

Tomaž Erjavec

Odsek za tehnologije znanja, Institut »Jožef Stefan«

Odprt dostop do podatkovne baze slovarja

Prispevek izpostavi pomen odprtega dostopa do načrtovanega slovarja slovenskega jezika, predvsem za namene jezikovnih tehnologij. V nasprotju s prostim dostopom, pri katerem sta uporabnikom omogočeni prebiranje in iskanje po slovarju na spletu, pomeni odprt dostop možnost prevzema celotne slovarske baze v izvornem zapisu, podobno kot pri bolj znani iniciativi za odprto programsko kodo. V prispevku argumentiramo potrebo po čim bolj odprtem dostopu, izpostavimo primere dobre in slabe prakse v Sloveniji in pregledamo različice licenc Creative Commons in kot najbolj ustrezno za pravni okvir razširjanja slovarja predlagamo licenco Creative Commons – Priznanje avtorstva (CC BY).

Open access to the lexical database of the dictionary

The paper discusses the importance of open access to the planned new dictionary of Slovene language, esp. for the purposes of language technologies. As opposed to free access, where the users can search and read the dictionary on-line, open access means that the complete lexical database can be downloaded in its source encoding, as is the case with programs developed under the better-known open source initiative. The paper motivates the need for open access, presents some cases of best and worst practices in Slovenia, overviews the Creative Commons licences and proposes the use of CC BY (Attribution) as the most appropriate one for the legal framework of the new dictionary.

Ključne besede: odprti dostop, Creative Commons, avtorske pravice

Keywords: open access, Creative Commons, copyright

1 Uvod

Prispevek povzema načela odprtosti jezikovnih virov slovenskega jezika, ki smo jih že izpostavili v Erjavec (2009), vendar so tu načela osredotočena na problematiko dostopnosti načrtovanega slovarja slovenskega jezika. V prispevku ne bo govora o *prostemu dostopu* do

slovarja, pri katerem je mogoče brezplačno in brez registracije videti geselske članke na spletu, saj takšen dostop predpostavljamo. Zato pa je manj jasno, do kakšne mere bo dostop do novega slovarja odprt. Pod pojmom *odprti dostop* razumemo možnost in dovoljenje, da lahko celoten slovar prevzamemo (prenesemo) na svoj računalnik, in to ne samo v kateri od oblik, predvidenih za predstavitev, kot sta HTML ali PDF, temveč kot kopijo slovarske baze v izvornem zapisu XML.

Razliko ilustrira primer ročno oblikoskladenjsko označenih korpusov slovenščine, katerih izdelava je po eni strani draga in zamudna, po drugi pa so nujni predpogoj za izdelavo označevalnikov in lematizatorjev za posamezni jezik, brez katerih je kakršnakoli bolj kompleksna obdelava jezika nemogoča. Takšen korpus, katerega izdelavo ocenjujemo na vsaj dve človek leti, je bil namreč izdelan že v 90-ih letih prejšnjega stoletja (Jakopin in Bizjak, 1997) in nato prosto dostopen v spletnem konkordančniku, vendar pa nikoli ni bil kot podatkovna baza dan na voljo drugim raziskovalcem. Šele projekt ARRS »Jezikoslovno označevanje slovenskega jezika« je izdelal primerljiv korpus *jos1M* (Erjavec in Krek, 2008) in ga naredil odprto dostopnega. Razvoj jezikovnih tehnologij za slovenski jezik je tako izgubil deset let, obenem pa je bilo zelo podobno delo z javnimi sredstvi financirano dvakrat.

Šele z možnostjo nadaljnjega kopiranja lahko novi slovar polno izpolni svoje poslanstvo, da ga bo na čim več možnih načinov uporabljalo čim večje število ljudi. Za razliko od materialnih dobrin, kot so npr. knjige, je kopiranje računalniških podatkov namreč zastonj, pri tem pa odprt dostop do slovarske baze nudi, odvisno od licence, vrsto izvedenih aplikacij s strani tretjih oseb, kot so sintetične jezikoslovne raziskave, prilagajanje slovarskih gesel za ciljne skupine uporabnikov, ponovne izdaje na novih platformah, prepis gesel v Wikislovar in druge odprte repozitorije in, ne nazadnje, predelavo slovarskih ravní v leksikone¹ za uporabo v jezikovnih tehnologijah slovenskega jezika. Z vsem tem spodbujamo informatizacijo slovenščine kot enega nujnih pogojev za preživetje in razvijanje jezika v digitalni dobi.

