Simon Krek
Amebis, d.o.o., Kamnik
Institut Jožef Stefan, Ljubljana

UDK 811.163.6'26:81'32

OD SSKJ DO SPLETNEGA PORTALA STANDARDNE SLOVENŠČINE

V prispevku obravnavamo stanje priročniških virov za slovenski jezik in orišemo predlog programa za izgradnjo portala medsebojno povezanih jezikovnih virov v elektronski obliki. Predlog je zasnovan v štirih sklopih, ki zajemajo področja jezikovnega opisa, standardizacije jezika, terminoloških in dvojezičnih virov. Na kratko orišemo razvoj podobnih virov za druge jezike in podajamo primere teh virov po posameznih področjih. Ugotavljamo, da je stanje glede razvitosti in dostopnosti virov za slovenščino nezadostno in da bi k izboljšanju lahko pripomoglo bolj domišljeno načrtovanje in obstoj osrednjega institucionalnega telesa z možnostjo realnega dodeljevanja finančnih sredstev.

Ključne besede: jezikovno načrtovanje, jezikovni viri, jezikovni priročniki, računalniško jezikoslovje, korpusno jezikoslovje

1 Uvod

Pokrajina področja jezikovnih priročnikov in virov za katerikoli jezik se je v zadnjih petindvajsetih letih dodobra spremenila. Množična raba osebnih računalnikov in drugih elektronskih naprav, pojavitev in silovit razvoj svetovnega spleta ter razmah informacijskih tehnologij bi vsakega uporabnika Commodorja 64 ali Atarija 800 iz orwellovega leta 1984, ki bi brez blažilnih vmesnih korakov preskočil v leto 2009, naravnost osupnila. Naravni jezik kot sredstvo medčloveške komunikacije je seveda izjemno pomemben del informacijskih in komunikacijskih tehnologij, zato ne čudi, da se je vzporedno z omenjenim razvojem dogajal tudi bliskovit napredek pri računalniškem obvladovanju naravnih jezikov, ki je generiral celo

vrsto podpodročij in aktivnosti, kjer se srečujeta računalnik in naravni jezik. Svet jezikovnih priročnikov in orodij je močno zaznamoval vstop statističnih metod strojnega učenja v analizo jezikovnih podatkov, ki so bile na to področje prenesene z drugih področij računalniške obdelave podatkov, in na tem danes temeljijo tako pomembna področja in aplikacije, kot so sinteza in analiza govora (ang. *speech generation*, *speech recognition*), strojno in strojno podprto prevajanje (ang. *machine translation*, *machine-aided translation*), dialoški sistemi (ang. *question answering*, *spoken dialog systems*), avtomatsko povzemanje (ang. *automatic summarization*), luščenje podatkov (angl. *information retrieval*), tehnologije obdelave besedil, kot so oblikoskladenjsko, skladenjsko in semantično označevanje in razdvoumljanje (ang. *tagging*, *parsing*, *word sense disambiguation*), prepoznava poimenovalnih izrazov (ang. *named entity recognition*), razreševanje medstavčnih referentov (ang. *anaphora resolution*), tehnologije semantičnega spleta (ang. *Semantic Web 2.0*) in še bi lahko naštevali. Nenazadnje je možnost obdelovanja velikih količin jezikovnih podatkov spremenila tudi samo jezikoslovje kot vedo, kjer se je zgodil radikalen premik od razmišljanja o jeziku kot sistemu, ki je v precejšnji meri zanemarjalo jezikovno realnost, k razmišljanju o jeziku kot diskurzu, kjer je poudarjena neločljivost različnih nivojev jezikovne analize (oblikoslovna, skladenjska, pomenska, pragmatična) in gradacijska narava obravnave jezikovnih pojavov s prototipičnimi primeri v središču in sivimi polji na robovih. To pa ob razvoju drugih ved, kot so psihologija, pedagogika, didaktika itd., ni ostalo brez vpliva na tradicionalna jezikovnopriročniška področja, kot so enojezična in dvojezična leksikografija, normativistika, terminografija, jezikovna didaktika za materni in tuje jezike itd.

Tako kot Gutenbergova tiskarska revolucija v 15. stoletju je tudi informacijska revolucija v 21. stoletju vsak jezik oziroma govorce tega jezika doletela ne glede na to, ali so bili na to pripravljeni ali ne. Povsem logično je, da so bili jeziki z velikim številom govorcev, obilico nakopičenega znanja in bogato finančno podporo pri tej revoluciji v prednosti, na čelu z angleščino kot jezikom globalne komunikacije. Ta prispevek želi zastaviti nekatera vprašanja o specifični situaciji slovenskega jezika v omenjenem kontekstu in nanje odgovoriti z enim od možnih načinov prilagajanja globalnim trendom, in sicer pri jeziku, ki ima svojo zgodovino, specifično zatečeno stanje pri jezikovnih priročnikih in virih ter omejene človeške in finančne vire za njihovo izdelavo.

2 Izhodišče

Primer dobre prakse kot izhodišče za sprehod skozi nujne sodobne vire in orodja za slovenski jezik lahko najdemo pri estonščini – jeziku, ki je imel ob spremembi družbenega sistema ob koncu prejšnjega stoletja verjetno celo slabše izhodišče in ima skoraj pol manjše število maternih govorcev kot slovenščina. Estonščina je bila do leta 1991 potopljena v izjemni večjezičnosti Sovjetske zveze, ki je uveljavljala politiko vsiljene dvojezičnosti pri nerusko govorečih prebivalcih, po vzpostavitvi neodvisnosti pa se je Estonija soočala z jezikovno

situacijo, ko kar precejšen del prebivalstva ni govoril po novem edinega uradnega jezika države. Danes je eden od uradnih jezikov Evropske unije s približno 1,25 milijona govorcev. Leta 2004 je vlada republike Estonije sprejela *Strategijo razvoja estonskega jezika 2004–2010*, v kateri je postavila izhodišča za nadaljnje financiranje dejavnosti v zvezi z estonščino. Dokument je do neke mere primerljiv z *Resolucijo o nacionalnem programu za jezikovno politiko 2007–2011*, ki je bila sprejeta maja 2007 v slovenskem parlamentu, zato si je zanimivo ogledati nekatere rešitve. Razlika, ki je vidna že na prvi pogled, je sistematična priprava na izdelavo Strategije, ki je trajala tri leta s študijami po posameznih področjih, in sicer: 1. analiza raziskovanja estonskega jezika, 2. področne študije o stanju standardne estonščine, 3. stanje pri načrtovanju estonščine za specializirana področja (LSP – *Language for Special Purposes*), 4. analiza načrtovanja glede rabe lastnih imen, 5. jezikovnotehnološki viri za estonski jezik, 6. estonski znakovni jezik, 7. materni jezik in drugi jeziki: estonščina kot drugi/tuji jezik in stanje glede tujih jezikov v Estoniji, 8. estonščina v srednjih šolah z ruskim učnim jezikom. V času priprave je bilo organiziranih sedem konferenc, tudi s tujimi predavatelji, predlog Strategije pa je bil po objavi nekaj manj kot leto dni v javni razpravi. Natančnejša obravnava estonske Strategije v okviru tega prispevka ni mogoča, čeprav bi bila zanimiva, pomembno pa je, da so bili v skladu s to strategijo sprejeti konkretni programi, ki so omogočili implementacijo njenih načel. Eden od takih programov je *Nacionalni program za estonske jezikovne tehnologije (2006–2010)*, katerega namen je »razviti jezikovne tehnologije za estonščino do stopnje, ki bo omogočila nemoteno delovanje jezika na področju informacijskih tehnologij in vzpostavitev sodobne jezikovnotehnološke infrastrukture. Podrejena cilja sta razviti potrebno programsko opremo za jezikovne tehnologije ter komplementarne jezikovne vire.«[1] Skupna vrednost programa je približno 4,5 milijona evrov, pri čemer je tretjina namenjena za ustvarjanje jezikovnih virov, dve tretjini pa za raziskave in razvoj programske opreme. Program upravlja in nadzoruje odbor z devetimi člani, v letu 2008 pa je bilo iz njega financiranih 23 projektov. Vsi rezultati, ki jih financira program, so prosto dostopni.

