ODPRTOST JEZIKOVNIH VIROV ZA SLOVENŠČINO

Tomaž Erjavec
Institut "Jožef Stefan", Ljubljana

UDK 811.163.6'322:004.738.52

Prispevek zagovarja odprtost jezikovnih virov kot enega od pogojev za napredek v poučevanju, proučevanju in razvoju jezikovnih tehnologij in jezikoslovja slovenskega jezika. V prispevku utemeljimo prednosti in izpostavimo probleme odprtih jezikovnih virov. Predstavimo vidike odprtosti, in sicer koncept predkonkurenčnosti, problem avtorskih pravic, standarde zapisa in najdljivosti, ter podamo primer dobre prakse. V zaključku podamo nekaj predlogov, kako povečati odprtost jezikovnih virov za slovenski jezik.

odprti jezikovni viri, računalniški korpusi, avtorske pravice, standardi zapisa

The paper discusses the importance of open language resources as a basis for progress in teaching, researching and developing language technologies and linguistic studies of Slovene. The pros and cons of open language resources are discussed. Issues such as pre-competitiveness, copyright, recording standards and ease of finding language resources are considered and an example of good practice is given. The paper concludes with some suggestions for increasing the openness of language resources for Slovene.

open language resources, computer corpora, copyright, encording standards

1 Uvod

Zadnji dve desetletji lahko, tako na področju jezikovnih tehnologij kot na področju jezikoslovja, poimenujemo kar obdobje jezikovnih virov. V jezikovnih tehnologijah se je v tem obdobju izoblikovalo in utrdilo stališče, da šele jezikovni viri omogočajo razvoj kakovostnih, robustnih in praktično uporabnih orodij za obdelavo naravnega jezika. Po eni strani je temu premiku botrovala uvedba pojma evalvacije, kjer se kakovost posameznih orodij ovrednoti na testni podatkovni množici, torej na nekem kakovostno označenem jezikovnem viru, po drugi strani pa izjemen napredek v strojnem učenju, kjer se računalniške modele jezika nauči neposredno iz jezikovnih virov. Tudi v jezikoslovju se je težišče raziskav premaknilo v korpusno (podprto) jezikoslovje, ki, za razliko od jezikoslovja, temelječega na intuiciji, potrebuje za svoje raziskave velike, po možnosti kakovostno izdelane in označene jezikovne vire.

Pojem »jezikovni viri« v prispevku opredelimo kot digitalno zapisane eno- ali večjezične jezikovne in jezikoslovne podatke, med katere sodijo:

– Korpusi pisnega ali govorjenega jezika, torej zbirke besedil, večinoma opremljene z metapodatki in jezikoslovno označene. Kot podzvrst lahko štejemo pomnilnike prevodov, do neke mere pa tudi digitalne knjižnice.

– Slovarji in računalniški leksikoni, pri čemer so prvi namenjeni branju in imajo tako veliko informacij predstavljenih kontekstualno, slednji pa računalniški uporabi in so torej precej bolj formalizirani. V prvo kategorijo sodijo tudi terminološki slovarji, v slednjo pa terminološke baze, semantični leksikoni in do neke mere tudi ontologije.

– Računalniški modeli jezika, ki predstavljajo osnovo za delovanje jezikovno-tehnoloških orodij, ki npr. identificirajo leme, oblikoslovne oznake, skladenjske strukture, termine, pomenske oznake, prevode ...

– Specifikacije za formate in nabore oznak, kot tudi bibliografske podatke, torej metapodatke o virih; sem štejemo npr. definicije oblikoslovnih oznak, ki se uporabljajo pri označevanjih korpusov, ali opis strukture nekega tipa virov.

Tako za korpusno jezikoslovje kot za jezikovne tehnologije velja, da jezikovne vire, ki jih raziskovalci potrebujemo za proučevanje, ali aplikacije lahko mogoče izdelamo sami, vendar to vodi k velikemu podvajanju dela in obremenjenosti raziskovalcev z zagotavljanjem potrebne podatkovne infrastrukture, ki, namesto da bi se posvetili predmetu raziskave, izgubijo veliko časa s pridobivanjem in pripravo podatkov.