Osnovna teza prispevka je, da morajo vsi jezikovni viri slovenskega jezika, ki nastanejo z javnim financiranjem, tudi biti javni in s tem maksimalno odprti, kolikor to pač dopuščajo predhodne avtorske pravice in varovanje pravice do zasebnosti, kar pa pri predvidenem slovarju ne bi smel biti problem. Zapiranje namreč koristi zgolj inštituciji oz. inštitucijam, ki

¹ Termin »leksikon« tu uporabljamo za slovarje oz. besedišča, ki so namenjeni za jezikovnotehnološke uporabe: imajo torej formalno dobro definirano strukturo, ki pa je enostavnejša od slovarske, saj tipično pokriva samo eno raven jezikoslovnega opisa.

si nad slovarjem lastijo avtorske pravice, uporabnikom, ki so slovar posredno financirali, pa samo škoduje. Ne nazadnje predvideva odprt dostop do rezultatov raziskav tudi Resolucija o nacionalnem programu za jezikovno politiko 2014–2018, pa tudi sicer se slovenska politika financiranja raziskav počasi, a vztrajno bliža zahtevi po obveznem odprtem dostopu rezultatov javno financiranih raziskav (Kotar, 2013), ki je npr. tudi obvezen pogoj za vse projekte Obzorja 2020.

V nadaljevanju prispevka najprej izpostavimo dosedanje dobre in slabe prakse v Sloveniji, nato podrobneje obdelamo variante licenc Creative Commons, ki so zaradi svoje uveljavljenosti verjetno najboljši kandidat za formalnopravno urejanje dostopa do slovarja, ter na koncu podamo nekaj zaključkov.

2 Slovenske prakse pri zagotavljanju odprtosti slovarskih jezikovnih virov

Pri obravnavi načrtov je koristno pogledati v preteklost in preučiti, kako je z odprtostjo že narejenih jezikovnih virov slovenskega jezika, saj nam to izpostavi uveljavljene prakse v Sloveniji, ki bi se lahko nadaljevale tudi v prihodnosti. V razdelku obravnavamo štiri slovarske oz. leksikonske vire s poudarkom na njihovi odprtosti in siceršnji dostopnosti.

2.1 Slovarji slovenskega knjižnega jezika

SSKJ je bil izdelan na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU (v nadaljevanju ISJ) z javnim financiranjem in nato, ravno tako z javnim financiranjem, pretvorjen v digitalno obliko. V digitalni obliki je bil najprej prodajan na CD-ROM, nato pa postavljen na splet v prostem dostopu. Spletna različica je namenjena zgolj branju, saj kolofon spletnega slovarja določa, da »Noben del te izdaje ne sme biti reproduciran, shranjen ali prepisan v katerikoli obliki oz. na katerikoli način«. SSKJ tako ni odprto dostopen niti v obliki HTML, še manj pa kot izvorna podatkovna baza. Kolikor nam je znano, dostop do te baze ni bil, enako kot za ročno označeni oblikoskladenjski korpus, nikoli omogočen niti v raziskovalne namene katerokoli drugi raziskovalni inštituciji v Sloveniji.