Estonski program za jezikovne tehnologije sestavljajo tri področja ukrepov:

1. govorne tehnologije, ki zajemajo tehnologije in vire za:
   – sintezo govora:[2] izdelava programske opreme za TTS (text-to-speech) in razvoj prototipa za avdiovizualno sintezo,

[1] <http://www.eki.ee/keelenoukogu/strat_en.pdf.>

[2] Govornih tehnologij v nadaljevanju ne bomo natančneje obravnavali, zato za lažjo orientacijo navajamo nekaj tovrstnih programov za slovenski jezik. Govorne tehnologije za slovenščino so razmeroma razvite, čeprav seveda daleč od stanja pri najbolj razširjenih jezikih. Primer sistema za sintezo govora je prosto dostopen program *Govorec* (<http://ai.ijs.si/govorec/>, <http://govorec.amebis.si/>, <http://www.rtvslo.si/mmcrtvgovorec>), ki je nastal v sodelovanju med Institutom Jožef Stefan in podjetjem Amebis, d.o.o. Še en program s podobno funkcionalnostjo je razvilo podjetje Alpineon, d.o.o., ki ga je mogoče preizkusiti na strani <http://www.alpineon.com/proteus/test/>.

- prepoznavo govora:[3] razvoj prototipa za ASR (automatic speech recognition) z omejenim slovarjem in razvoj jezikovno specifičnih metod za prepoznavo govora brez slovarskih omejitev,
- dialoški sistemi:[4] izdelava področno omejenih inteligentnih sistemov, ki so zmožni nadomestiti rutinska dela.

2. tehnologije za obdelavo pisnega jezika, ki vsebujejo:
   - metode procesiranja jezika:[5] postopki za avtomatsko obdelavo jezika na različnih ravneh (oblikoslovje, skladnja, semantika, pragmatika), modeliranje in izdelava ustreznih prototipov,
   - strojno prevajanje:6 razvoj metod za prevajanje v estonščino ali iz nje, sestavljanje večjezičnih slovarskih baz in razvoj mehanizmov za pretvorbo skladenjskih struktur; razvoj prototipa za estonsko-angleško-estonsko strojno prevajanje.
3. jezikovni viri, kar zajema:
   - ustvarjanje infrastrukture za zbiranje različnih jezikovnih virov in upravljanje z njimi,
   - zbiranje različnih tipov jezikovnih virov: govorni in pisni korpusi in slovarske baze podatkov.

V prispevku nas glede na izbrano jezikovnopriročniško tematiko najbolj zanima tretji del oziroma tretje področje ukrepov – jezikovni viri. Če si torej ogledamo, kaj program predvideva na tem segmentu (gl. **Dodatek 1**) in na kratko analiziramo primerjavo med predpostavljenim stanjem estonskih jezikovnih virov glede na njihov nacionalni program financiranja in realno stanje slovenskih jezikovnih virov, ugotovimo predvsem to, da: (1) so že izdelani slovenski viri v precejšnem delu plačljivi in da (2) za slovenščino obstaja izrazita vrzel pri enojezičnih

---

[3] Prepoznava govora je od dveh temeljnih postopkov govornih tehnologij nedvomno težji del. Prosto dostopnih splošnih programov za slovenščino ni, čeprav že obstajajo integrirani sistemi, ki uporabljajo omejene slovarje, npr. Kolosejeva M-vstopnica. Do neke mere je za individualno rabo mogoče uporabiti tudi jezikovno neodvisne sisteme, ki so učljivi, npr. *Dragon NaturallySpeaking* (<http://www.nuance.com/naturallyspeaking/>). Bralca, ki ga zanimajo podrobnosti postopkov in stanje na področju, lahko napotimo na zbornike konferenc *Jezikovne tehnologije* na strani Slovenskega društva za jezikovne tehnologije (<http://www.sdjt.si/>).

[4] Tudi dialoški sistemi so za slovenščino že na voljo. Verjetno trenutno najbolj znan delujoč sistem je *Virtualna davčna asistentka Vida*, ki se ji je v letu 2009 pridružila še *Telekomova interaktivna asistentka Tia*. Oba sistema uporabljata tehnologijo mednarodnega podjetja *Artificial Solutions* švedskega izvora. Drugi sistem, ki je bolj splošne narave, je program *Kolos* oziroma *Klepec* (<http://klepec.amebis.si/>) podjetja Amebis, d.o.o.

[5] Programska oprema, ki jo običajno asociiramo s temi postopki, so predvsem čim bolj uspešni (oblikoslovni) označevalniki (ang. *tagger*) in (skladenjski) razčlenjevalniki (ang. *parser*). Za slovenščino imamo dva označevalnika, ki jih lahko preizkusimo, na Institutu Jožef Stefan (<http://nl.ijs.si/jos/analyse/>) in na ZRC SAZU (<http://bos.zrc-sazu.si/dol_lem1.html>), pri čemer zadnji ne razdvoumlja med več možnimi osnovnimi oblikami in oblikoskladenjskimi oznakami glede na sobesedilo. Označevalnik (in razčlenjevalnik) je tudi del programa *BesAna* (<http://besana.amebis.si/>) za slovnično pregledovanje besedil podjetja Amebis, d.o.o., z njim sta bila med drugim označena korpusa *FIDA* (<http://www.fida.net>) in *FidaPLUS* (<http://www.fidaplus.net>).

[6] Poleg sistema strojnega prevajanja podjetja Google, ki je v letu 2008 ponudilo brezplačni sistem tudi za slovenščino (<http://translate.google.com/>) – ta deluje na statističnih modelih strojnega prevajanja – je za slovenščino na voljo še sistem *Presis* (<http://presis.amebis.si/>) podjetja Amebis, d.o.o., ki deluje po metodi prevajanja na podlagi pravil.

in dvojezičnih leksikalno-skladenjsko-semantičnih virih, kot so leksikalno-skladenjska baza, leksikalno-semantična baza, tezaver in angleško-slovenska slovarska baza. Kar pravzaprav ne čudi, saj so ravno tovrstni viri najbolj zahtevni z izvedbenega in finančnega stališča. V nadaljevanju nas bodo torej zanimali predvsem leksikalni viri za slovenščino, pretekli in morebitni bodoči, njihova povezljivost in dostopnost.

3 Obstoječi leksikalni viri za slovenščino

Besedilnih korpusov kot primarnega vira za izdelavo jezikovnih priročnikov v tem prispevku ne bomo podrobneje obravnavali, tematika je znana in obdelana v literaturi (Gorjanc 2005; Gorjanc 2006; Arhar in Gorjanc 2007; Jakopin in Michelizza 2007/2008; Vintar 2008; Vintar in Logar 2008). Zanima nas stanje glede sekundarnih, iz primarnih »izvedenih« priročniških virov, kot so različni slovarji oziroma podatkovne baze slovarskega tipa (splošne, dvojezične, terminološke, pravopisne itd.), slovnice in didaktična gradiva – predvsem v kontekstu v uvodu zastavljenega vprašanja – ter njihova dostopnost.

