Trajnostni načrt razvoja jezikovnih virov in tehnologij

Razvoj slovenščine v digitalnem okolju (RSDO)
DS1: Vzdrževanje in nadgradnja korpusov (jezikovni viri)

V okviru projekta Razvoj slovenščine v digitalnem okolju (RSDO) (2020-2023) je v Delovnem sklopu 1 (DS1) potekala izdelava več temeljnih jezikovnih virov (korpusov in učnih množic) ter temeljnih tehnologij (predvsem orodij za jezikovno označevanje virov). Rezultati projekta so razvite vire in tehnologije postavili na bistveno višjo raven, kot so bili v začetku projekta. S skokovitim razvojem tehnologij umetne inteligence, katere eno glavnih področij je tudi obdelava naravnega jezika, pa se kaže potreba po nadaljnjem vzdrževanju in nadgradnji razvitih virov ter razvoju novih virov in tehnologij za slovenščino. Pričujoči dokument predstavlja trajnosti načrt za nadaljnji razvoju jezikovnih virov slovenskega jezikav prihodnjih letih. Ker upošteva rezultate projekta RSDO in skokoviti napredek na področju umetne inteligence v zadnjih letih, je sodobnejši od obstoječih nacionalnih dokumentov. Predstavlja sintezo vidikov več znanstveno-raziskovalnih področij: jezikoslovnega, družboslovnega in področja umetne inteligence.

Trajnostni načrt je razdeljen na dva razdelka, ki ustrezata vsebinsko zaokroženim celotam. V razdelku 1 predstavi orodja in tehnologije za osnovno obdelavo in temeljno označevanje besedil: modele strojnega učenja, učne množice, ki so potrebne za njihovo učenje, ter leksikon slovenskega besedišča Slokes, ki prispeva k temeljnemu procesiranju besedil. Razdelek 2 načrtuje nadaljnji razvoj temeljnih nacionalnih jezikovnih zbirk: korpusa standardne in spletne slovenščine Gigafida in Janes, korpusa govorjene slovenščine GOS, korpusov šolskih besedil Šolar in KOST, korpusa parlamentarnega govora siParl in metakorpusa, ki omogoča poizvedovanje po vseh ostalih korpusih. Opredeli tudi nove, pomembne, vendar v slovenščini še ne obstoječe korpuse. V zadnjem, 3. razdelku, povzamemo najpomembnejše ugotovitve in orišemo povezave z drugimi delovnimi sklopi projekta RSDO.

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.

Kazalo

1. Osnovna obdelava in jezikoslovno označevanje besedil

Osnovna obdelava naravnega jezika potrebuje orodja, ki omogočajo razbitje besedila na odstavke, stavke in besede, lematizacijo besed, oblikoskladenjsko označevanje stavkov in njihovo skladenjsko razčlenjevanje, označevanje imenskih entitet, označevanje semantičnih vlog in iskanje koreferenc. Za vsa našteta orodja je projekt RSDO zgradil cevovod, katerega nadaljnji razvoj načrtujemo v razdelku 1.1. Večino orodij v tem cevovodu uporablja modele strojnega učenja, ki temeljijo na globokih nevronskih mrežah. Modele strojnega učenja je potrebno za njihovo delovanje naučiti, za kar so potrebne ročno označene učne množice. Nekatera orodja, predvsem lematizator, potrebujejo tudi leksikon, ki vsebuje besede v vseh oblikah. Načrt za nadaljnji razvoj učnih množic je opisan v razdelku 1.2, leksikona pa v razdelku 1.3.

1.1 Cevovod za obdelavo in označevanje besedil

V okviru projekta RSDO je bil razvit cevovod CLASSLA-Stanza za osnovno obdelavo in temeljno jezikoslovno označevanje slovenskih besedil. Programska koda cevovoda CLASSLA-Stanza je izšla (angl. fork) iz mednarodnega projekta Stanza, ki združuje enotno osnovo za jezikoslovno procesiranje več sto jezikov in temelji na zdaj že nekoliko starejši arhitekturi nevronskih mrež LSTM (angl. Long short term memory networks), ki jo v zadnjem času nadomešča arhitektura transformer.
Za nadaljnji razvoj temeljne obdelave je potrebno naslednje.

- **Vzdrževanje trenutne programske kode cevovoda CLASSLA-Stanza.**
  Cevovod CLASSLA-Stanza je dosegel visoko stopnjo zrelosti, ki jo je za nekatere naloge (segmentacija, lematizacija, oblikoskladenjsko označevanje) mogoče le še malo izboljšati in se že bliža točnosti človeškega jezikoslovnega označevanja. Pri drugih, naprednejših označevalnih nalogah (prepoznavanje imenskih entitet, označevanje udeleženskih vlog, določanje koreferenčnosti), je prostora za napredek še precej. Zaradi združljivosti za nazaj in hitrosti procesiranja (trenutna arhitektura CLASSLA-Stanza zahteva manj računske moči kot novejša

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.

arhitektura transformer) je potrebno cevovod CLASSLA-Stanza redno vzdrževati, posodabljati z novimi verzijami mednarodne osnove Stanza in odpravljati v bodoče odkrite napake.

