KORPUS STAREJŠE GAJIČNE SLOVENŠČINE

Tomaž Erjavec, Katja Zupan
Institut »Jožef Stefan«, Mednarodna podiplomska šola Jožefa Stefana, Ljubljana

UDK 811.163.6'373.44"1845/1918"

V prispevku predstavimo nabor jezikovnih virov iz obdobja 1845–1918, ki predstavlja gajični del virov starejše slovenščine IMP in vsebuje tri korpuse (ročno, delno ročno in avtomatsko označenega) ter besedje, izdelano iz ročno označenih besed pojavnic v korpusih. Korpus izbranega obdobja je zanimiv za obogatitev SSKJ pa tudi zato, ker se na njegovem začetku – kljub uporabi istega črkopisa kot danes – zapis besed še vedno močno razlikuje od sodobne norme, ob njegovem koncu pa te razlike že skorajda izginejo.

starejša slovenščina, gajica, korpus, besedje, posodabljanje besed

The paper presents a collection of language resources from the period 1845–1918 that consists of the IMP resources for historical Slovene written in the Gaj alphabet. These resources contain three corpora (annotated by hand, semi-automatically and automatically) and a lexicon built from the hand-annotated words in the corpora. The corpora from the chosen period are interesting as a source of enrichment of the Dictionary of Standard Slovene, and for the study of differences in orthography from today's standard: at the beginning of the observed period the spelling was still quite different from the one used today but by its end the differences in spelling had practically disappeared.

historical Slovene, Gaj alphabet, corpus, vocabulary, word modernisation

1 Uvod

Jezikovni viri IMP (Erjavec 2012)[1] vsebujejo večjo zbirko besedil (digitalno knjižnico), jezikoslovno označene korpuse in besedje slovenskega jezika med letoma 1584 in 1918. Zbirka je bila prvotno zgrajena za razvoj jezikovnih tehnologij starejšega slovenskega jezika, kot podatkovna baza za optično razpoznavanje znakov in iskanje po besedilu digitalnih knjižnic. Za uporabno se je izkazala tudi v digitalni humanistiki, kjer omogoča učiteljem, učencem, študentom, jezikoslovcem in drugim uporabnikom spletni dostop in raziskovanje besedja naše pisne kulturne dediščine (Erjavec, Fišer 2014).

Za novi slovar slovenskega jezika sta bila do zdaj izdelana dva koncepta (Krek idr. 2013; Gliha Komac idr. 2015), ki jima je skupno, da sta osredotočena na sodobni jezik in da v središče slovarske podstati postavljata jezikovne korpuse, ne samo kot vir geslovnika in pomagalo slovaropiscem, temveč tudi s predvidevanjem, da bodo imeli geselski članki spletnega slovarja neposredne povezave do korpusov. Pri tem se

1 http://nl.ijs.si/imp

postavi vprašanje, kako naj uporabniki dostopajo do starejše leksike, ki je sodobnemu bralcu lahko še vedno zanimiva, predvsem ker jo najdemo tudi v slovenski leposlovni klasiki. Ta se pospešeno seli na splet (Hladnik 2009), pa tudi v splošnem postajajo »po zaslugi računalniške tehnologije in svetovnega spleta tudi slovenska besedila, nastala v preteklih stoletjih, vse lažje dostopna«, potrebi po njihovem razumevanju pa bodo lahko zadostili »široko obvestilni slovarski priročniki« (Merše 2009: 255).

Za starejšo leksiko Krek idr. (2013: 27) predvidevajo, da je »podatke o starinskih oblikah mogoče pridobiti avtomatsko že na ravni iztočnice, kjer črpamo iz korpusa in Slovarja starejšega slovenskega besedja«, s čimer sta tu mišljena korpus in besedje IMP. Po drugi strani Gliha Komac idr. (2015: 5) predlagajo, da bodo »[b]esede, besedne zveze in pomeni, ki jih sodobni viri ne vključujejo […], na voljo v starejših splošnih in specializiranih slovarskih priročnikih ter drugih jezikovnih virih in priročnikih«. Pomanjkljivost tega predloga je, da dostop do vse starejše leksike prepušča tehnično vedno bolj zastarelim obstoječim slovarjem: medtem ko naj bi novi slovar omogočal povezave s primeri iz korpusov, bo starejše besedje dostopno samo npr. v SSKJ, kjer ni nobenih korpusnih povezav ali navedb virov, s čimer bo vsa nekoliko starejša leksika bistveno deprivilegirana glede na sodobno. Vendar ima tudi predlog Kreka idr. (2013) pomanjkljivost, ki jo izpostavlja Ahačič (2014), in sicer da za starejša besedila viri IMP niso dovolj reprezentativni, zaradi česar dajejo izkrivljeno podobo slovenščine tistega časa. Kritika pa se nanaša na vire izpred 1850, saj je v IMP-u samo 43 starejših besedil, mlajših pa kar 615. Gajični del zbirke IMP je tako razmeroma velik in vsebuje več kot 13 milijonov besed, ob tem pa je tudi raznovrsten, saj vsebuje (Erjavec idr. 2011):

– stvarna, nabožna in leposlovna dela, ki so bila prevedena iz nemščine ter korigirana in obdelana v projektu AHLib (Prunč 2007);
– izvirna slovenska, večinoma leposlovna besedila, zajeta iz Wikivira slovenske leposlovne klasike[2] ali obdelana v NUK-u v okviru projekta EU IMPACT;
– izbrane številke *Kmetijskih in rokodelskih novic*.