Intenzivnejše ukvarjanje z digitalnimi slovarskimi viri se je vsaj pri angleškem jeziku začelo že v zgodnjih osemdesetih letih prejšnjega stoletja s prvimi večjimi digitalizacijami obstoječih slovarjev,[7] kot so LDOCE, OALD, OED itd. in njihovim podrobnim raziskovanjem (Boguraev in Briscoe 1989) oziroma kot bi rekli danes – luščenjem skladenjskih, semantičnih in drugih informacij iz njih. Britanski digitalni projekt osemdesetih let, ki ga težko obidemo, je tudi korpus in korpusni slovar (ter korpusna slovnica) *Cobuild*,[8] na ameriški strani pa se je leta 1985 denimo začel danes izjemno razširjen projekt *WordNet*.[9] V devetdesetih letih so zahodnoevropski jeziki v digitalno dobo vstopili na vso moč. Takrat še Evropska gospodarska skupnost in posamezne močnejše evropske države so jezikovne tehnologije in vire za »svoje« jezike izdatno raziskovalno podprle in v tem času – od leta 1993 v okviru Evropske unije – so bili za leksikalne vire jezikov EU izdelani predlogi standardov (npr. *EAGLES: Expert Advisory Group on Language Engineering Standards – 1993-1996*) ter digitalni leksikalni viri v okviru projektov, kot so ACQUILEX (*Acquisition of Lexical Knowledge – 1989-1995*), MULTILEX (*A Multi-Functional Standardised Lexicon for European Community Languages –1990-1993*), GENELEX (*GENEric LEXicon – 1990-1994*), CEGLEX (*Central European Genelex Model – 1995-1996*), DELIS (*Descriptive Lexical*

[7] Hkrati s širjenjem rabe računalnikov in digitalizacijo virov se je pojavila tudi potreba po standardizaciji zapisa digitaliziranih virov. Leta 1986 je bil sprejet ISO standard SGML (*standard generalized markup language*), ki je zdaj vseprisoten na svetovnem spletu in drugod v svojih izvedenkah HTML (*hypertext markup language*) in XML (*extended markup language*).

[8] »Leksikografi, ki so delali pri prvi izdaji slovarja Cobuild so bili med privilegiranimi. Sploh prvič so imeli dostop do dokaznega gradiva, ki jim je omogočilo opazovati, kako se besede obnašajo v svojem okolju. Odpreti novo konkordančno listo za vse rabe neke besede v 7,3-milijonskem korpusu je bilo tako, kot bi nekega sončnega zimskega jutra odprli okno z razgledom na pokrajino, prekrito s svežim snegom.« (Hanks 2007: *John Sinclair Obituary*, *Euralex Newsletter*.)

[9] Gl. <http://wordnet.princeton.edu/>.

Specifications and Tools for Corpus-based Lexicon building – 1993-1995), MECOLB (*Multilingual Environment for Corpus-Based Lexicon Building – 1994-1995*), PAROLE (*Preparatory Action for Linguistic Resources Organization for Language Engineering –1996-1998*) in njegovo nadaljevanje SIMPLE (*Semantic Information for Multifunctional Plurilingual Lexicons – 1998-2000*), MULTEXT (*Multilingual Text Tools and Corpora – 1994-1996*), Multext-East (*Multilingual Text Tools and Corpora for Central and Eastern European Languages – 1995-1997*), ELAN (*European Language Activity Network – 1998-1999*), CONCEDE (*Consortium for Central European Dictionary Encoding – 1998-2000*) in TELRI (*Trans European Language Resources Infrastructure, I – 1995-1998 in II – 1999-2001*). Hkrati so se oblikovala telesa in organizacije za hrambo in distribucijo leksikalnih virov: RELATOR (*A European Network of Repositories for Linguistic Resources – 1993-1995*) in ELRA (*European Language Resources Association – 1995-*).

Raziskovalno dogajanje na področju digitalnih leksikalnih virov je slovenščino v prvi polovici devetdesetih let bolj ali manj obšlo, deloma najbrž zaradi razburljivega političnega dogajanja in s tem povezane nejasne (tudi finančne) prihodnosti, s spremenjenim družbenim sistemom in novo proevropsko usmerjenostjo pa je v evropske raziskovalne projekte prvič vstopila leta 1995 v projektu TELRI I, pri katerem sta sodelovala Institut Jožef Stefan in ZRC SAZU, kasneje še v projektih Multext-East (IJS), TELRI II (IJS in ZRC SAZU), CONCEDE (IJS) in ELAN (IJS). V okviru projekta Multext-East so bile tako izdelane oblikoskladenjske specifikacije oziroma nabor oznak za jezikoslovno označevanje slovenščine, manjši govorni korpus branih besedil (2.000 besed), leksikon besednih oblik (15.000 leksikonskih enot), oblikoskladenjsko označen prevod romana *1984* Georgea Orwella (100.000 besed) in manjši primerljivi korpus (proza 100.000 besed, periodika 100.000 besed).[10] V okviru projekta TELRI je nastal večjezični arhiv računalniških orodij in virov z imenom TRACTOR,[11] v katerem so bili zbrani večinoma korpusni viri za slovenski jezik: slovenska proza v formatu HTML, korpusni viri projekta Multext-East, korpus časopisnih besedil (270.000 besed), Kosmačev korpus (18 literarnih del Cirila Kosmača), časopis DELO (111.000 besed) in prevod Platonove *Republike*. Rezultat projekta ELAN je bil vzporedni korpus ELAN,[12] v projektu CONCEDE pa je bila izdelana formalna struktura enojezičnih in dvojezičnih slovarjev več jezikov.[13] Kot lahko razberemo iz zgoraj navedenega, je pri teh projektih večinoma šlo za standardizacijo formatov in postopkov ter za korpusne vire, kar je logično, saj je bila strojna obdelava in dodajanje jezikoslovnih podatkov velikim količinam pisnih besedil novost, standardizacija pa je vedno potrebna po začetku uvajanja novih tehnologij in je bila pred tem v precejšnji meri že opravljena za »razvitejše« jezike.

[10] Gl. <http://nl.ijs.si/ME/>.
[11] Gl. <http://tractor.bham.ac.uk/tractor/catalogue.html>.
[12] Gl. <http://nl.ijs.si/elan/>.
[13] Gl. <http://www.itri.brighton.ac.uk/projects/concede/>, <http://nl.ijs.si/telri/>.

Po razmeroma obetavnem začetku razvoja novih digitalnih virov za slovenščino v drugi polovici devetdesetih let[14] po letu 2000 lahko zaznamo presenetljiv zastoj. Ker gre pri večjih virih leksikalnega oziroma slovarskega tipa vedno za projekte, ki zahtevajo ogromno človeškega napora in s tem povezanih finančnih virov, je zanje značilno dvoje: (1) če se le da, novi viri temeljijo na že opravljenem delu predhodnikov (= »kopičenje«), (2) temeljne raziskave, ki lahko vodijo tudi do neuspešnih rezultatov, ko se tehnologija še vzpostavlja, se ponavadi dogajajo pri večjih jezikih, ker si jih jeziki z manjšim številom govorcev težko privoščijo, saj imajo bistveno manj raziskovalnih potencialov. Kakorkoli se to zdi oportuno, se mali jeziki navadno zgledujejo po preverjenih rešitvah za velike (= »prevzemanje«).

Na problematiko tako lahko pogledamo z dveh strani:

- za računalniško obdelavo naravnih jezikov (ang. *natural language processing* – NLP) potrebujemo podatke (oblikoslovne, skladenjske, pomenske, pragmatične), ki jih v drugačni, ljudem namenjeni obliki že vsebujejo predhodno obstoječi viri, npr. enojezični, dvojezični slovarji, ti pa morajo biti spremenjeni tako, da čim bolj olajšamo strojno branje in »izključimo« človeškega uporabnika in
- iz različnih jezikovnih virov (korpusi, obstoječi slovarji, zdaj svetovni splet, katerikoli računalniško berljivi vir ...) lahko s pomočjo postopkov računalniške obdelave naravnih jezikov dobimo jezikovne podatke (oblikoslovne, skladenjske, pomenske, pragmatične), ki jih uporabimo v leksikalnih virih za jezikovne priročnike, eksplicitno namenjene človeškemu uporabniku.