Pri SSKJ je pogosto omenjan razlog za zaprtost dejstvo, da si avtorske pravice nad slovarjem delijo ZRC in avtorji oz. njihovi dediči, ki naj bi se jim ne hoteli odpovedati. Vendar se lahko upravičeno vprašamo, ali je to resnično glavna ovira, saj si avtorji verjetno od SSKJ ne morejo več obetati komercialne koristi, po drugi strani pa takšen razlog ne drži za druge

slovarje, izdelane na ISJ, pa ti tudi niso odprti. Tako je Slovar novejšega besedja slovenskega jezika – edini nosilec avtorskih pravic je ZRC SAZU – sicer digitalno dostopen v okviru Google Books, in to celo pod licenco Creative Commons, vendar pa v njeni najbolj omejeni različici (samo nekomercialna uporaba in prepoved predelav), predvsem pa je tam na voljo samo v obliki PDF, ki je primerna samo za ogled oziroma tisk, ne pa npr. za iskanje po slovarju, kaj šele za kakršnokoli nadaljnje izkoriščanje podatkov iz slovarja. Tudi pri novi izdaji SSKJ, ki naj bi bila končana v 2014, je bilo na dogodku, ki ga je pred kratkim organiziral ISJ, rečeno, da se ne predvideva odprt dostop in da bo predvidoma tudi dostop preko spleta omogočen samo kupcem knjižne izdaje.

2.2 Leksikon besednih oblik MULTEXT-East

V sredini 90-ih let prejšnjega stoletja je Slovenija (IJS in podjetje Amebis, d. o. o.) sodelovala v evropskem projektu MULTEXT-East,² v okviru katerega smo naredili prvi, čeprav za sedanje razmere majhen jezikoslovno označeni korpus (roman »1984«) in oblikoslovni leksikon slovenskega jezika (Dimitrova in dr. 1998). Leksikon z oblikoslovnimi paradigmi 15.000 lem je prispevalo podjetje Amebis in je odprto dostopen. Uporabljen je bil za raznovrstne namene, npr. za šolanje lematizatorja za slovenski jezik, bil pa je tudi osnova za po številu uporabnikov enega uspešnejših odprtokodnih virov za slovenski jezik, in sicer leksikon besednih oblik za črkovalnik Aspell.³ Ta leksikon (Košir, 2002) je nastal v sodelovanju Društva uporabnikov Linuxa LUGOS, podjetja Amebis, d. o. o., in Inštituta »Jožef Stefan« v okviru projekta takratnega Ministrstva za informacijsko družbo. Aspell (in sedaj Hunspell⁴) je s slovarji za skoraj sto jezikov dostopen pod GNU General Public License in se uporablja za slovenski črkovalnik v OpenOffice, LibreOffice, Mozilla Firefox in Thunderbird, Mac OS X itd., ki skupaj pokrivajo bolj ali manj vse uporabnike, ki za pisanje svojih besedil ne uporabljajo sistema Windows.

2.3 Oblikoskladenjski leksikon Sloleks

Projekt »Sporazumevanje v slovenskem jeziku« (SSJ) je izdelal vrsto korpusov, kot tudi Sloleks,⁵ obsežen oblikoskladenjski leksikon za slovenski jezik (Arhar, 2009), in večina teh

² <http://nl.ijs.si/ME/>.

³ <http://aspell.net/>.

⁴ <http://hunspell.sourceforge.net/>.

⁵ <http://www.slovenscina.eu/sloleks>.

virov je ne samo prosto, temveč tudi odprto dostopna. Odprti viri projekta SSJ so omogočili pravo renesanso na področju jezikovnih tehnologij slovenskega jezika, saj so služili za izdelavo orodij za oblikoskladenjsko označevanje, za lematizacijo, za skladenjsko analizo itd.

Vseeno pa z dostopnostjo leksikona Sloleks ne moremo biti povsem zadovoljni, saj je odprto dostopen samo v nekomercialne namene. Takšna licenca je bila postavljena s strani takratnega ministrstva MIŠZ, kljub temu da za to omejitev ni videti posebnih razlogov, ravno obratno: ministrstvo, ki je lastnik izdelanih virov, verjetno ne računa na to, da jih bo prodajalo komercialnim uporabnikom, ob tem pa s tako omejitvijo izrazito omejuje uporabo izdelanega leksikona, saj ga ni moč vgraditi v nobeno aplikacijo, ki bi se nato prodajala na trgu, s čimer se umetno zavira razvoj jezikovnih tehnologij za slovenski jezik.