Če za »velike« jezike vsaj do neke mere velja, da za zagotavljanje jezikovnih virov lahko uporabimo tržne vzvode, je situacija drugačna za jezike, kot je slovenščina, saj je malo virov, kjer ne bo razkorak med sredstvi, potrebnimi za njihovo izdelavo, in potencialnim trgom zanje prevelik, da bi se stroški razvoja lahko povrnili. V vsakem primeru je koncept prodaje virov primeren predvsem v poslovnem kontekstu. S strani univerz in drugih akademskih institucij pa je plačljivost nekega vira velika ovira za njegovo uporabo. V tem prispevku nas bo zanimal predvsem model »raziskovalci raziskovalcem«, kjer je jezikovni vir izdelala raziskovalna institucija, uporabnik tega vira pa je ravno tako (druga) raziskovalna institucija, pri čemer kot raziskovalne institucije vzamemo univerze, raziskovalne inštitute in tudi podjetja, ki izvajajo raziskovalno dejavnost. Namembnost takih jezikovnih virov je po eni strani poučevanje slovenskega jezika, prevodoslovja in jezikovnih tehnologij, po drugi strani pa raziskovanje na teh področjih, torej korpusno jezikoslovje in razvoj jezikovnih tehnologij.

Teza prispevka je, da bi morali jezikovni viri slovenskega jezika, ki nastanejo s pomočjo javnih sredstev, biti javni oz. odprti, saj se s tem najbolj pripomore k informatizaciji in tudi kvalitetnemu študiju slovenskega jezika. V Sloveniji je za odprtost jezikovnih virov do zdaj še največ naredilo Slovensko društvo za jezikovne tehnologije (SDJT),[1] predvsem z organizacijo konferenc o jezikovnih tehnologijah (1998–2008, vsaki dve leti, zborniki so dostopni na spletu), v zadnjem času pa tudi z organizacijo serije sestankov, posvečenih jezikovim virom.[2]

2 Dimenzije odprtosti

Kako naj razumemo »odprtost« jezikovnih virov? Odgovor ni enoznačen, temveč večdimenzionalen, kar obravnavamo v tem razdelku. Zgodovinsko gledano je pojem prišel iz gibanja za odprto kodo (OSI), ki se zavzema za dostopnost izvorne kode programske opreme, saj je na ta način omogočena ne samo splošna uporaba programov, temveč tudi javno popravljanje napak, prilagajanje in nadgrajevanje. Odprta koda ima natančne pravne formulacije in obstaja v več različicah. Tu se ne bomo spuščali v podrobnosti definicij licenc za odprte kode, ker so predvsem usmerjene v programsko opremo in ne v podatke (jezikovne vire), vseeno pa omenimo po številu uporabnikov enega uspešnejših odprtokodnih virov za slovenski jezik, in sicer slovar besednih oblik za odprtokodni program Aspell, (GNUsl), ki je nastal v sodelovanju Društva

1 Kjer je v prispevku kratica podana v oklepajih, je spletni vir zanjo na koncu prispevka v seznamu spletnih virov.

2 Prispevek nima namena podajati pregleda jezikovnih virov za slovenščino, čeprav bi bilo to zanimivo s stališča vpeljane tipologije. Sezname in opise pomembnejših virov najdemo na spletnih straneh SDJT, predvsem na http://www.sdjt.si/sdjt-www.html in http://www.sdjt.si/dogodki/LJ2008/.

uporabnikov Linuxa LUGOS, podjetja Amebis, d. o. o. in Inštituta Jožef Stefan v okviru projekta takratnega Ministrstva za informacijsko družbo. Aspell je s slovarji za skoraj sto jezikov dostopen pod GNU General Public Licence in se uporablja za slovenski črkovalnik v OpenOffice in Mac/iWork, ki skupaj pokrivata bolj ali manj vse ne-Windows uporabnike.

V prispevku odprtost jemljemo precej širše kot diskreten prostor v več dimenzijah, in sicer dejansko, pravno, tehnološko in najdljivo. Jezikovni vir bo v raziskovalnih okvirih minimalno odprt že s tem, da lahko uporabniki iz nematične institucije do vira sploh dostopajo, ne glede na način dostopa, restriktivnost pogojev uporabe, formata virov ali njihove kakovosti. Po drugi strani bo maksimalno odprt vir z minimalno truda omogočal komurkoli izvedeti zanj, za njegovo vsebino in strukturo, ga v izvornem in standardiziranem formatu prenesti na svoj računalnik ter ga zastonj uporabiti v poljubne namene.