Priti do realne možnosti uspešne uporabe istih podatkov za oba namena pa nikakor ni enostavno, saj sta tako narava kot zapis teh zelo različna.

Če se za hip ozremo v nasprotno smer, tj. na pot od predhodno obstoječih knjižnih leksikalnih virov do njihove digitalizacije in potencialne rabe za namen računalniške obdelave naravnih jezikov, lahko ugotovimo, da se je digitalizacija slovenskih leksikalnih virov za splošno rabo na osebnih računalnikih pravzaprav zgodila razmeroma hitro, raba teh virov za namen računalniške obdelave naravnih jezikov pa je bila v vseh skoraj dvajsetih letih njihovega obstoja takorekoč marginalna. Če kot vzorčni primer digitalizacije vzamemo v obzir usodi edinega knjižnega enojezičnega slovarja in do nedavna največjega tujejezično-slovenskega slovarja (Grad, Škerlj in Vitorovič 1978), lahko ugotovimo, da so se prve (piratske) verzije angleško-slovenskega slovarja za rabo na osebnem računalniku z operacijskim sistemom DOS pojavile že ob koncu osemdesetih let.[15] Ista slovarska baza je bila nekaj časa na voljo tudi na spletu, potem pa jo je leta 1997 nadomestila legalna verzija na disketah in kasneje na plošči CD-ROM za osebno oziroma mrežno

[14] Treba je denimo omeniti, da je slovenščina 100-milijonski oblikoslovno označen in zvrstno uravnotežen referenčni korpus *FIDA* dobila že leta 2000, le pet let po objavi korpusa BNC (*British National Corpus*) in takorekoč hkrati z neprimerno bolj tehnološko razvito in podprto češčino (*Czech National Corpus*).

[15] Avtor tega prispevka je prvo verzijo slovarske baze za DOS, prepisane iz angleško-slovenskega slovarja Grad, Škerlj in Vitorovič 1978, dobil leta 1990, takrat še na 5¼-palčnih disketah.

rabo. SSKJ je po izidu zadnje knjige (1991, T–Z) leta 1994 dočakal izdajo v eni knjigi, za katero je bilo treba digitalizirati celotno besedilo slovarja. Na disketah in za splošno osebno rabo je bil slovar prvič objavljen v digitalni obliki leta 1997, tako kot angleško-slovenski slovar v slovarskem iskalnem programu ASP podjetja Amebis. Od leta 2000 je SSKJ prosto dostopen tudi na spletu, v preprostejšem iskalnem programu za množično rabo.

Kar ob digitalizaciji podatkov v SSKJ in dvojezičnih slovarjih v zadnjih dvanajstih letih njihovega obstoja bije v oči, je predvsem to, da takorekoč ni poročil, da bi v raziskovalnih projektih ali drugih (komercialnih) projektih ti slovarji nastopali kot podatkovna zbirka (ang. *data set*), iz katerih bi s pomočjo računalniške obdelave izločili in uporabili podatke za druge (računalniške leksikalne) namene. Pri SSKJ sta sicer postopkovno neproblematični izjemi obrnjeni slovar (Holz in dr. 1996) in spletni slovar slovenskih besed,[16] pri dvojezičnih slovarjih pa vsaj v primeru novega *Velikega angleško-slovenskega slovarja* obrat ene smeri slovarja za pripravo obratne slovensko-angleške strani (Krek 2005–2006; Kocjančič idr. 2008). Iz poročil (Jakopin in Bizjak Končar 2008) lahko dodatno sklepamo, da je bil geslovnik SSKJ uporabljen za leksikon besednih oblik, ki ga na ZRC SAZU uporabljajo za oblikoslovno označevanje besedil, vendar leksikon ni prosto dostopen ali podrobneje opisan. Resnejše in bolj množično raziskovalno računalniško luščenje prevodnih, skladenjskih, semantičnih in drugih jezikovnih podatkov iz že obstoječih slovarjev se v slovenščini za razliko od zgodnjih poskusov za druge jezike torej nikdar ni zgodilo.[17]

Prvi in najbrž odločilni razlog za to je lastništvo podatkovnih baz in njihove avtorske pravice. Če se ozremo na usodo SSKJ in njegovega *copyrighta* oziroma lastništva avtorskih pravic, se je po založniških kolofonih različnih izdaj spreminjalo na zanimiv in po svoje predvidljiv način. Prva izdaja v petih knjigah, ki je izhajala v letih 1970–1991, oznake za *copyright* nima – v kolofonu je preprosto navedeno, da je knjigo izdala SAZU oziroma kasneje ZRC SAZU, založila pa Državna založba Slovenije. To je bil čas socialistične družbene ureditve in zasebna lastnina, s tem tudi avtorske pravice, ni bila na dobrem glasu. Prvič se oznaka za *copyright* pojavi leta 1994 že v novem družbenem sistemu, ob izdaji v eni knjigi – torej takrat, ko je bil slovar tudi digitaliziran in pripravljen za računalniško luščenje podatkov, kar je tudi opisano v predgovoru v to izdajo – in tam so kot lastniki avtorskih pravic navedeni »*Inštitut za slovenski jezik Frana Ramovša ZRC SAZU in avtorji*«. Dodatek »*in avtorji*« je zanimiv zato, ker bi se avtorskopravno gledano glede potencialnih sprememb ali »računalniške rabe« slovarske baze morali (pisno) strinjati vsi v kolofonu navedeni avtorji slovarja ali njihovi dediči, ob stotinah sodelavcev bi bil to prav gotovo izjemen logističen napor, ki bi mu lahko že vnaprej napovedali neuspeh. Dodatek »*in avtorji*« kasneje doživi spremembo: pri elektronski izdaji SSKJ (ponatis 2005) se na CD-ju dodatek pojavi, v kolofonu

[16] Temu bi morda lahko dodali še projekt SI-PRON (Jakopin idr. 2006), o katerem pa je bilo le poročano, rezultati pa v času pisanja članka niso na voljo.

[17] Glede raziskovalnega izkoriščanja komercialnih slovarskih virov gl. tudi Fontenelle 1997.

iste publikacije pa je kot lastnik avtorskih pravic samo »*SAZU in ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša*«. Enaka formulacija je navedena tudi v kolofonu v izdaji časnika *Dnevnik* iz leta 2008. Iz tega bi bilo mogoče sklepati, da je v tem trenutku le od odločitve ZRC SAZU odvisno, ali bo SSKJ kdaj splošno dostopen tudi kot podatkovna zbirka za (računalniške) raziskovalne namene.

Pri dvojezičnih slovarjih je lastništvo avtorskih pravic še preprostejše: če imamo pri SSKJ na strani kot lastnika akademsko raziskovalno ustanovo, ki se po naravi stvari lahko odloči za predajo podatkovne zbirke v javno (raziskovalno) rabo, so dvojezični slovarji večinoma v lasti zasebnih založb oziroma avtorjev ali dedičev avtorjev in kot podatkovne zbirke brez dobro načrtovanega sodelovanja med javno raziskovalno in zasebno sfero za raziskovalno rabo bolj ali manj izgubljeni tudi v prihodnje.