2.4 Besedišče starejše slovenščine IMP

Kot zadnji primer izpostavimo besedišče starejše slovenščine IMP,⁶ del jezikovnih virov, ki so bili izdelani v sklopu projekta EU IMPACT in Google nagrade in zajemajo predvsem besedila iz XIX. stoletja (Erjavec, 2012). Ker so izvirnim besedilom avtorske pravice že potekle in so se dajalci digitalnih izvirnikov strinjali z odprto uporabo, so vsi jezikovni viri projekta odprto dostopni, vključno s komercialno rabo. To besedišče bi izpostavili kot primer dobre prakse, saj ima njegovo razširjanje eno samo omejitev: da se mu prizna izvorno avtorstvo.

3 Ustvarjalna gmajna: licence CC

Pojem odprti dostop je zgodovinsko gledano prišel iz gibanja za odprto kodo (OSI), ki se zavzema za dostopnost izvorne kode programske opreme, saj je na ta način omogočena ne samo široka uporaba programov, temveč tudi javno popravljanje napak, prilagajanje in nadgrajevanje. Tu se ne bomo spuščali v podrobnosti definicij licenc za odprto kodo, ker so usmerjene predvsem v programsko opremo, in ne v podatke (jezikovne vire), se je pa na tej osnovi začelo tudi gibanje za odprte podatke v najširšem pomenu te besede, kot so npr. podatkovne baze, znanstveni članki ali pa referenčni viri, od katerih je najbolj znana Wikipedija. Tudi za odprt dostop do podatkov obstaja večje število licenc, vendar so se od

⁶ <http://nl.ijs.si/imp/>.

vseh najbolj uveljavile licence Creative Commons⁷ oz. Ustvarjalna gmajna.⁸ Licence CC so namenjene urejanju dostopa do (umetniških, znanstvenih) avtorskih del in uporabniku podajo ustrezno informacijo, imajo pravno veljavo, so strojno čitljive in kot take povsem primerne tudi kot pravni okvir za odprt dostop do slovarjev. Zadnja različica licenc CC je »4.0 International«, ki se od prejšnjih razlikuje mdr. po tem, da za vse države velja ista formulacija licenc (po možnosti prevedena v posamezne jezike). To je poenostavilo njihovo uporabo.

Licence CC dovoljujejo prevzem avtorskih del in njihovo nadaljnje razširjanje, vendar obstajajo v več različicah, pri katerih lahko medsebojno kombiniramo posamezne omejitve odprtosti; vsega skupaj obstaja šest različnih kombinacij. Če npr. zahtevamo priznanje avtorstva (BY) in obenem prepovedujemo komercialno izkoriščanje (NC, »non-commercial«), bomo delo ponudili pod licenco CC BY-NC. V nadaljevanju tega razdelka obravnavamo posamezne omejitve in jih na kratko komentiramo v povezavi z odprtostjo načrtovanega slovarja.

3.1 Priznanje avtorstva: CC BY

Ker je velika večina ponudnikov podatkov pod licencami CC zahtevala priznanje avtorstva, je sedaj (verzija 4) to postalo kar obvezen del licenc CC. Formulacija »avtorstva« je prepuščena tistemu, ki delo ponuja v odprt dostop. Za Slovar, kot kolektivno delo, je verjetno najprimernejša kar navedba imena slovarja skupaj z naslovom URL, ki predstavlja njegovo domačo stran. Če pa je slovar uporabljen v raziskavi, ki je nato objavljena v znanstvenem članku, naj se avtorstva priznajo tudi skozi citiranje članka oz. člankov, ki opisujejo izdelavo in zgradbo tega slovarja. Korektno citiranje je še posebej pomembno, ker je merljiv kazalec raziskovalne uspešnosti, zaradi česar bi se tudi moralo dosledno izvajati.

3.2 Deljenje pod enakimi pogoji: CC SA

Deljenje pod enakimi pogoji prepoveduje, da bi prejemnik podatkov pri nadaljnjem razširjanju podatkov spremenil licenco. Tako npr. tudi pri izvedenih komercialnih produktih ni dovoljeno zakleniti vira, pač pa mora ta, četudi je predelan, biti na voljo pod enako licenco kot izvirnik. Namen določila je, da so tudi vse dopolnitve oz. izboljšave vira na voljo tretjim osebam. To je sicer zelo koristno, povzroča pa probleme podjetjem, ki morajo zato ponujati

⁷ <http://creativecommons.org/>.