- **Razvoj nove generacije označevalnih orodij arhitekture transformer.**
  Za večino nalog obdelave naravnega jezika od leta 2019 naprej arhitektura nevronskih mrež transformer nadomešča prejšnje arhitekture. Tudi za slovenščino je potrebno izvesti ta prehod, saj so nekateri posamezni testi pokazali, da so modeli arhitekture transformer, npr. SloBERTa (razvita v projektu RSDO) uspešnejši, še posebej za zahtevnejše naloge razumevanja jezika. Trenutno se kot solidna osnova za nadaljnji razvoj slovenskih orodij kaže odprtokodno orodje Trankit, ki uporablja modele arhitekture transformer. Novo generacijo orodij bo potrebno naučiti na učni množici SUK (razdelek 1.2) in njenih bodočih izboljšavah.
- **Ansambli napovednih modelov za označevanje**
  Različne generacije tehnologij bodo pričakovano komplementarne, zato bo potrebno zgraditi modele, ki bodo izkoriščali dobre lastnosti vseh obstoječih modelov (t.i. ansambelski modeli).
- **Modularni razvoj novih orodij**
  Razvoj jezikovnih orodij je skokovit in v prihodnosti lahko pričakujemo, da bo vse več orodij dovolj zrelo za vključitev v osnovni cevovod. Zaradi tega je treba novo generacijo orodij zasnovati modularno, da bo mogoče vključiti tudi nova orodja in nove ravnine jezikoslovnega označevanja (npr. razdvoumljanje besed, metaforičnost, sentiment, koreferenčnost itd.).
- **Tehnološko neodvisni programski vmesnik**
  Za trenutni cevovod CLASSLA-Stanza, nov cevovod na arhitekturi transformer in potencialne nove nevronske tehnologije je potrebno razviti skupni programski vmesnik (API) in spletni programski vmesnik (REST API), ki bosta omogočala uporabo, neodvisno od trenutno prevladujoče tehnologije, kar bo omogočilo dolgoročno trajnost programske opreme, ki bo uporabljala te vmesnike. Bodoče postopne izboljšave in nadgradnje temeljnega cevovoda za jezikovno obdelavo tako ne bodo vplivale na strukturo uporabniške programske opreme, bodo pa izboljšale njeno delovanje.
- **Vključevanje novih orodij v primerjalnik SloBENCH**
  Za primerljivost različnih tehnoloških pristopov in orodij je bil zgrajen mednaroden odprt primerjalnik slovenskih jezikovnih tehnologij SloBENCH. Za vse novo razvite pristope je potrebno zahtevati njihovo vključitev v SloBENCH, kar bo pokazalo, kako uspešne so v primerjavi z obstoječimi. Za povsem nove tehnologije, ki jih še ni v okolju SloBENCH, je potrebno v okolju vzpostaviti nove primerjalne okvirje.
- **Vzdrževanje in razvoj spletnih vmesnikov**
  Tehnologije za jezikovno označevanje je potrebno narediti dosegljiva splošnim uporabnikom oz. uporabnikom brez tehničnega znanja, ne le programerjem. Primer takšne dobre prakse je oblikoskladenjski označevalnik na https://orodja.cjvt.si/oznacevalnik/slv/. Najlažji način za dosego tega cilja so spletna orodja, ki jih lahko uporabniki dosegajo brez instalacije. Zanje je potrebno razviti in vzdrževati spletne vmesnike.
- **Kvalitativne analize in razlage odločitev modelov**
  Globoke nevronske mreže, na katerih temeljijo orodja za obdelavo naravnih jezikov, so po svoji naravi netransparentne in ne znajo razložiti svojih odločitev. Razložljivost odločitev je pomembna prvina orodij umetne inteligence in ena od zahtev za njihovo rabo v EU. Za obstoječe in bodoče modele za jezikoslovno označevanje je zato potrebno financirati kvalitativne analize njihovega delovanja, ki bodo pokazale, kje so njihove napake. Financirati je potrebno raziskave na področju razložljive umetne inteligence (ang. Explainable AI), ki bodo generirale razlage odločitev globokih nevronskih mrež specializiranih za jezik in še posebej za slovenščino. Take metode so trenutno v razvoju in med drugim omogočajo odkrivanje napak in izboljšanje transparentnosti napovednih modelov.
- **Sistematično popravljanje napak**
  Dosedanji razvoj je pokazal, da pri nekaterih nalogah jezikovne obdelave prihaja do napak, npr. zaradi premajhne vsebovanosti raznolikih primerov v učni množici modeli za

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.

oblikoskladenjsko označevanje CLASSLA-Stanza niso ločili med enakopisnima oblikama je-jesti in je-biti. Projekt RSDO je pokazal, da je takšne napake mogoče odpravljati z dodajanjem novih primerov v učne množice in dodatno učenje napovednih modelov. Na podlagi kvalitativne analize modelov je k odpravljanju zaznanih pomanjkljivosti potrebno pristopiti sistematično.