Eden od pokazateljev, kako pomembno je vprašanje lastništva avtorskih pravic pri izkoriščanju jezikovnih podatkovnih zbirk za namene računalniške obdelave jezika, je usoda zbirk pravnih besedil, ki so bila prevedena v procesu pridruževanja srednje- in vzhodnoevropskih držav Evropski uniji. Iz njih je takorekoč »nehote« nastal tudi ogromen vzporedni korpus prevedenih besedil in ker so ta besedila večinoma v lasti državnih institucij in Evropske unije, so prosto dostopna tudi za računalniško obdelavo in tudi že obilno izkoriščena (Erjavec idr. 2005; Erjavec 2006; Erjavec in Bence 2006; Steinberger 2006; Tufis idr. 2008).[18] Podobno usodo doživlja tudi *Wikipedija* kot čedalje večja zbirka besedil v različnih jezikih, ki je kot podatkovna zbirka dostopna pod licenco *GNU Free Documentation Licence*,[19] ki v osnovi določa, da je podatke mogoče prosto uporabljati, dokler uporabnik daje na voljo predelave pod istimi pogoji in priznava avtorstvo originala. Za slovenščino je bila denimo *Wikipedija* na ta način uporabljena v evropskem projektu SMART.[20] Na splošno torej velja, da se viri, ki imajo urejen prost dostop in pogoje uporabe, tudi uporabljajo za namen računalniške obdelave naravnih jezikov.

Kljub vsemu pa težave z lastništvom avtorskih pravic niso edini razlog za pičlo bero računalniško berljivih slovenskih leksikalnih virov. Leksikalni podatki v slovarjih za končne »človeške« uporabnike, tako v knjižni kot digitalni obliki, niso neposredno uporabni za namene računalniškega procesiranja naravnih jezikov. Predelava teh podatkov je zahtevna, njihova narava drugačna in zato se pogosto dogaja, da so strojno berljivi leksikalni viri tudi pri drugih jezikih narejeni povsem na novo. Tipično je, da leksikalni viri nastajajo na podlagi izkušenj pri drugih jezikih ali se celo lokalizirajo v okviru enakega sistema. Izjemno razširjen primerek semantične leksikalne baze, ki je presegel meje omejenosti na angleščino, je ameriška baza *WordNet*, ki se je sčasoma razširila v mednarodno mrežo (*Global*

[18] Gl. <http://langtech.jrc.it/DGT-TM.html>, <http://evrokorpus.gov.si/>, <http://dcl.bas.bg/ssbc/home.html>.

[19] Ta način licenciranja je sicer znan tudi pod morda rahlo humornim angleškim poimenovanjem *copyleft*.

[20] Gl. <http://www.smart-project.eu/>.

WordNet, *EuroWordNet*, *MultiWordNet*, *BalkaNet*), *WordNet* pa od leta 2007 obstaja tudi za slovenski jezik (Fišer in Erjavec 2006). Pri bazi *WordNet* se za razdvoumljanje pomena uporabljajo serije sinonimov, t. i. sinseti, ki so organizirani v ontologijo z odnosi, kot je hiponimija, hipernimija, meronimija itd. Podobne tendence razširjanja na druge jezike kaže tudi projekt *FrameNet* (<http://framenet.icsi.berkeley.edu/>) s širitvijo na španščino, italijanščino, nemščino, japonščino itd., nenazadnje tudi na slovenščino.[21] Če se torej izkazuje, da stava na ontološko organizirane semantične informacije preko prevodnih ustreznic prinaša rezultate, je problem bistveno večji pri skladenjskih leksikalnih bazah tipa *VerbNet*, *Vallex*, *STO* itd.,[22] ki beležijo tipične skladenjsko-semantične ali vezljivostne informacije, organizirane v strojno berljivi obliki. Skladenjske informacije so bolj jezikovno specifične in zato težje prenosljive z jezika na jezik.

Za naš namen je bistveno poudariti naslednje: leksikalne baze, bodisi skladenjske ali semantične, v knjižni obliki niso nikoli obstajale, kljub vsemu pa lahko vsebujejo podatke, ki jih obstoječi viri imajo, npr. slovarske definicije ob posameznih elementih ontološke mreže, prevodne ustreznice v večjezičnih ontologijah in podobno. Bistvena značilnost teh baz je, da so prilagojene strojni berljivosti podatkov in da vsebujejo zavestno predelane leksikalne informacije jezikoslovne narave. Za dosedanje dogajanje s slovenskimi leksikalnimi viri v digitalni obliki je značilno, da so predhodno obstoječi digitalizirani leksikalni viri redko uporabljeni za namen izdelave strojno berljivih leksikalnih baz, uporabnih za računalniško procesiranje, nove leksikalne baze pa nastajajo večinoma z neposrednim prenosom ali lokalizacijo obstoječih baz za druge jezike, tipično za angleščino.

4 Bodoči leksikalni viri za slovenščino

Za vzpostavitev »sodobne jezikovnotehnološke infrastrukture«, če si sposodimo formulacijo iz estonskega nacionalnega programa, bi bilo v smislu načrtovanja gradnje ali sestavljanja jezikovnih virov za slovenščino torej potrebno predvsem zagotoviti možnost splošne in čim bolj razširjene uporabe že obstoječih jezikovnih virov oziroma njihovih delov, kar pomeni predvsem, da bi bili ti digitalizirani na standarden in poenoten način (po vsej verjetnosti večinoma v formatu XML, v standardni strukturirani obliki, npr. LMF),[23] in da bi bila v zvezi z avtorskimi pravicami njihova raba za različne namene urejena na način, ki bi zagotovil čim boljši izkoristek virov glede na možnosti dogovora z njihovimi lastniki. Taka ureditev verjetno predpostavlja predvsem spodbujevalno intervencijo zainteresiranih institucij, ki bi zagotovila prost dostop do virov za raziskovalne namene ter možnost uporabe virov za druge namene preko povezovalne institucije,

[21] Še en ontološko organiziran sistem, namenjen predvsem razvoju tehnologij za t. i. *Semantic Web* ali *Web 3.0*, ki kaže tendence po prenosu na druge jezike, med drugim tudi slovenščino, je projekt *Cyc* (<http://www.cyc.com/>) s svojo evropsko razvojno enoto v Ljubljani.

[22] Gl. <http://verbs.colorado.edu/verb-index/>, <http://ufal.mff.cuni.cz/vallex/>, <http://english.cst.ku.dk/sto_ordbase/>.

[23] <http://www.lexicalmarkupframework.org/>.

ki bi ponujala jezikovne vire pod dogovorjenimi pogoji in nadzorovala njihovo uporabo.

Po drugi strani bi bilo smiselno spodbuditi razvoj novih tipov jezikovnih virov in hkrati zagotoviti njihovo kompatibilnost, kar pomeni, da bi bila namerno spodbujana možnost uporabe podatkov iz enega vira, ki imajo podobno naravo, tudi v drugih, vsebinsko drugačnih virih. Zahteva po kompatibilnosti verjetno ne bi smela biti končna in absolutna, ker prisilno (vsebinsko) poenotenje lahko zaduši raznoličnost pristopov in potencialno ubija raziskovalno radovednost. Kljub temu pa sedanje stanje virov kaže, da bi predvsem zaradi omejenih možnosti financiranja celotna slovenska jezikoslovna skupnost precej pridobila, če bi bila kompatibilnost in povezljivost virov na višji ravni kot sedaj.