Uporabo jezikovnih virov lahko razumemo na dva zelo različna načina. Po eni strani je lahko na spletu dostopen skozi program, ki nam omogoča iskanje in izpis delov vira. Tipičen primer bi bili spletni konkordančniki, ki omogočajo dostop do korpusa, kot je npr. FidaPLUS (Arhar, Gorjanc, 2007). Ta vrsta uporabe je jezikoslovno usmerjena, neprimerna pa je za jezikovnotehnološke raziskave, saj zanje potrebujemo celotno podatkovno zbirko, torej možnost prenosa vira. Takšna uporaba je tudi tema prispevka, ki pa seveda ni omejena samo na jezikovne tehnologije – tudi jezikoslovci, posebno tisti, ki imajo vsaj osnovna računalniška znanja, lahko izkoristijo možnost uporabe celotnega vira.

2.1 Predkonkurenčnost in znanstvena etika

Prva ovira pri dostopnosti jezikovnih virov je odločitev avtorjev, da zainteresirani javnosti ne omogočijo dostopa do vira. Matična institucija ima lahko za zaklepanje izdelanih virov različne razloge. Predvsem korpusi, lahko pa tudi leksikoni (slovarji), imajo tipično pripisane predhodne avtorske pravice na izvorno besedilo, ki lahko preprečujejo naknadno razširjanje. Vendar se ta ovira velikokrat zdi večja, kot je v resnici (več o tem v razdelku 2.2), obstaja pa tudi precej virov, katerih odprtost je odvisna predvsem od njihovih avtorjev. Tako je npr. pri nas vse več doktorskih disertacij, v sklopu katerih so izdelani zanimivi specializirani korpusi slovenskega jezika, ki nikoli niso objavljeni. Tu je verjetno glavna ovira nevednost študentov, pa tudi mentorjev, da je to lahko koristno in da je v odprtost vredno vložiti nekaj dodatnega truda pri izdelavi vira.

V slovenskem prostoru, pa tudi širše, so viri pogosto zaklenjeni, da bi matična institucija s tem pridobila ali ohranila konkurenčno prednost. Ta pristop je povsem legitimen v primeru podjetij, ki sama financirajo svojo raziskovalno oz. razvojno dejavnost, veliko manj tam, kjer raziskave delno financirajo javna sredstva, povsem neprimeren pa, kadar viri nastanejo v okviru državnih institucij in v celoti z javnimi sredstvi.[3] Rezultati takšnega dela bi morali biti usmerjeni v čim širšo uporabo, konkurenčna prednost pa bi se morala izraziti skozi publikacije; po eni strani takšne, kjer avtorji vir in postopek izdelave opišejo, po drugi pa skozi citate v publikacijah, ki opisujejo raziskave, ki te vire uporabljajo. Citiranje je še posebej pomembno, ker je merljiv

3 Najbolj znan primer zaklenjenega jezikovnega vira (v smislu dostopa kot podatkovne baze) je Slovar slovenskega knjižnega jezika, katerega izdelava je bila v celoti financirana iz javnih sredstev in ki je bil nato s pomočjo nadaljnjih javnih sredstev spremenjen v podatkovno bazo, vendar je skrbno varovan v okviru Inštituta za slovenski jezik Frana Ramovša. Napredek jezikovnih tehnologij za slovenski jezik je bil s tem bistveno upočasnjen, saj so bile ostale raziskovalne institucije prisiljene – večinoma spet z javnimi sredstvi – računalniške leksikone slovenskega jezika razvijati povsem na novo.

kazalec raziskovalne uspešnosti, zato bi se tudi moralo dosledno izvajati. Žal pa to ni v navadi pri citiranju publikacij o jezikovnih virih: vse prepogosto se nek vir omeni samo po imenu ali pa se v najboljšem primeru doda njegov spletni naslov, namesto da bi se v virih citiralo publikacijo, kjer je vir prvotno opisan. Za razvoj slovenskega jezika bi bilo tako koristno, da so vsaj jezikovni viri, ki nastanejo z javnimi sredstvi, potem tudi zares javni, bolje kot da utrjujejo monopole na področju slovenskega jezika. Po drugi strani pa bi bilo nujno, da raziskovalci v svojih publikacijah skozi bibliografske citate korektno navedejo, katere vire so uporabili.