⁸ <http://creativecommons.si/>.

svoje produkte na kompleksnejši način, odpovedati pa se morajo tudi konkurenčni prednosti, ki bi sledila iz izboljšave vira. Zato podjetja včasih celo raje kupijo licenco za zaprto uporabo podatkov od avtorjev, kot da prevzamejo odprte podatke pod licenco SA. V primeru slovarja bi, glede na to, da je delo financirano iz javnih sredstev, bila ta določba mogoče smiselna, nikakor pa ne nujna, saj, kot rečeno, otežuje predvsem komercialne uporabe slovarja.

3.3 Brez predelav: CC ND

Omejitev prepoveduje nadaljnje razširjanje dela, ki je bilo tako ali drugače predelano. Ta omejitev je smiselna za posamezne vrste del, kot so npr. umetniški izdelki ali znanstveni članki, saj tu avtor upravičeno zahteva, da se njegovo delo širi v nespremenjeni obliki. Pri slovarski bazi je stanje manj jasno. Po eni so slovarska gesla plod dela visoko usposobljenih slovaropiscev in zato vlada prepričanje, da bi kakršnakoli predelava pokvarila njihovo delo, s tem da bi slovar bodisi okrnila bodisi vanj vnesla napake. To velja toliko bolj za normativne dele slovarja, na katere se uporabniki zanašajo npr. pri reševanju pravopisnih zadreg.

Po drugi strani pa z omejitvijo ND odvezamemo možnost dopolnitev in tudi popravkov s strani tretjih oseb, ki bodo v veliki večini primerov prej koristne kot pa škodljive, kar najbolje izpričuje razvoj Wikipedije. Potrebno se je tudi zavedati, da predelan slovar ne bi nadomestil izvirnega, pač pa bi predstavljal samo dodatno ponudbo, ki bi morala po CC BY jasno navesti, kaj original je in kje ga najdemo. Poleg tega predelava ne pomeni nujno, da se celotni slovar tako ali drugače prilagodi, npr. za drugačne ciljne uporabnike, pač pa tudi, da se iz njega izvede npr. leksikon, ki je nato uporaben bodisi za jezikovnotehnološke raziskave bodisi za vgrajevanje v produkte. Dostopnost takšnih predelanih del zato lahko bistveno pripomore k širši uporabnosti izvirnega slovarja tudi za namene jezikovnih tehnologij. Zaradi teh dejstev bi bilo priporočljivo, da se omejitve ND nad slovar ne postavlja.

3.4 Nekomercialno: CC NC

Omejitev prepoveduje uporabo dela v komercialne namene in je verjetno najbolj kontroverzna med določili licenc CC, predvsem zaradi tega, ker se jo velikokrat doda skoraj refleksno (glej primer leksikona Sloleks v razdelku 2.3), po drugi pa ima mnogo negativnih posledic. Razlog, zakaj avtorji oz. lastniki avtorskih pravic zahtevajo NC, je, da se sicer strinjajo, da je njihovo delo odprto in brezplačno dostopno, ne pa, da bo neka tretja oseba kovala iz njega dobiček – če dobiček že bo, hočejo, da je njihov oz. da bodo v njem vsaj soudeleženi. Takšen odnos je

upravičen pri delih, ki so jih avtorji naredili zasebno oz. v zasebnih podjetjih, ne pa, če gre za delo, narejeno v javnih inštitucijah, kjer so avtorji za svoje delo že bili plačani iz javnih sredstev in kjer je osnovni namen dela, da je maksimalno izkoriščeno.