V diagramu v **Dodatku 2** predstavljamo enega od možnih razmislekov o morebitnih bodočih jezikovnih virih za slovenščino in njihovi povezanosti. Vsi omenjeni viri so v prvi vrsti računalniški (spletni) viri in torej v elektronski obliki, njihova razvrstitev pa poteka po štirih vertikalnih linijah, ki jih vidimo kot razmeroma zaključene in povezane z enim od obsežnejših področij srečevanja splošne ali strokovne javnosti z jezikovnimi vprašanji. Prva od teh je *jezikovni opis*, ki zajema tako raziskovalno kot uporabno predelovanje podatkov o dejanski rabi sodobnega jezika, ki jih najdemo v različnih vrstah besedilnih korpusov, v jezikovne priročnike za splošno ali strokovno rabo. Druga vertikalna linija, poimenovana *standardizacija*, zajema normativnejši del ukvarjanja z jezikom, kjer je močnejši poudarek na jezikovnem standardu in težavah pri učenju ali usvajanju slovenskega jezika oziroma pri vseh vprašanjih, kjer se govorci znajdejo v zadregi glede trenutnega konsenza o standardni varianti slovenščine. Tretja linija – *terminologija* – je povezana z reševanjem jezikovnih vprašanj na specializiranih področjih, ki so manj stvar obče, nediferencirane jezikovne skupnosti, kot njenih segmentov, ki se ukvarjajo z enim od specializiranih področij védenja, na katerih potrebujejo nadzorovano urejanje poimenovalnih možnosti. Zadnja linija, imenovana *večjezičnost*, zajema srečanja slovenščine s tujimi jeziki in – gledano z uporabniškega stališča – predvsem pomoč pri usvajanju tujih jezikov in zadreg, ki izhajajo iz medjezičnih kontrastivnih razlik.

1. Jezikovni opis

Leksikalna baza: baze v strukturirani elektronski obliki s povezanimi oblikoslovnimi, skladenjskimi in semantičnimi informacijami so bile ena od prioritet jezikovnotehnoloških projektov v devetdesetih letih, omenjenih višje v prispevku. Baze so lahko namenjene najrazličnejšim končnim uporabnikom, kombinirani rabi za namen jezikoslovnih raziskav ali (leksikografskih) opisov jezika, ali primarno za strojno procesiranje naravnih jezikov. Vsebujejo informacije jezikoslovne narave (oblikoslovne, skladenjske, semantične), od njihovega končnega namena pa je v precejšnji meri odvisna organiziranost in narava teh podatkov. Če se omejimo na nekaj tujejezičnih primerov, bi kot zanimive lahko izpostavili denimo *Base lexicale*

du français (BLF),[24] ki je organizirana okrog elektronskega slovarja DAFLES (Selva idr. 2002) in namenjena predvsem učenju francoskega jezika. Primer kombinirane baze, namenjene leksikografom in za strojno procesiranje, je denimo nizozemska *Referentiebestand Nederlands (RBN)*,[25] zanimive so tudi sintaktično naravnane španske baze Grial[26] in ADESSE,[27] danski STO[28] itd. Za naš namen je bistveno poudariti, da bi bili podatki iz slovenske leksikalne baze, nakazane v naši shemi, uporabni za sestavljanje različnih vrst jezikovnih priročnikov, kot so enojezični slovarji za materne govorce slovenščine, pedagoški slovarji za šolsko rabo ali za učenje slovenščine kot tujega jezika, slovarji sinonimov in različne vrste slovnic ter dvojezične slovarske baze in slovarji, kar kažejo tudi puščične povezave na diagramu.

Splošni enojezični slovar, *enojezični pedagoški slovar/slovar za tujce*, *slovar sinonimov*: ko kot del jezikovnega portala omenjamo navedene slovarje, ki so v knjižni obliki znani že dolgo časa, s tem merimo seveda na njihove elektronske različice, ki so lahko po videzu dokaj »neslovarske«, a jih od leksikalnih baz loči predvsem izrazita naravnanost k znanemu končnemu uporabniku in njegovim potrebam. Te potrebe so lahko zelo različne – šolarji bodo potrebovali precej drugačen priročnik pri učenju tujega ali maternega jezika, odrasli pri spraševanju o pomenu nerazumljivega izraza, pisci pri iskanju alternativne ubeseditve (slovar sinonimov ali tezaver) itd.[29] Primeri takšnih spletnih slovarjev so znani, od »očeta vseh slovarjev«, *Oxford English Dictionary*, do prosto dostopnih slovarjev za učenje angleščine, poleg slovarjev za različne jezike, ki so se domala vsi v zadnjem času preselili na splet, bodisi v plačljivi ali prosto dostopni različici.[30] V kontekstu enotnega portala slovenskih jezikovnih virov se zdi potrebno poudariti, da za slovenščino obstaja le en enojezični slovar in da bi povezljivost različnih jezikovnih virov na spletnem portalu omogočila hitrejše sestavljanje, predvsem

[24] <http://ilt.kuleuven.be/blf/>.

[25] <http://tst.inl.nl/producten/rbn/>.

[26] <http://grial.uab.es/>.

[27] <http://adesse.uvigo.es/>.

[28] <http://cst.ku.dk/sto_ordbase/>.

[29] Zanimivo je, da se smiselna organizacija opisa pomenskih odtenkov nekega izraza v splošni rabi do določene mere upira trendu, ki je v dobi interneta zelo razširjen in je dokazal, da altruistični prispevek energije in časa zaradi skupne in splošne dobrobiti lahko prevlada nad družbeno spodbujanim egoizmom: če je uspeh prosto dostopne in decentralizirane spletne enciklopedije *Wikipedija* z Microsoftovo odločitvijo o ukinitvi Encarte dokazal, da je sestavljanje enciklopedičnih del v zaprtih uredniških skupinah do neke mere stvar preteklosti, ima *Wictionary* kot slovarski antipod *Wikipedije* drugačno usodo in je bistveno manj razširjen, čeprav ima enake tehnične možnosti in široko spletno skupnost, ki tehnično obvladuje to orodje. Razširjenost klasičnih enojezičnih slovarjev na spletu predvsem pri splošnem besedišču nakazuje, da je za predstavljanje pomenskih odtenkov splošnega besedišča morda potreben nek profesionalni presežek, prek praga katerega povprečni konzument neomejenih spletnih možnosti težko stopi.

[30] <http://www.oed.com/>, <http://www.oxfordreference.com/>, <http://www.ldoceonline.com/>, <http://www.macmillandictionary.com/>, <http://www.merriam-webster.com/>, <http://www.dwds.de/>, <http://g3.spraakdata.gu.se/saob/>, <http://gtb.inl.nl/, http://buscon.rae.es/draeI/> itd.

pa hitrejši dostop[31] do tako potrebnih slovarjev, kot so pedagoški slovar, slovar za tujce, sploh pa že dolgo pogrešani slovar sinonimov.

Slovnični opis: obravnava slovnice je mišljena v najširšem smislu kot zbirka različnih pristopov k opisovanju morda predvsem skladenjske ravni jezikovnega opisa, v neizogibni kombinaciji z drugimi ravnmi. Kot nekakšen zgled za ta del spletnega portala bi lahko vzeli češki projekt *Študije o sodobni češki slovnici*, ki ga izvajajo na Inštitutu za češki jezik v Pragi,[32] druge možne smeri raziskovanja, če omenimo samo nekatere, zajemajo denimo še *Head-Driven Phrase Structure Grammar* (HPSG), *Lexical Functional Grammar* (LFG), Langackerjevo *Cognitive (Construction) Grammar* in mnoge druge.[33] Ta del jezikovnega portala bi bil izraziteje namenjen ožji strokovni publiki.

2. Standardizacija jezika

Pravopisna pravila, *leksikon besednih oblik in spletni servis:* enega od uspelejših srečanj računalniške analize jezika in končnih uporabnikov jezikovnih priročnikov predstavljajo podatki o pregibanju besed oziroma t. i. leksikoni besednih oblik. Predvsem pri slovanskih jezikih z množico pregibnih oblik se uporabniki nemalokrat znajdejo v zadregi glede standardne oblike, te podatke pa potrebujejo tudi sistemi za avtomatsko označevanje in razčlenjevanje besedil. Podobno funkcijo so v predračunalniški dobi opravljali slovarski deli pravopisov in iz obeh so se razvile spletne aplikacije, ki ponujajo tovrstne podatke za končnega uporabnika, temu so navadno dodane tudi druge pravopisne informacije in v nekaterih primerih tudi spletni servisi za pomoč uporabnikom pri jezikovnih vprašanjih. Tipična in s slovenščino primerljiva primera sta denimo servisa za češki in slovaški jezik, podobno za nizozemščino, francoščino itd.[34]

Pedagoška slovnica: slovnice, namenjene učenju jezika, so pravzaprav zelo primeren priročnik za splet. Hitra navigacija s hiperpovezavami omogoča dobro povezljivost vsebin in interaktivnost (vaje, kvizi itd.). Za večje jezike so takšne slovnice prosto dostopne, precej jezikov pa ima tudi pedagoški slovnični opis v angleščini.[35] Predpostavljamo, da bi bila spletna pedagoška slovnica predvsem pripomoček pri učenju slovenščine v zadnji triadi osnovne šole, v srednjih šolah in deloma na jezikoslovnih fakultetnih programih.