2.2 Avtorske pravice

Povsem realen in zelo kompleksen problem pri nadaljnjem razširjanju jezikovnih virov so avtorske pravice, vezane na njihov izvor, saj je malo virov, ki nastanejo povsem na novo. Posebej je ta problem pereč pri korpusih sodobnega jezika; ti vsebujejo izvorna besedila, ki so po definiciji avtorsko zaščitena, pojavljajo pa se tudi pri drugih virih, če je npr. računalniški leksikon narejen na osnovi slovarja. Po drugi strani imajo tudi neposredni izdelovalci vira avtorske pravice na dodano vrednost, npr. na jezikoslovne oznake, posebej če so bile ročno zapisane. V ekstremih obstajata dva načina spopadanja s tem problemov: vir enostavno ni javno dostopen in sta uporaba ter razširjanje urejena neformalno ali pa je z vsakim nosilcem avtorskih pravic podpisana pogodba, vsaj digitalno pa podpisana tudi pogodba z vsakim uporabnikom. V prvem primeru je krog uporabnikov seveda majhen, slednji model pa zahteva veliko investicijo časa, saj je do podpisov težko priti. Pravno korektni pristop do besedilodajalcev je verjetno potreben pri referenčnih, posebej nacionalnih korpusih, in seveda nujen pri virih, ki vsebujejo osebne podatke. Za ostale vire, ki bi jih hoteli odpreti, predvsem korpuse, pa bi zagovarjali bolj sproščen odnos. Posebej za nekomercialno uporabo je npr. rešitev, da se besedila pridobijo brez pogodb (s spleta, neposredno od avtorjev oz. založb z ustnim pristankom) in se jih vključi v korpus, na kasnejšo željo avtorjev pa se njihova besedila iz korpusa izbrišejo.

Druga plat avtorskih pravic je pogodba, pod katero neka institucija ali oseba dobi odprti vir v uporabo, predelavo ali nadaljnje razširjanje. Če uporabnikove pravice vsebujejo slednje, je tak vir res odprt v smislu odprte kode, ki je dobila svojo ustreznico za besedila v licencah Creative Commons (CC), kar je verjetno največ, kar lahko dosežemo v smislu odprtosti z dobro pravno podlago. Obstaja več vrst licence CC, ki nalagajo različne omejitve pri uporabi oz. razširjanju: da je potrebno navesti avtorja dela, da je uporaba dovoljena samo za nekomercialne namene in da je delo dovoljeno razširjati samo brez predelav. Kjer je mogoče, bi zagovarjali uporabo licence CC, saj tako vir lahko doseže največjo možno odmevnost.

2.3 Standardi zapisa

V razdelku opišemo tehnični, pa tudi jezikoslovnoformalni vidik odprtosti jezikovnih virov. Ta temelji na uporabi standardov, mednarodnih priporočil ali dobrih praks, kar naj bi izboljšalo kvaliteto, dostopnost in trajnost virov. Pri tem se je treba zavedati, da je »krasna stvar pri standardih, da jih je toliko« in da dostikrat ne obstaja enoumna rešitev, katerega, če sploh katerega, uporabiti. Posebej pri kompleksnejših načinih označevanja lahko obstaja več neskladnih priporočil, pa drugi strani pa standardi po definiciji omejujejo in lahko silijo v »nenaravne« rešitve za opis lastnosti posameznega jezika.

Osnovni standard za zapis jezikovnih virov je priporočilo konzorcija za svetovni splet W3C, in sicer Extensible Markup Language (XML), ki je poleg tega, da je splošno uveljavljen, tudi razmeroma enostaven za uporabo. Ker je dokumente XML mogoče programsko validirati za strukturo in tudi za pravilen zapis

znakov, uporaba tega standarda zagotavlja pravilnost zapisa na formalni ravni, pomaga pa tudi pri uporabi vira, saj je XML mogoče razmeroma enostavno pretvoriti v poljubne ciljne formate, ki jih nato uporabljajo konkretna orodja.

Pristop, kjer za osnovni zapis jezikovnega vira uporabimo XML, je zelo drugačen od tistega, ki vir vidi samo v kontekstu uporabe nekega konkretnega orodja. Zaradi formata konkretnega orodja se namreč hitro zgodi, da so pomembni (meta)podatki o viru nezabeleženi, zapis znakov pa na razne načine pokvarjen, s čimer kvaliteta viru občutno pade.