Razlog za prepoved komercialnega izkoriščanja virov ISJ, ki je bil večkrat poudarjen, je, da za svoje delo nimajo zadostnih sredstev in so zato prisiljeni uporabljati tržne mehanizme, da ta manjkajoča sredstva nadomestijo, ter da ga k takšnemu ravnanju financer celo spodbuja. S komercializacijo rezultatov raziskav se na prvi pogled res prihranijo javna sredstva, vendar samo, če gledamo na problem zelo ozko in predvsem s stališča izdelovalcev slovarja. Slovarji niso sami sebi namen, pač pa se osmislijo šele z uporabo. Predvsem slovarska baza je posredno uporabna v množici jezikovnotehnoloških aplikacij, ki bi lahko vodile v pospešeno informatizacijo slovenskega jezika. V trenutku, ko se na slovar doda prepoved uporabe v komercialne namene, bodo od podjetij, ki bi sicer slovar uporabila v svojih produktih, ostala le tista, ki bi ob sredstvih, ki bi jih že tako ali tako morala vlagati v razvoj, našla še dodatna sredstva, da bi si zagotovila licenco za komercialno uporabo slovarja. Pri tem se je seveda potrebno zavedati, da je trg za slovenski jezik zelo majhen in s tem tak tudi pričakovani zaslužek takih podjetij. Dodatno obremenjevanje tistih redkih, ki bi vlagala v razvoj tega sektorja, zato vodi v še manjšo ponudbo produktov opremljenosti slovenskega jezika.

Slovar predstavlja jezikovno infrastrukturo, v katero država vlaga tudi zato, da omogoča razvoj in konkurenčnost gospodarstva, in zato omejevanje odprtosti slovarja s prepovedjo komercialne rabe ni smiselno.

4 Zaključek

V prispevku smo obravnavali odprtost načrtovanega slovarja. Najprej smo predstavili dosedanjo prakso pri odprtosti slovarskih oz. leksikalnih virov. Tu se pokaže, da imamo v Sloveniji dva, skoraj diametralno nasprotna pristopa. Na eni strani so leksikoni in drugi jezikovni viri, ki so nastali za bolj (vendar ne zgolj) jezikovnotehnološke in korpusno-jezikoslovne namene, v največji meri s strani podpornikov novega predloga Slovarja slovenskega jezika. Na drugi strani so slovarji in drugi priročniki za slovenski jezik, ki so nastali v ISJ. Prvi pristop, pri katerem so, zanimivo, pogosto sodelovali tudi zasebni partnerji, odlikuje brezplačen in odprt dostop do izdelanih jezikovnih virov. Drugi temelji na njihovem zaklepanju, pri čemer so izvirne baze vseh slovarjev in drugih izdelanih virov in celo

slovarske sheme (koncepti) skrbno varovana poslovna skrivnost in niso dostopne niti v raziskovalne namene drugim raziskovalnim institucijam v Sloveniji. S tem se, vsaj do sedaj, osrednja slovenska institucija, ustanovljena za to, da skrbi za razvoj in opremljenost slovenskega jezika, in financirana z izdatnimi javnimi sredstvi, izpostavi kot največja ovira pri diseminaciji in široki uporabnosti virov, ki so ključni za informatizacijo slovenskega jezika.

V drugem delu prispevka obravnavamo različne pravne formulacije odprtosti. Osredotočili smo se na licence Creative Commons. Tu se zavzemamo za uporabo licence CC BY (priznanje avtorstva). Nadaljnje omejitve, predvsem CC BY-ND (brez predelav) in CC BY-NC (nekomercialno) ali celo CC BY-ND-NC, bi bistveno zmanjšale uporabnost in s tem uporabo slovarja. Omejitvev ND na primer pomeni, da bi bilo potrebno, podobno kot je bilo to z oblikoskladenjsko označenim korpusom, izdelavo odprte predelane slovarske baze za specifične jezikovnotehnološke namene financirati povsem na novo. To bi vodilo v dvojno javno financiranje podobnih jezikovnih virov, ob tem pa dodatno upočasnilo razvoj jezikovnih tehnologij za slovenski jezik.