---

[31] Hitrejši dostop pomeni, da lahko uporabniki na spletu sproti dostopajo do delov slovarske baze, ki je v procesu kompiliranja – takšen pristop so denimo uspešno uveljavili pri *Oxford English Dictionary Online*.

[32] <http://www.ujc.cas.cz/>, <http://mam.ujc.cas.cz/>.

[33] <http://hpsg.stanford.edu/>, <http://www.essex.ac.uk/linguistics/LFG/>.

[34] <http://prirucka.ujc.cas.cz/>, <http://slovnik.juls.savba.sk/>, <http://gtb.inl.nl/>, <https://ilt.kuleuven.be/blf/>, <http://www.la-conjugaison.fr/>.

[35] <http://hypermedia.ids-mannheim.de/grammis/>, <http://hypermedia.ids-mannheim.de/programm/>, <http://grammar.ccc.commnet.edu/grammar/>, <http://fr.tsedryk.ca/>, <http://grammaire.reverso.net/>, <http://www.dutchgrammar.com/>.

3. Terminologija

Terminološki portal: terminološke zbirke so po svoji naravi izjemno primerne za spletno okolje, ker terminologija v osnovi teži k enopomenskosti pri poimenovanju konceptov, obseg besedišča posameznih področij je navadno dokaj velik, hkrati pa pri iskanju pomena ali prevoda posameznega termina ni pomembno, ali se nahaja v zbirki s petimi ali dvajset tisoč iztočnicami. To se je hitro izkazalo tudi v praksi, ker so se na spletu zelo hitro pojavile večje ali manjše terminološke zbirke, ki so za slovenščino denimo zbrane v imeniku <http://evroterm.gov.si/slovar/slovar.html>. V okviru predloga spletnega portala bi razmišljali predvsem o terminološkem portalu, ki bi obsegal prosto dostopni uredniški sistem za poljubno število slovarjev z možnostjo uvoza in izvoza podatkov v standardnih formatih ter dodatnimi možnostmi luščenja terminov iz specializiranih korpusov in podobno. Podobni sistemi obstajajo denimo za slovaščino, litvanščino, estonščino itd.[36]

4. Večjezičnost

Dvojezične ali večjezične slovarske baze: kot smo ugotovili že v uvodnem delu, pri dvojezičnih virih za slovenski jezik vlada hudo pomanjkanje prosto dostopnih virov, predvsem na področju splošnega, najbolj razširjenega besedišča. S kvalitetnimi prosto dostopnimi elektronskimi enojezičnimi viri (leksikalna baza, različni slovarji itd.) bi bili vzpostavljeni tudi pogoji za izdelavo prav tako prosto dostopnih dvojezičnih slovarskih baz, pri katerih bi bilo podatke iz enega, npr. slovensko-angleškega slovarja, mogoče uporabiti tudi za druge slovarje in na ta način (lahko tudi s pomočjo podatkov iz lokaliziranih baz *WordNet*, *FrameNet* ali *Cyc*),[37] ustvariti prosto dostopni večjezični del portala s podatki, ki bi jih bilo mogoče izvoziti s spleta in uporabiti za raziskovalne, pedagoške ali druge namene.

5 Zaključek

Predlog načrta spletnega portala z jezikovnimi viri za slovenščino je namenjen predvsem razmisleku strokovne javnosti, morda tudi akterjev v okviru institucij, ki usmerjajo razvoj in odločajo o financiranju dejavnosti, povezanih s slovenščino in njenimi jezikovnimi viri v elektronski obliki. Predlog bi moral biti v končni fazi mnogo bolj domišljen in razdelan, menimo pa, da dokumenti, kot je denimo *Resolucija o nacionalnem programu za jezikovno politiko 2007–2011*,[38] ali nenazadnje publikacije, ki se eksplicitno ukvarjajo z bodočim načrtovanjem leksikalnih virov, kot je *Strokovni posvet o novem slovarju slovenskega jezika* (Perdih 2009) ali *Novejša slovenska leksika (v povezavi s spletnimi jezikovnimi viri)* (Gložančev idr. 2009), ne upoštevajo dovolj, do kako radikalnih sprememb

[36] <http://www.termnet.lv/>, <https://data.juls.savba.sk/std/>, <http://www.eurotermbank.com/>.

[37] <http://lojze.lugos.si/~darja/slownet.html>, <http://framenet.icsi.berkeley.edu/>, <http://www.cyc.com/>.

[38] <http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/Podrocja/slovenski_jezik/RESOLUCIJA_.pdf>.

pri gradnji in uporabi jezikovnih virov je prišlo v zadnjih dvajsetih letih in kako radikalno drugačne so nove možnosti uporabe računalniških tehnologij tako pri sestavljanju teh virov kot pri njihovi »vizualizaciji« na spletu ali v drugih računalniških okoljih, sploh ob izrabi možnosti drugih delov jezikovnih tehnologij, ki so naštete na začetku prispevka. Kratek in nepopoln pregled virov za druge, predvsem finančno izdatneje podprte jezike, kaže, kako daleč je slovenščina na tem področju celo od estonščine, desetletja dela pa jo ločijo od skandinavskih jezikov, nizozemščine in drugih jezikov, ki nenazadnje nimajo nepredstavljivo več govorcev od slovenščine. Menimo, da bi k izboljšanju stanja lahko precej pripomoglo domišljeno načrtovanje in osrednje institucionalno telo z možnostjo realnega dodeljevanja finančnih sredstev.

Literatura

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

Boguraev, Bran (ur.), in Briscoe, Ted (ur.), 1989: *Computational Lexicography for Natural Language Processing*. Longman.

Erjavec, Tomaž, 2006: The English-Slovene ACQUIS corpus. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'06*. Pariz: ELRA.

Erjavec, Tomaž, in Sarossy, Bence: Morphosyntactic Tagging of Slovene Legal Language. *Informatica 30*. 483–488.

Erjavec, Tomaž, Ignat, Camelia, Pouliquen, Bruno, in Steinberger, Ralf, 2005: Massive multi-lingual corpus compilation: Acquis Communautaire and totale. *Proceedings of the 2nd Language & Technology Conference, April 21-23, 2005*. Poznan. 32–36.

Erjavec, Tomaž, in Fišer, Darja, 2006: Building Slovene WordNet. *Proceedings of the 5th International Conference on Language Resources and Evaluation LREC'06*. Genoa.

Fišer, Darja, 2007: A multilingual approach to building Slovene Wordnet. *Proceedings of the workshop on A Common Natural Language Processing Paradigm for Balkan Languages held within the Recent Advances in Natural Language Processing Conference RANLP'07, 26 September 2007*. Borovets.

Fišer, Darja, 2007: Leveraging Parallel Corpora and Existing Wordnets for Automatic Construction of the Slovene Wordnet. *Proceedings of the 3rd Language and Technology Conference L&TC'07, 5-7 October 2007*. Poznan.