Na nivojih zapisa, višjih od XML, konsenz o uporabi standardov oz. priporočil hitro pada, vseeno pa je nekaj izjem. Za »jezikovne vire«, kot so digitalne zbirke historičnih besedil, ki vsebujejo kompleksna razmerja med faksimili in prepisi, se skoraj povsod uporabljajo priporočila konzorcija za zapis besedil (TEI), to pa vsaj do neke mere velja tudi za slovarje in označene korpuse. Na področju pomnilnikov prevodov in terminoloških baz sta se uveljavila TMX in TBX, oba kot priporočili Localization Industry Standards Organization (LISA), na področju računalniških leksikonov pa je bil nedavno sprejet standard ISO-24613:2008, Lexical Markup Framework (LMF), ki ga že uporablja več EU projektov, ni pa še jasno, kako široko se bo res uveljavil.

Vsa ta priporočila se nanašajo predvsem na zapis strukture, ne pa na jezikoslovno označevanje in klasifikacijo. Tu do neke mere vlada konsenz samo za področje oblikoslovnih oznak, kjer se je za več jezikov, v veliki meri tudi za slovenščino, uveljavil sistem MULTEXT oz. MULTEXT-East, ki vsebuje priporočila za veliko število, predvsem evropskih jezikov. Po sistemu MULTEXT-East so označeni korpusi Fida in FidaPLUS, kot tudi drugi, kjer je označevanje izvedlo podjetje Amebis, d. o. o ali Institut Jožef Stefan. Primer prvega je KoRP, Korpus besedil odnosov z javnostmi (Logar, 2007), drugega pa iKorpus (Vintar, Erjavec 2008). Iz tega sistema označevanja korpusov slovenskega jezika najbolj izstopa nabor oblikoslovnih oznak, razvit na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU (Jakopin, Bizjak 1997), vendar pa so tudi tu v teku raziskave, ki naj bi omogočile prehajanje med obema sistemoma (Jakopin, Bizjak, 2008).

2.4 Najdljivost

Tudi če je nek vir odprt, kako naj potencialni uporabniki zanj zvedo? Slovenski prostor je v tem pogledu še dovolj majhen, da poznanstva, dopisni seznami[4] in objave do neke mere zadostujejo, čeprav seveda ni nujno, da so potencialni uporabniki iz Slovenije, da razumejo slovenski jezik; v prihodnosti pa bo isto lahko veljalo tudi za proizvajalce virov slovenskega jezika. Zato je vsaj za mednarodno odmevnost pomembno, da so viri slovenskega jezika dokumentirani ne samo v slovenskem, temveč tudi v angleškem jeziku.

Nekateri avtorji distribucijo virov prepustijo eni od agencij, ki se ukvarjajo s tem, kar sta zaenkrat Linguistic Data Consortium (LDC) v ZDA in Evaluations and Language resources Distribution Agency (ELDA) v Evropi.[5] Obstaja tudi več spletnih strani, ki naj bi zaobjele vire za posamezne jezike – za slovenskih jezik sta bili dve taki že omenjeni – njihov problem pa je, da so dostikrat pomanjkljive in hitro zastarajo. V zadnjem času se za iskanje jezikovih virov razvijajo tudi posebni standardi in protokoli, ki definirajo zajem, zapis in uporabo metapodatkov o virih, kot je npr.

4 Med slovenskimi dopisnimi seznami področje jezikovih virov pokrivata zelo aktiven seznam za slovensko literarno vedo slovlit@ijs.si in nekoliko manj dejaven seznam SDJT, sdjt-l@ijs.si, mednarodno pa corpora@uib.no.

5 Za slovenske vire to možnost za zdaj izkorišča samo Laboratorij za digitalno procesiranje signalov Fakultete za elektrotehniko, računalništvo in informatiko Univerze v Mariboru, ki skozi agencijo ELDA prodaja pet virov za slovenski jezik.

Open Archive Initiative (OAI). Katalog in mreža ponudnikov virov, ki naj bi pokrivala jezikovne vire za področje humanistike (torej bolj kot za razvoj jezikovnih tehnologij), sta v delu v projektu Common Language Resources and Technology Infrastructure (CLARIN), ki se tudi sicer trudi vzpostaviti modele dobre prakse na področju zapisa in javne uporabe jezikovnih virov.