- **Priprava variantnih modelov za specifične jezikovne potrebe**
  Trenutno so modeli za jezikoslovno označevanje zgrajeni predvsem za sodobni pisni standardni jezik. Za druge vrste jezikovne rabe , kot so nestandardni spletni jezik (npr. spletne uporabniške vsebine, kot so komentarji ali tviti), starejši jezik, narečni jezik, govorjeni jezik, zamejski jezik itd. je potrebno zgrajena orodja prilagoditi na podlagi učnih množic (glej razdelek 1.2), ki bodo pripravljene specifično v te namene.§

1.2 Učne množice za izdelavo računalniških modelov

Modele strojnega učenja, ki jih vsebuje cevovod za jezikoslovno označevanje, je potrebno za njihovo delovanje naučiti na človeško označenih učnih množicah. V okviru projekta RSDO sta bila nadgrajena učni korpus ssj500k (po novem Slovenski učni korpus *SUK 1.0*) za učenje označevanja standardnega jezika ter učni korpus Janes-Tag (po novem *Janes-Tag 2.0*) za učenje označevanja nestandardnega jezika. Za njen nadaljnji razvoj in nadaljnji razvoj jezikoslovne obdelave je potrebno naslednje.

- **Kontinuirane izboljšave učnih korpusov SUK in Janes-Tag**
  Učna korpusa SUK in Janes-Tag sta temeljna učna korpusa za označevalne modele strojnega učenja, zato ju je potreno kontinuirano izboljševati in dopolnjevati z novimi primeri ter novimi označevalnimi ravninami. Posodabljati je potrebno tudi smernice za njuno označevanje ter tako smernice kot učna korpusa usklajevati z razvojem mednarodnih standardov.
- **Učne množice za sistematično popravljanje napak**
  Pri nekaterih nalogah jezikovne obdelave prihaja do napak, ki so posledica premajhne vsebovanosti nekaterih jezikovnih fenomenov v učnih množicah (redkejše dvoumne besedne oblike, redkeje zastopane kategorije, kot sta srednji spol, dvojina in podobno). Na podlagi kvalitativne analize modelov je k odpravljanju zaznanih pomanjkljivosti potrebno pristopiti sistematično in dopolniti učne množice s primeri, ki naslavljajo te pomanjkljivosti. Dopolnjene učne množice se po evalvaciji vključijo v osnovne učne korpuse, npr. SUK ali Janes-Tag.
- **Učne množice za specifične jezikovne potrebe**
  Trenutne učne množice (SUK 1.0), pokrivajo predvsem sodobni standardni pisni jezik. Treba je pokriti tudi druge oblike jezikovne rabe in zanje pripraviti ločene manjše učne množice, ki bodo omogočale podporo jezikoslovnemu cevovodu za:
  - starejši jezik od 16. pa do konca 19. stoletja, kot tudi slabo pokriti del 20. stoletja do leta 1900 do 1990, npr. specifično jugoslovansko besedišče;
  - narečni jezik, kjer je potrebno določiti vrstni red, po katerem se bodo obravnavala posamezna narečna besedišča;
  - zamejski jezik, npr. jeziki slovenskih skupnosti na avstrijskem Koroškem, Italiji, Argentini in Severni Ameriki;
  - govorjeni jezik, ki je za nekatere naloge, npr. segmentacijo stavkov in skladenjsko razčlenjevanje, bistveno težji od standardnega jezika;
  - določene domene pisnega jezika, ki se glede na raziskave jezikovnih značilnosti pomembneje razlikujejo od splošne rabe, npr. pravna in uradovalna besedila, znanstveni jezik, pisanje učečih se.
- **Učne množice za označevanje semantičnih ravnin jezika**
  Dosedanje učne obstoječe množice v zelo omejeni obliki pokrivajo nekatere naprednejše semantične koncepte, kot so relacije, koreference, razdvoumljanje, metafore itd. Tudi te nivoje je potrebno pokriti s kakovostnimi in dovolj velikimi učnimi množicami.
- **Učne množice za označevanje diskurznih ravnin jezika**
  Dosedanje učne obstoječe množice ne pokrivajo diskurznih ravnin jezika, kot je npr. ugotavljanja mnenj v diskurzu, povzemanje diskurza in tudi posebnih lastnostih spletnega

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.

diskurza, kot so oblike žaljivega in sovražnega govora. Zanje je potrebno po zgledu mednarodnih iniciativ pripraviti nove učne množice.