Fišer, Darja, 2008: Using Multilingual Resources for Building SloWNet Faster. *Proceedings of the 4th International WordNet Conference GWC'08, 22-25th January 2008*. Szeged.

Fontenelle, Thierry, 1997: *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Max Niemeyer Verlag.

Gložančev, Alenka, 2009: *Novejša slovenska leksika (v povezavi s spletnimi jezikovnimi viri)*. Ljubljana: Založba ZRC, ZRC SAZU.

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

Gorjanc, Vojko, 2006: Korpusno jezikoslovje in leksikalni opisi slovenskega jezika. *Slovensko jezikoslovje danes*. Ljubljana: Slavistično društvo Slovenije. 146–148.

Grad, Anton, Škerlj, Ružena, in Vitorovič, Nada 1978: *Veliki angleško-slovenski slovar*. Ljubljana: Državna založba Slovenije.

Hajnšek-Holz, Milena, in Jakopin, Primož, 1999: *Odzadnji slovar slovenskega jezika po Slovarju slovenskega knjižnega jezika*. Ljubljana: Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti.

Jakopin, Primož, in Bizjak Končar, Aleksandra, 2008: Part-of-speech tagging of Slovenian, 12 years after. *Zbornik Šeste konference Jezikovne tehnologije, 16. do 17. oktober 2008.* Ljubljana: Institut Jožef Stefan. 104–109.

Jakopin, Primož, in Michelizza, Mija, 2007/2008: Besedilni korpus Nova beseda. *Mostovi* 41/1–2. 165–176.

Kocjančič, Polonca, Krek, Simon, in Šorli, Mojca: The Funny Mirror of Language: the Process of Reversing the English-Slovenian Dictionary to Build the Framework for Compiling the New Slovenian-English Dictionary. *Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008)*. Barcelona: Documenta Universitaria: Institut universitari de lingüística aplicada, Universitat Pompeu Fabra. 202–203.

Krek, Simon (ur.), 2005–2006: *Veliki angleško-slovenski slovar Oxford*. Ljubljana: DZS.

Logar, Nataša, in Vintar, Špela, 2008: Korpusni pristop k izdelavi terminoloških slovarjev: od besednih seznamov in konkordanc do samodejnega luščenja izrazja. *Jezik in slovstvo* 53/5. [3] –17.

Perdih, Andrej (ur.), 2009: *Strokovni posvet o novem slovarju slovenskega jezika, 23. in 24. oktober 2008*. Ljubljana: Založba ZRC, ZRC SAZU.

Selva, Thierry, Verlinde, Serge, in Binon, Jean, 2002: Le DAFLES, un nouveau dictionnaire électronique pour apprenants du français. Braasch, Anna, in Povlsen, Claus (ur.): *Proceedings of the 10th EURALEX International Congress, EURALEX 2002*. Copenhagen: Center for Sprogteknologi I. 199–208.

Steinberger, Ralf, Pouliquen, Bruno, Widiger, Anna, Ignat, Camelia, Erjavec, Tomaž, Tufis, Dan, in Varga, Daniel, 2006: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'06*. Pariz: ELRA.

Vintar, Špela, 2008. *Terminologija: terminološka veda in računalniško podprta terminografija*. Ljubljana: Znanstvena založba Filozofske fakultete.

Žganec Gros, Jerneja, Cvetko-Orešnik, Varja, in Jakopin, Primož, 2006: SI-PRON: a comprehensive pronunciation lexicon for Slovenian. *Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 44–49.

Dodatek 1: *Načrt razvoja jezikovnih virov za estonščino (Program NPELT).*[39]

leto	načrt razvoja jezikovnih virov za estonščino	načrt razvoja jezikovnih virov za slovenščino	obstoj	dostopnost
2010				
	baza podatkov za avdiovizualno sintezo govora	–	ne	–
2009				
	odvisnostna drevesnica (100.000 besed)	v okviru projekta *Jezikovni viri za slovenščino* nastaja odvisnostna drevesnica s 100.000 besedami, izdelana bo do konca leta 2009	ne	–
2008				
	tezaver oz. slovar sinonimov	–	ne	–
	dialoški korpus z 1 milijonom besed	–	ne	–
2007				
	estonsko-angleška slovarska baza	Evroterm bi pogojno lahko razumeli kot zametek takšne bazo – ker so v njem samo zakonodajna besedila, je besedišče relativno specifično	da/ne	–
	leksikosemantična baza podatkov	slovenski Wordnet kot ena od možnosti takšne baze obstaja v relativno omejenem obsegu (20.000 literalov)	da/ne	–
	popolna transkripcija govornega korpusa (0,1 milijona besed)	govorni korpus *Broadcast News*	da	plačljiv (6.000 EUR nekomercialno)
2006				
	odvisnostna drevesnica (50.000 besed)	korpus s 30.000 besedami, označen po sistemu praške odvisnostne drevesnice: <http://nl.ijs.si/sdt/>.	da	prosto (nekomercialno)
	leksikogramatična baza podatkov	leksikogramatična baza podatkov tipa VerbNet, VALLEX, SIMPLE, PAROLE, CLIPS, ADESSE itd. ne obstaja	ne	–
	osnovna transkripcija govornega korpusa (0,1 milijona besed)	govorni korpus *Broadcast News*	da	plačljiv (6.000 EUR nekomercialno)
	dialoški korpus (500.000 besed)	–	ne	–

[39] <http://www.keeletehnoloogia.ee/projects/npelt-projects-in-2008?set_language=en>.

	govorni korpus (1 milijon besed)	govorni korpus *Broadcast News* s transkribiranimi 268.000 besedami je dostopen preko agencije ELRA	ne	plačljiv (6.000 EUR nekomercialno)
2005				
	paralelni korpus (angleško-estonski, vsak jezik 10 milijonov besed)	desetmilijonski korpus prevodov zakonodajnih besedil Evrokorpus: <http://evrokorpus.gov.si/>.	da	prosto (spletni dostop)
	dialoški korpus (100.000 besed)	–	ne	–
	površinsko skladenjsko označen korpus (50.000 besed)	korpusa, označenega na besednozvezni ravni (*chunking*), za slovenščino še ni	ne	–
2004				
	pisni korpus (80 milijonov besed)	dva večja pisna korpusa: *FidaPLUS* (<http://www.fidaplus.net>), 620 milijonov besed; *Nova beseda* (<http://bos.zrc-sazu.si/s_beseda.html>), 240 milijonov besed	da	prosto (nekomercialno, spletni dostop)
	semantična baza podatkov (estonski *Wordnet* s 15.000 pomeni)	slovenski *Wordnet* s 17.000 sinseti in 20.000 literali: <http://lojze.lugos.si/~darja/slownet.html>.	da	prosto (nekomercialno)
	razdvoumljeni korpus z besednimi pomeni	–	ne	–
	estonsko-angleški vzporedni korpus (2 milijona besed)	več manjših korpusov vzporednih besedil na naslovu: <http://nl2.ijs.si/index-bi.html>.	da	prosto (nekomercialno)
	baza podatkov tipa BABEL	baza je primerljiva in »vsebovana« v bazi *Speech Dat(II)*	-	-
	baza podatkov tipa *Speech Dat*	baza podatkov *Speech Dat(II)* za slovenščino je dostopna preko evropske agencije za jezikovne vire (ELRA), vsebuje posnetke 1.000 govorcev	da	plačljivo (14.000 EUR)
	slovarji v elektronski obliki (rusko-estonski, finsko-estonski, angleško-estonski itd.)	tiskani slovarji v elektronski obliki so večinoma v lasti založb, največja prosto dostopna dvojezična terminoška podatkovna zbirka je Evroterm (<http://evroterm.gov.si/>)	da	plačljivo (Evroterm prosto dostopen)

Dodatek 2: *Jezikovni viri za slovenščino – predlog*