge pravice na delu, ki bi naročniku preprečevale nameravano uporabo dela po tej pogodbi.

5. Določbe te pogodbe ne vplivajo na prenos moralnih avtorskih pravic, ki so v skladu z določbami ZASP neprenosljive.

6. Naročnik se zavezuje:

a) da bo delo naložil v spominske enote, namenjene za gradnjo **korpusa Gigafida**, morebitne začasne nosilce dela pa bo nato na zahtevo imetnika pravic ali izbrisal ali uničil ali vrnil imetniku pravic;
b) da bo delo konvertiral ter uporabljal izključno za namene gradnje **korpusa Gigafida** in v skladu s to pogodbo;
c) da bo morebitni začasni nosilec dela v času od prejema do njegovega brisanja ali uničenja ali vrnitve skrbno varoval, da ga ne bo dal na razpolago komu tretjemu z namenom, da bi ga ta uporabljal izven obsega te pogodbe.

7. Za namen izvrševanja te pogodbe obe pogodbeni stranki imenujeta skrbnika pogodbe:

- za imetnika pravic: _____
- za naročnika: _____

8. Pogodbeni stranki soglašata, da se za vse, kar v tej pogodbi ni urejeno, uporabljajo določila ZASP in Obligacijskega zakonika. Morebitne spore, izvirajoče iz te pogodbe, bosta pogodbeni stranki reševali na sporazumen način. Če to ni mogoče oziroma do sporazumne rešitve ne pride, je za reševanje spornih zadev pristojno sodišče v Ljubljani. Pogodba je sestavljena v dveh (2) izvodih, od katerih prejme vsaka izmed pogodbenih strank po en (1) izvod. Pogodba prične veljati z dnem podpisa obeh pogodbenih strank,

V Ljubljani, dne _____ _____

Naročnik: **Imetnik pravic**:

_____ _____

Priloga k pogodbi*:* Seznam avtorskih del oziroma predmetov sorodnih pravic imetnika pravic

Ime, naslov ali oznaka dela