3 Odprti viri za oblikoslovno označevanje slovenščine

V tem razdelku predstavljamo jezikovni vir, ki se trudi čim bolj upoštevati zgoraj navedene kriterije, in sicer oblikoslovno označena korpusa slovenskega jezika jos100k in jos1M, ki sta nastala v okviru projekta Jezikoslovno označevanje slovenščine (JOS). Oblikoslovno označevanje in z njim povezana lematizacija sta osnovni in izredno pomembni predobdelavi besedila, pri čemer se sodobni označevalniki naučijo modela nekega jezika neposredno iz pravilno označenega korpusa. Za oblikoslovno označevanje slovenskega jezika potrebujemo čim večji in čim bolj raznovrsten korpus, ki je kakovostno (ročno) označen. Obstaja odprti označeni korpus MULTEXT-East (Dimitrova idr., 1998), ki ima tudi slovenski podkorpus, vendar ta vsebuje eno samo besedilo, roman *1984* Georga Orwella, zaradi česar se iz njega ni bilo moč naučiti kvalitetnih označevalnih modelov.[6]

Projekt JOS zapolnjuje to vrzel s korpusoma jos100k in jos1M. Prvi ima 100.000 besed in je ročno oblikoslovno označen in lematiziran, drugi pa šteje milijon besed z delno ročno preverjenimi oblikoslovnimi oznakami in lemami. Korpusa vsebujeta naključno izbrane odstavke iz korpusa FidaPLUS. Korpusa sta dostopna preko konkordančnika in tudi neposredno za prenos pod licenco Creative Commons BY NC, kar pomeni, da je treba pri uporabi navesti avtorje vira in da se lahko uporablja za nekomercialne namene.

Korpusa sta namenjena za raziskave, publikacije, vezane na uporabo korpusov, pa so pogojene s korektnim citiranjem vira. Avtorske pravice na predhodnem označevanju iz korpusa FidaPLUS so bile urejene s pogodbo med nosilcem projekta, IJS, in projektnimi partnerji FidaPLUS. Avtorske pravice nad izvornimi besedili so urejene preko besedilodajalske pogodbe s FIDO oz. FidoPLUS tako, da korpusi JOS za vsak odstavek navedejo vir besedila, ter z dejstvom, da vsebujejo samo posamezne odstavke besedil – da iz njih torej ni mogoče v celoti zajeti izvornih besedil. S strani besedilodajalcev je pomirjujoče tudi dejstvo, da IJS nudi korpusa brezplačno. Korpusa sta zapisana na standardiziran način v XML po smernicah TEI P5. Specifikacije za oblikoslovne oznake so tudi javno dostopne in ravno tako zapisane v XML/TEI P5, obenem pa dosegljive tako v slovenskem kot angleškem jeziku. Dodatno je na splet postavljen oblikoskladenjski označevalnik za slovenski jezik, naučen na korpusih JOS, s katerim je moč preko spleta označiti poljubno slovensko besedilo in shraniti rezultate, ki so nato primerni za uvoz v konkordančnik ali drugo orodje.

4 Zaključek

V prispevku smo poskusili utemeljiti potrebo po odprtih jezikovnih virih, obravnavali več vidikov odprtosti in nakazali rešitve, ki bi pripomogle k večji odprtosti slovenskih jezikovnih virov. V prispevku so bile omenjene tudi prepreke pri odpiranju virov, ki bi jih na kratko lahko razdelili v psihološke, pravne in tehnične; poleg teh pa je prepreka seveda tudi to, da je za zagotavljanje odprtosti nekega vira potrebno vložiti dodaten trud in ga izpostaviti kritični javnosti. Vseeno pa bi težnja po odprtosti morala biti nujna lastnost vseh virov, ki

6 Za slovenski jezik je bil sicer razvit večji oblikoslovno označeni korpus na ISJ ZRC SAZU (Jakopin, Bizjak, 1997), vendar pa je žal, tako kot SSKJ, zaklenjen.

nastajajo kot plod raziskav, financiranih z javnimi sredstvi. V zaključku podajmo še nekaj konkretnih predlogov, kako spodbuditi odprtost jezikovnih virov za slovenščino:

– Uvajanje pogoja odprte uporabe (ali vsaj dodatnih točk pri evalvaciji) pri razpisih ARRS in drugih javnih razpisih za predloge projektov, kjer je namen izdelovanje jezikovnih virov, kot je že zdaj v navadi pri razpisih za raziskovalno-razvojne projekte EU.