- **Podpora učnih množic kakovostni in javno dostopni evalvaciji**
  Za evalvacijo orodij za jezikoslovno označevanje smo zgradili portal SloBENCH, ki podatke za vsako evalvacijsko nalogo razbije na javni in skriti del. Javni del učne množice je na voljo javnosti za učenje modelov, pri skritem delu pa so na voljo le učni primeri brez oznak. Manjkajoče oznake napovedo naučeni modeli, sistem SloBENCH pa jih primerja s skritimi pravilnimi oznakami in na tej podlagi ovrednoti in rangira različne pristope k danemu problemu. Vključitev v SloBENCH in podobno delitev na javni in skriti del za potrebe evalvacije je potrebno zagotoviti tudi za vse nove javno financirane učne množice. V SloBENCH je potrebno vključiti tudi tiste ravnine jezikovnega označevanja, ki niso bile pokrite v projektu RSDO, npr. JOS-SYN skladnja in udeleženske vloge.
- **Harmonizacija smernic za jezikoslovno označevanje**
  Priprava učnih korpusov na RSDO je razkrila določene nekonsistentnosti med različnimi nivoji jezikoslovnega označevanja, ki jih je v nadaljevanju treba jezikoslovno raziskati in odpraviti. Obstoječe smernice je potrebno poenotiti, z njimi uskladiti obstoječe učne množice in jih vključiti v enotne povečane učne množice.
- **Mednarodna usklajenost in sodelovanje**
  Učne množice in označevalne sheme je potrebno usklajevati s standardizacijskimi pobudami v mednarodnem prostoru ter sodelovati pri njihovem nastajanju. Mednarodno standardizirane učne množice omogočajo, da se za slovenščino razvijajo orodja v okviru mednarodnih konzorcijev in da je slovenščina del večjezikovnih evalvacijskih množic.

1.3 Računalniški leksikon Sloleks

Sloleks je leksikon besednih oblik za slovenski jezik. V strukturirani bazi podatkov vsebuje osnovne podatke o slovenskih besedah, predvsem v katero besedno vrsto spadajo in kakšne so njihove slovnične lastnosti. Pri vsaki besedi so v bazi zabeležene vse njene pregibne oblike, skupaj z izgovorjavo. Sloleks, kot obsežna zbirka podatkov, je koristen tako za jezikovne govorce, ki jih zanima pregibanje in izgovor besed v slovenščini, kot za razvijalce jezikovnih tehnologij. Kot je pokazal projekt RSDO, je Sloleks koristen pri nekaterih nalogah, kot je lematizacija. V nadaljnjem razvoju je za Sloleks potrebno naslednje:

- **Širjenje s skupinami besed, ki trenutno niso pokrite**
  V centralni del leksikona Sloleks je treba dodati predvsem zapis števil in lastnih imen, obenem pa je v istem formatu treba razviti tudi specializirane leksikone za nestandardno besedišče (npr. spontani govor ali računalniško posredovana komunikacija), za starejše besedišče, za različna terminološka področja (npr. medicina, pravo, informacijske tehnologije) itd.
- **Sprotne letne objave novih verzij**
  Ker se jezik neprestano razvija in pridobiva nove besede, je treba za Sloleks zagotoviti sprotne letne posodobitve (npr. na podlagi spremljevalnega korpusa Trendi in drugih specializiranih korpusov) in objave novih verzij.
- **Zlogovanje**
  V oblikoslovni leksikon je treba dodati zlogovane oblike (npr. mi-za, o-pe-ra-ci-ja, u-zur-pi-ra-ti), ki se lahko uporabijo za razvoj delilnikov besed in za druge jezikovnotehnološke aplikacije, kot so npr. sistemi zaznavanja napačnega zapisa skupaj ali narazen ("na glas" in "naglas"). Za dodajanje zlogovanih oblik je treba razviti in evalvirati zlogovalnik, ki bo dovolj natančen tudi za tvorjenje zlogovanih oblik pri novo dodanih besedah.
- **Povezane iztočnice in besedne družine**
  V Sloleksu 2.0 in 3.0 imajo le nekatere iztočnice navedene tudi povezane iztočnice – načeloma gre za besedotvorno povezane besede (npr. pisati – pisatelj, maček – mačka), a so trenutno navedene povsem nesistematično, med različnimi vrstami povezav pa ni nobenih razlik; npr. na enak način sta obravnavana pisatelj – pisateljica kot pisati – pisatelj (več o tem v Čibej

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.

2021). Razviti je treba sistem, ki bo omogočal učinkovito strojno povezovanje besedotvorno povezanih iztočnic in avtomatsko tvorjenje besednih družin ter morfoloških derivacijskih drevesnic, s katerimi bo mogoče v leksikonu strojno povezovati novo dodane iztočnice. Za primer lahko vzamemo besedišče, ki je nastalo v času pandemije: *korona*, *koronski*, *protikoronski* ter *covid*, *coviden*, *postcoviden* ipd. Sistem bi tako omogočal širjenje leksikona s skupkom povezanih iztočnic, ne le s posameznimi nepovezanimi iztočnicami. S tem bi se olajšalo tudi leksikografsko delo, saj bi leksikografom omogočalo hitro obravnavo več pomensko podobnih iztočnic naenkrat.