Kot je bilo pojasnjeno, ima tudi omejitvev NC negativne posledice za uporabo slovarja in s tem za informatizacijo slovenskega jezika. Prihodki nosilca oz. nosilcev avtorskih pravic nad slovarjem, ki jih lahko pridobi iz prodaje podatkovne baze posameznim podjetjem, so majhni, vsaj v primerjavi z javnimi sredstvi, potrebnimi za izdelavo samega slovarja. Po drugi strani pa bi dejstvo, da je podatke potrebno plačati, odgnalo večino že tako redkih podjetij, ki se trudijo razvijati informacijske produkte, vezane na uporabo slovenščine. Res bodo takšni produkti plačljivi in avtorji slovarja od teh prihodkov ne bodo dobili nič, zato pa vsi, ki so skozi davke financirali izdelavo slovarja, dobijo možnost, da kupijo uporabo slovenskega jezika, kjer je sicer ne bi bilo, obenem pa se spodbuja kompetitivnost domačega gospodarstva.

Gornje misli o nujnosti licence CC BY za načrtovani slovar strne citat iz poročila Evropske komisije (European Commission, 2012), ki se v prevodu glasi:

Celovitejši in širši dostop do znanstvenih publikacij in podatkov bo pripomogel k:

- *pospeševanju inovacij (hitreje do trga = hitrejša rast);*
- *spodbujanju sodelovanja in izogibanju ponavljanja dela (povečana učinkovitost);*
- *dograjevanju predhodnih rezultatov raziskav (izboljšana kvaliteta rezultatov);*
- *vključevanju državljanov in družbe (povečana transparentnost znanstvenega procesa).*

Literatura

ARHAR, Špela, 2009. Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo* 54/3–4, 43–56. <http://www.jezikinslovstvo.com/pdf/2009-03-04-Razprave-Spela-Arhar.pdf>.

DIMITROVA, Ludmila, ERJAVEC, Tomaž, IDE, Nancy, KAALEP, Heiki-Jan, PETKEVIČ, Vladimir, TUFİŞ, Dan, 1998: MULTEXT-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. Zbornik *COLING-ACL '98*, 315–319. <http://aclweb.org/anthology//P/P98/P98-1050.pdf>.

ERJAVEC, Tomaž, 2009. Odprtost jezikovnih virov za slovenščino. V: Zbornik Simpozija Obdobja »Infrastruktura slovenščine in slovenistike« Ljubljana: Znanstvena založba Filozofske fakultete, str. 115–121.

<http://www.centerslo.net/files/file/simpozij/simp28/Erjavec.pdf>.

ERJAVEC, Tomaž, KREK, Simon, 2008: Oblikoskladenjske specifikacije in označeni korpusi JOS. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Zbornik Šeste konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 49–53. <http://nl.ijs.si/is-ltc08/IS-LTC08-Proceedings.pdf>.

ERJAVEC, Tomaž, 2012. Jezikoslovni viri starejše slovenščine. *Knjižnica* 56/3, tematska številka »Ljubljana v BiTiH - BiTi v Ljubljani : prispevki iz prvega ljubljanskega kongresa digitalizacije kulturne dediščine«. str. 205–221. <http://knjiznica.zbds-zveza.si/index.php/knjiznica/article/view/151>.

EUROPEAN COMMISSION, 2012. *Towards better access to scientific information: Boosting the benefits of public investments in research*. http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf.

JAKOPIN, Primož, BIZJAK, Aleksandra, 1997. O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična revija* 1997/3–4, str. 513–532. [URN:NBN:SI:DOC-GIBTLWVE](http://nbn-resolving.org/urn:nbn:si:doc-gibtlwve).

KOŠIR, Aleš, 2002. Sodobni prosti črkovalniki in baze pravih besednih oblik. *Zbornik konference Jezikovne tehnologije*, IJS. <http://nl.ijs.si/isjt02/zbornik/sdjt02-D06kosir.pdf>.

KOTAR, Mojca, 2013: Odprti dostop v Evropski uniji in v Sloveniji. *Knjižničarske novice* 23/10. <http://www.nuk.uni-lj.si/knjiznicarskenovice/v2/podrobnostClanek.aspx?id=778>.

(Dostop do vseh spletnih virov 26. 2. 2014.)