– Upoštevanje (ne)korektnega citiranja uporabe jezikovnih virov pri recenziranju člankov s strani programskih odborov.

– Vzpostavitev registra (odprtih) jezikovnih virov za slovenski jezik v okviru iniciative CLARIN.

– Pedagoške vsebine na univerzah, kjer se študentje naučijo pisanja člankov (reference), osnove standardov računalniškega zapisa besedil (Unicode, XML) in znanj, ki jim omogočajo samostojno ne samo uporabljati, temveč tudi najti ali narediti odprte jezikovne vire.

Literatura in spletni viri

ARHAR, Špela, GORJANC, Vojko, 2007: Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2. 95–110.

DIMITROVA, Ludmila, ERJAVEC, Tomaž, IDE, Nancy, KAALEP, Heiki-Jan, PETKEVIČ, Vladimır, TUFIŞ, Dan, 1998: Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. Zbornik *COLING-ACL* '98. 315–319.

ERJAVEC, Tomaž, KREK, Simon, 2008: Oblikoskladenjske specifikacije in označeni korpusi JOS. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Zbornik Šeste konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 49–53.

JAKOPIN, Primož, BIZJAK, Aleksandra, 2008: Part-of-speech tagging of Slovenian, 12 years after. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Zbornik Šeste konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 104–109.

JAKOPIN, Primož, BIZJAK, Aleksandra, 1997: O oblikoslovnem označevanju slovenskega besedila. *Slavistična revija* 45/3–4, [513]–532.

LOGAR, Nataša, 2007: Spričevalo stroke: njen slovar. Postružnik, Natalija, Kušar, Matej (ur.): *Zbornik 11. slovenske konference o odnosih z javnostmi*. Ljubljana: Slovensko društvo za odnose z javnostmi. 38–45.

LOGAR, Nataša, 2008: *Pregled korpusov za slovenščino*. Predavanje na posvetu »Est modus in korpus: ni korpusov brez divizij« SDJT. Ljubljana, IJS, oktober 2008. http://www.sdjt.si/dogodki/LJ2008/sdjtLJ08.htm

TEI Consortium, 2007: TEI P5: *Guidelines for Electronic Text Encoding and Interchange*.

VINTAR, Špela, ERJAVEC, Tomaž, 2009: iKorpus in luščenje izrazja za Islovar. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.). *Zbornik Šeste konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, 2008, 65–69.

CC – Creative Commons: http://creativecommons. org/

CLARIN – Common Language Resources and Technology Infrastructure, http://www.clarin.eu/

ELDA – Evaluations and Language resources Distribution Agency, http://www.elda.org/

FidaPLUS – http://www.fidaplus.net/

GNUsl – http://nl.ijs.si/gnusl/

iKorpus – Korpus informacijskega izrazja, http://nl2. ijs.si/dsi.html

JOS – Jezikoslovno označevanje slovenščine, http://nl. ijs.si/jos/

KoRP – Korpus besedil odnosov z javnostmi, http:// www.korp.fdv.uni-lj.si/

LDC – Linguistic Data Consortium, http://www.ldc. upenn.edu/

LISA – Localisation Industry Standards Organisation, http://www.lisa.org/

LMF – Lexical Markup Framework, http://www.lexicalmarkupframework.org/

MULTEXT- East, http://nl.ijs.si/ME/

Nova beseda – http://bos.zrc-sazu.si/

OAI – Open Archives Initiative, http://www.openarchives.org/

OSI – Open Source Initiative, http://www.opensource .org/

SDJT – Slovensko društvo za jezikovne tehnologije, http://www.sdjt.si/

SSJ – Sporazumevanje v slovenskem jeziku, http:// www.slovenscina.eu/

TEI Consortium – http://www.tei-c.org/

XML – Extensible Markup Language, http://www. w3.org/XML/

ZRC SAZU – Oblikoslovno označevanje na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU http://bos.zrc-sazu.si/oblikoslovno_oznacevanje.html