- **Dodajanje podatkov o korpusnih frekvencah**
  V različicah 2.0 in 3.0 Sloleks za vse iztočnice in njihove pregibne oblike vsebuje podatke o njihovi absolutni pogostosti v korpusu Gigafida 2.0 – s temi podatki lahko npr. ugotovimo, kako pogosto se samostalnik »hči« v korpusu pojavlja v rodilniku ednine, kar je koristno za uporabnike_ce. Obenem nam to omogoča, da se v primeru, da se moramo za specifičen razpoznavalnik govora zaradi tehničnih omejitev omejiti na določen nabor najpogostejših besed, ki jih bo razpoznavalnik lahko razpoznaval, odločimo, da besede z izjemno redkimi pojavitvami izločimo. Sloleksu je zato treba dodati tudi podatke o absolutnih (in relativnih) frekvencah iz drugih korpusov; tako bo npr. pri gradnji razpoznavalnika govora specifično za akademski kontekst mogoče dati prednost besedam, ki so v akademskem diskurzu pogostejše oz. bolj pričakovane.
- **Popravki na nivoju naglašenih oblik**
  V različici 3.0 Sloleks vsebuje naglašene oblike (ročno potrjene za prvi 100.802 iztočnici in avtomatske za ostale iztočnice), iztočnice same pa nimajo pripisane naglasnih vzorcev na podoben način, kot so pripisani oblikoslovni vzorci (ki povedo, kako se beseda pregiba, npr. računalnik-0, računalnik-a, računalnik-u, računalnik-0, …). Na enak način je treba izdelati tudi sistem pripisovanja naglasnih vzorcev, s katerimi lahko kategoriziramo besede glede na to, kako se naglašujejo (npr. nepremični in nespremenljivi naglas: míza, míze, mízi, mízo …; nepremični in spremenljivi naglas: Rožlè, Rožléta, Rožlétu, …; premični in spremenljivi: móž, možá, móžu, možá, …). Izdelava naglasnih vzorcev v strojno berljivi obliki bo omogočila tudi izboljšave trenutnega naglaševalnika, ki je zasnovan tako, da naglašuje vsako besedno obliko posebej. Naglaševalnik na podlagi naglasnih vzorcev bo iztočnice naglaševal bolj konsistentno, prilagojen pa bo tudi dvo- in večnaglasnim besedam (npr. mákroekonómika), na katerih trenutni naglaševalnik ni bil naučen.
- **Kratice**
  Kratice (npr. RTV, FBI, SCT) so poseben primer in jih trenutna različica Pregibalnika ne podpira, saj zahtevajo posebno obravnavo; pri kraticah je namreč pregibna oblika odvisna od načina izgovorjave, ki pa ni razviden iz same leme. Kratico "OZN" npr. lahko pregibamo kot "OZN-a", če je izgovorjava /ozéêna/, ali kot "OZN-ja", če je izgovorjava /ozənəja/, s polglasnikom. Za Pregibalnik je torej treba razviti ločen modul, ki ustrezno tvori oblike in izgovorjave za kratice.
- **Večbesednost**
  V različici 3.0 Sloleks vsebuje zgolj enobesedne iztočnice, v prihodnjih različicah pa je nujno treba določiti, kako so v njem obravnavane tudi večbesedne enote, npr. večbesedne imenske entitete (Slovenj Gradec, New York) in druge večbesedne enote (streljati kozle, iti rakom žvižgat) oz. enote, ki jih glede na tokenizacijska pravila obravnavamo kot večbesedne (npr. DMA-krmilnik). Ko bo način obravnave večbesednih enot določen, bo vse, kar izpolnjuje kriterije za ločeno iztočnico, v Sloleksu vneseno pod lasten vnos.

2. Nacionalno pomembni korpusi

Kot je pokazal dosedanji razvoj jezikovnih virov in tehnologij, so veliki korpusi besedil ključni za strojno učenje velikih jezikovnih modelov, ki so trenutno najuspešnejši pristop k obdelavi in razumevanju

naravnega jezika. Nekateri korpusi (npr. Gigafida, Janes, GOS, itd.) so še posebej pomembni in je za njih potrebna stalna skrb. Po zgledu drugih EU držav je potrebno za te nacionalno pomembne korpuse stalno skrbeti, jih vzdrževati, dopolnjevati in izboljševati. V nadaljevanju za nacionalno pomembne korpuse predstavimo načrt razvoja. Skupno vsem naštetim korpusom je, da je potrebno zagotoviti stalen delotok zbiranja pokritih vsebin, njihovo jezikoslovno označevanje, prisotnost ustreznih metapodatkov in dostopnost pod danimi licenčnimi pogoji na repozitorju CLARIN.SI. Nacionalno pomembnost besedilnih korpusov naj določi Center za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT UL), kot temeljna inštitucija, ki se ukvarja z razvojem in vzdrževanjem korpusov v slovenskem prostoru. Na koncu razdelka predstavimo načrt razvoja metakorpusa, ki omogoča zainteresirani javnosti enoten dostop do vseh pomembnih slovenskih korpusov, in orišemo še neobstoječe, vendar potrebne in pomembne korpuse.

2.1 Referenčni korpus Gigafida

Korpus Gigafida je referenčni korpus pisne standardne slovenščine, ki vsebuje različna besedila splošne jezikovne rabe, , od dnevnih časopisov, revij, knjižnih publikacij vseh vrst (leposlovje, učbeniki, stvarna literatura) do spletnih besedil. Trenutno vsebuje skoraj 1,2 milijarde besed. Korpus zaradi vsebovanosti avtorsko zaščitenih del ne more biti javen, je pa prosto dostopen za raziskovalne namene. Korpus Gigafida je treba zaradi potreb po učenju velikih jezikovnih modelov bistveno razširiti. V ta namen je potrebno vzpostaviti:

- **Stalen delotok zbiranja pisnih izdaj**
  V sodelovanju z NUK in založniki medijskih vsebin je potrebno vzpostaviti delotoke za zbiranje standardnega jezika. Nove verzije korpusa Gigafida je potrebno pripraviti na vsaka tri leta.
- **Redno zbiranje spletnih vsebin**
  Po vzoru EU projekta MaCoCo je potrebno najmanj vsaki vsaki dve leti zajeti vse spletne vsebine v slovenščini in jih dodati v korpusa standardnega jezika Gigafida in spletnega jezika Janes.
- **Razvoj verzije korpusa Gigafida za učno rabo**
  Številni nacionalni referenčni korpusi imajo verzije prilagojene za učenje jezika (korpusi serije SkeLL), kjer je večji poudarek na pedagoškosti in pedagoški občutljivosti vsebine. Takšno učno inačico korpusa Gigafida je potrebno izdelati tudi za učenje slovenščine kot prvega in tujega jezika.

2.2 Korpus uporabniških vsebin Janes

Korpus Janes je zbirka javno objavljenih uporabniških spletnih vsebin, in sicer tvitov, forumskih sporočil, blogov, komentarjev na novice in pogovornih strani na slovenski Wikipediji. Podobno kot velja za razvoj Gigafide, je tudi za Janes potrebno zagotoviti

- **Stalen delotok zbiranja relevantnih spletnih vsebin**
  Trenutno razvite delotoke in orodja za zbiranje, čiščenje, normalizacijo in jezikoslovno označevanje je potrebno vzdrževati in razvijati. Nove verzije korpusa Janes je potrebno pripraviti na vsaka tri leta.
- **Periodične nadgradnje podkorpusa Twitter**
  Potrebno je vzdrževati in dopolnjevati orodje TweetCaT, ki kontinuirano zbira tvite slovenskih uporabnikov. S tem orodjem je potrebno sprotno zajemanje objav in letno ali dvoletno dopolnjevati korpus Janes.
- **Vključitev še nepokritih družbenih omrežij**
  Zgradi je potrebno orodje za zajem drugih, doslej nepokritih družbenih platform, ki že imajo (npr. Facebook) ali šele pridobivajo zadostno število slovenskih uporabnikov (npr. Mastodon). Javna ali le raziskovalna dostopnost novih korpusov bo odvisna od značaja in licenčni pogojev posameznih platform.

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.

2.3 Govorni korpus GOS

GOS je korpus GOvorjene Slovenščine in obsega transkripcije posnetkov (po)govora v najrazličnejših situacijah. V slovenskem prostoru zasledujemo cilj, da potrebe po govornih virih v različnih vedah sinergično usklajujemo in skušamo z istim govornim virom pokriti čim več različnih potreb, zlasti tudi za razvoj razpoznavanja govora. Govorni korpus GOS tako ne predstavlja samo osrednjega referenčnega govornega korpusa, ampak načrti za njegov nadaljnji razvoj vključujejo razmislek o potrebah po govornih virih za slovenščno v najširšem smislu. Po koncu projekta RSDO se razmisleki in raziskave na temo govornih virov nadaljujejo v okviru ARRS projekta Temeljne raziskave za razvoj govornih virov in tehnologij (MEZZANINE), ki pa je temeljni raziskovalni projekt in ne vključuje širjenja obsotječega korpusa GOS ali drugih govornih virov.

Trajnostni načrt razvoja korpusa GOS je del Delovnega sklopa 2, ki se ukvarja z govornimi tehnologijami. Ker gre za nacionalno pomemben korpus, je potrebno:

- Izdelati podroben izvedbeni načrt dolgoročnega zagotavljanja posnetkov (tudi video) in transkripcij na ekonomično vzdržen, a učinkovit način (npr. prek množičenja in spodbujanjem uporabnikov za sodelovanje z nagradami)
- Zagotavljati pravne podlage za zbiranje in distribucijo govornih gradiv, skladne z najnovejšo področno zakonodajo in drugimi predpisi, ter izkoriščanje že obstoječih arhivov posnetkov v medijih, javni upravi, zdravstvu, sodstvu ipd.
- Zagotoviti in posodabljati tehnično infrastrukturo za izdelavo govornih virov (npr. portal za oddajanje posnetkov, strojna in programska oprema za hranjenje in urejanje posnetkov in transkripcij)
- Definirati in vzpostaviti delotok, ki bo zagotavljal pregledne in varne procese dela (preglednost podatkov in varnost pred njihovo izgubo ali vnosom napak sta bistvena elementa pri razvoju govornega vira)
- Razvijati in nadgrajevati najrazličnejša orodja za označevanje in analizo govornih podatkov na vseh ravneh, od fonetične/fonemske prek oblikoslovne, skladenjske, semantične in pragmatične do označevanje neverbalnih dimenzij govorne komunikacije
- Zagotoviti vzdržen, a stabilen dolgoročni vir financiranja, ki se ne bo zaključil po nekaj letih

2.4 Pedagoški korpusi Šolar in KOST

Za namene proučevanja in izboljšav procesa učenja jezikov sta pomembna obstoječa korpusa Šolar in KOST, za namene slovničnih popravkov pa korpus Lektor. Poleg razvoja teh korpusov je potrebno izdelati tudi nove pedagoške korpuse.
Korpus šolskih pisnih izdelkov Šolar vsebuje besedila, ki so jih učenci slovenskih osnovnih in srednjih šol samostojno tvorili pri pouku. Besedila v njem predstavljajo dejansko šolsko produkcijo učencev in dijakov, kar omogoča razvoj ustreznih orodij za jezikovne popravke.
KOST je korpus slovenščine kot tujega jezika in vsebuje besedila, ki so jih napisali odrasli govorci, za katere slovenščina ni prvi jezik. Ponuja vpogled v slovenščino, kakršno tvorijo tisti, ki se jo še učijo kot drugi ali tuji jezik, predvsem v najpogostejše napake, ki nastajajo pri tem.
Korpus Lektor je zbirka lektoriranih avtorskih besedil in prevodov, ki vsebuje neliterarna, povečini strokovna in poljudnoznanstvena besedila, ki so jih lektorirali različni lektorji. Korpus vsebuje besedila v dveh različicah, in sicer izvorno avtorsko besedilo in lektorirano besedilo. Korpus omogoča uvid v najpogostejše jezikovne napake in predloge za njihove popravke.
Za nadaljnji razvoj specializiranih pedagoških korpusov je potrebno

- **Razviti nov korpus šolskih učbenikov**

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.

Posamezni primerki šolskih učbenikov so že vključeni v korpus Gigafida, vendar je zaradi specifičnosti in jezikovne pomembnosti učbeniškega besedišča potrebno vzpostaviti samostojen korpus učbenikov, zagotoviti sproten prenos vsebin od založnikov in skrbeti za razvoj tega korpusa. Skrbeti je treba tudi za razvoj korpusa učbenikov slovenščine kot drugega/tujega jezika KUUS.

- **Nadaljnji razvoj obstoječih pedagoških korpusov Šolar in KOST**
  Obstoječa korpusa sta dokaj majhna in bi bilo potrebno za izdelavo kakovostnih orodij za detekcijo in popravljanje šolskih slovničnih napak na podlagi pristopov strojnega učenja oba korpusa povečati in poskrbeti za njuno občasno, npr. triletno obnavljanje.
- **Izboljšati in povečati korpus lektoriranih besedil Lektor**
  Za razvoj in evalvacijo orodij za detekcijo in popravljanje slovničnih napak na podlagi pristopov strojnega učenja je potrebno povečati in izboljšati korpus lektor in njegovo označevanje napak harmonizirati s korpusoma Šolar in KOST.
- **Izdelati nov korpus lahkega branja**
  Za potrebe poenostavljanja besedil z orodji strojnega učenja je potrebno razviti korpus, ki bo poleg kompleksnih besedil vseboval tudi njihove poenostavljene različice. Tovrstna poenostavljena besedila, ki se imenujejo "lahko branje", služijo osebam z zmanjšanimi sposobnostmi in tistim, ki se šele učijo slovenskega jezika.
- **Specializirana orodja za napredno izrabo korpusnih podatkov**
  Razviti je treba orodja za enostavno vizualizacijo in izvažanje gradiva z označenimi jezikovnimi popravki. Splošni konkordančniki ne izkoriščajo celotnega potenciala, ki ga tovrstni bogato označeni viri ponujajo za razvoj učnih gradiv, raziskave in posredno razvoj jezikovne didaktike.
- **Izboljšati metodologijo priprave korpusov z optičnim branjem**
  Za hitrejši in celovitejši razvoj korpusov je treba dodati optično branje ročno napisanih besedil in razviti postopke, ki bi slednje prilagodili za šolsko situacijo upoštevanje značilnosti šolskega pisanja.

2.6 Parlamentarni korpus siParl

Slovenski korpus parlamentarnega govora siParl 3.0, ki je izšel v sklopu projekta RSDO, vsebuje bogato označene prepise (200 milijonov besed) parlamentarnih govorov od 1990 do 2022. Korpus siParl 2000-2022 je tudi del družine evropskih parlamentarnih korpusov, ki se s podporo CLARIN ERIC zbirajo in na enoten način pripravljajo v projektu ParlaMint, v katerem imamo slovenski raziskovalci vodilno vlogo. Parlamentarni korpusi vsebujejo vpliven segment javnega diskurza in nam skozi jezikovne vzorce pomagajo razumeti in analizirati politična stališča in trende v Sloveniji in Evropi.
V prihodnje je za vzdrževanje tega korpusa potrebno:

- **Vzdrževanje in nadgradnje delotoka za zbiranje in transkripcijo slovenskega parlamentarnega govora**
  Potrebno bo vzdrževati delotok za gradnjo novih različic korpusa in podpreti ročno delo, potrebno za izdajo novih različic. Delotok bi bilo potrebno še bolj avtomatizirati, da se zagotovi pogostejše posodabljanje korpusa, tudi skozi sodelovanje z Državnim zborom Republike Slovenije.
- **Razširitve korpusa s starejšimi podatki**
  Parlamentarni korpusi poleg raziskav sodobnega parlamentarnega diskurza predstavljajo podlago tudi za raziskave na drugih znanstvenih področjih, predvsem zgodovinopisju. Da lahko omogočimo raziskovalno dejavnost tudi na teh področjih, bi bilo potrebno korpus siParl razširiti tudi s parlamentarnimi podatki pred letom 1990.

2.7 Novi korpusi

Zgoraj opisani korpusi predstavljajo jedro dosedaj razvitih jezikovnih zbirk. V tem razdelku pa navajamo nacionalno pomembne nove korpuse, ki jih je v prihodnje potrebno načrtovati in izdelati.

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.

- **Razvoj zamejskih korpusov**
  Trenutni korpusi ne vsebuje zamejske slovenščine. V bodoče je potrebno izdelati tudi te korpuse in jih vključiti med nacionalno pomembne korpuse. To bo omogočilo izdelavo in prilagajanje orodij za jezikovno označevanje pa tudi drugih orodij, kot so slovnični popravljalniki in prepoznavalniki govora ter omogočilo jezikoslovne raziskave glede rabe in razvoja slovenskega jezika v teh okoljih.
- **Razvoj novih specializiranih korpusov**
  Za razvoj terminologije za različna strokovna področja ter za izdelavo področno specializiranih velikih jezikovnih modelov (npr. za medicino, farmakologijo, pravo itd.), je potreben razvoj novih specializiranih korpusov. Potrebno je določiti prioriteto področij glede na dostopnost virov besedil.
- **Razvoj korpusa 20. Stoletja**
  Besedila 20. stoletja, do leta 1990, ko se začelo osamosvajanje Slovenije in se je spremenil družbeni sistem, so zelo slabo pokrita v obstoječih korpusih. V sodelovanju z NUK-om jih je potrebno zbrati in digitalizirati ter iz njih izdelati besedilni korpus.

2.8 Metakorpus

Metakorpus oz. združeni korpus zbranih korpusov slovenskega jezika, narejen v okviru projekta RSDO, omogoča uporabnikom dostop do vseh nacionalno pomembnih jezikovnih korpusov na enem mestu. S tem jezikovni uporabniki hitro in enostavno dostopajo do koristnih jezikovnih informacij znotraj največjega korpusa slovenskega jezika, ki jih lahko uporabijo za razumevanje jezika, prevajanje, razvijanje aplikacij za obdelavo jezika, šolsko delo itd.
Za nadaljnji razvoj metakorpusa je potrebno naslednje.

- **Vključitev novonastalih korpusov v metakorpus**
  Vsi slovensko nacionalno pomembni in specializirani korpusi, bodisi nove različice obstoječih, oziroma povsem novi, naj se sproti vključujejo v metakorpus.
- **Posodabljanje postopka gradnje metakorpusa**
  Programsko opremo za izdelavo metakorpusa je potrebno vzdrževati in dopolnjevati. Zagotoviti mora učinkovito deduplikacijo.

2.9 Strategija razvoja in dostopnosti korpusov

Za javno financirane splošne in specializirane korpuse je potrebno zagotoviti njihovo čim večjo javno dostopnost in tudi čim večjo dostopnost za raziskovalne namene. V ta namen je potrebno izdelati strateški dokument ravnanja s temi korpusi, ki bo opredelili smernice za njihovo javno dostopnost, in določitev o njihovi nacionalni pomembnosti.

3. Zaključki in povezave z drugimi sklopi

Jezikovni korpusi predstavljajo temeljni informacijski vir za razvoj jezikovnih tehnologij, ki temeljijo na tehnologijah strojnega učenja in umetne inteligence, katere eno glavnih področij je prav obdelava naravnega jezika. V dokumentu smo orisali načrt razvoja temeljnih nacionalno pomembnih slovenskih korpusov in orodij za njihovo osnovno jezikoslovno označevanje za naslednjih nekaj let. Skupna točka je zagotovitev stabilnega financiranja za vzdrževanje in nadaljnji razvoj korpusov, učnih množic, modelov strojnega učenja in programske opreme za jezikoslovno označevanje.

Korpusi in temeljno jezikoslovno označevanje predstavljajo podstat za delovne sklope 2, 3, 4 in 5 projekta RSDO.Potrebni so za razvoj jezikovnih modelov za prepoznavanje in generiranje govora, kar

opisujemo v Delovnem sklopu 2. Korpusi omogočajo razvoj semantičnih jezikovnih virov in tehnologij, katerih razvoj načrtujemo v Delovnem sklopu 3. Gre predvsem za tehnologije za globlje razumevanje jezika, ki brez korpusne osnove in na njih temelječih velikih jezikovnih modelov ne delujejo, npr. povzemanje, odgovarjanje na vprašanja in generiranje jezika. Korpusi in jezikoslovno označevanje se preko velikih medjezikovnih in mnogojezikovnih modelov povezujejo s strojnim prevajanjem, katerega razvoj opišemo v Delovnem sklopu 4. Iz korpusov se črpa tudi terminologija, ki jo obravnavamo v Delovnem sklopu 5.

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.