V prispevku se posvetimo gajičnemu, torej najmlajšemu delu zbirke IMP. Ta del z začetkom v 1845 je dovolj nov, da je njegovo besedje še vedno aktualno tudi za nejezikoslovce. Zanimiv je tudi jezikovnotehnološko, saj je bilo, glede na besedila v bohoričici, v tem obdobju natisnjenih mnogo več slovenskih besedil, ki so večinoma tudi že digitizirana, zato imajo orodja, razvita za obdelavo takega jezika, bistveno širšo uporabnost.

2 Gajični del korpusov IMP

Osnova korpusov in besedja IMP je zbirka besedil, iz katere je bila izdelana spletno dostopna digitalna knjižnica IMP. Zbirka vsebuje faksimile posameznih del in njihove pregledane prepise. Iz te zbirke so bili zgrajeni trije korpusi, ki se razlikujejo

2 http://sl.wikisource.org/wiki/Wikivir:Slovenska_leposlovna_klasika

po namenu, deležu ročne obdelave in velikosti. V tem razdelku jih predstavimo in podamo kvantitativno analizo podkorpusov njihovih gajičnih delov.

2.1 Trije korpusi

Najmanjši korpus je goo300k, ki je bil narejen z naključnim vzorčenjem strani iz podmnožice besedil v zbirki IMP. Tu je bil zapis vsake besede kot tudi njenih jezikoslovnih oznak pregledan ročno, da bo korpus služil kot referenčni vir starejšega jezika, primeren tudi za urjenje jezikovnotehnoloških orodij. Približno desetkrat večji korpus je foo3M, ki ravno tako vsebuje vzorčene strani iz večjega nabora besedil. Ta korpus je avtomatsko jezikoslovno označen; ročno so pregledane samo oznake besednih oblik, ki se ne pojavijo v goo300k. Namen izdelave korpusa je bil povečanje ročno pregledanega besedja, obenem pa uporaba v obliki testne množice za orodja, razvita s korpusom goo300k. Največji korpus je imp15M, ki vsebuje besedila celotne zbirke IMP, a so ta označena samo avtomatsko.

Vsaka posamezna stran v goo300k in foo3M oz. besedilo v imp15M je označeno glede na to, ali je napisano v bohoričici oz. gajici, zato je iz njih enostavno izluščiti gajične podkorpuse, ki jih imenujemo goo300k-gaj, foo3M-gaj in imp15M-gaj. Ker obstajajo tudi besedila, ki so napisana v obeh črkopisih, bodisi so iz prehodnega obdobja bodisi vsebujejo kar nekaj citatov v bohoričici, smo iz korpusa imp15M-gaj izločili pet besedil.

2.2 Struktura korpusov

Vsi korpusi IMP so izvorno zapisani v formatu XML, skladnem s shemo TEI,[3] kar omogoča formalno preverjanje strukture, dokumentiranost in trajnost zapisa ter razmeroma enostavno pretvorbo v izvedene formate (npr. v vertikalni format za konkordančnik).

Goo300k in foo3M sta strukturirana v posamezna besedila, ki vsebujejo kolofon, faksimile in besedila, sestavljena iz vzorčenih posameznih strani, znotraj teh pa iz generičnih elementov na ravni odstavkov, ki pa so prek atributa klasificirani npr. v naslove in odstavke. Po drugi strani vsebuje imp15M celotna dela, z vsemi izvirnimi oznakami TEI, kot so npr. razdelki, slike, opombe, odstavki, pesmi oz. kitice itn.

Vzorčene strani oz. besedila imajo pripisane bogate metapodatke, poleg črkopisa, signature, letnice, avtorja, naslova v izvirniku in posodobljenega naslova tudi umestitev v taksonomije besedilnih zvrsti, prenosnika in prevodnega statusa. Vsaki besedilni enoti je dodana hiperpovezava na ustrezno stran v digitalni knjižnici oz. neposredno na faksimile strani.

Jezikoslovne oznake vsebujejo oznake za posamezne povedi, ki so bile za vse korpuse določene avtomatsko in zato vsebujejo tudi določeno število napak. V povedih je nato označena vsaka pojavnica (bodisi beseda ali ločilo) kot tudi presledki med pojavnicami.

3 http://www.tei-c.org

Besede razdelimo v sledeče skupine (glej tudi Erjavec 2012):

– *Sodobne besede*, ki se pišejo enako kot danes, pa tudi oblikoskladenjske lastnosti in pomen se jim niso bistveno spremenili.
– *Starinske besede*, ki se jim sicer tudi niso spremenile oblikoskladenjske lastnosti oz. pomen, a so se zapisovale drugače kot danes (npr. *merzel*, *ostanjki*). Posebno težavo predstavljajo besede, ki so se nekoč pisale narazen, zdaj pa skupaj, ali obratno, npr. *naj lepši*; te se vedno obravnavajo kot starinske.
– *Zastarele besede*, ki so se jim, glede na sodobno normo, bodisi spremenile oblikoskladenjske lastnosti (*bandero* → *bandera*), se jim je bistveno spremenil pomen (*edinec* → *posameznik*) ali se danes ne uporabljajo več (*baja* → *dojilja*).

Glede na to razdelitev so besedne pojavnice v korpusih označene z naslednjimi jezikoslovnimi podatki:

– *posodobljena oblika*, ki je za sodobne besede identična pojavnici (razen da je vedno zapisana z malimi črkami), starinskim besedam poda sodobno ustreznico, zastarele besede pa posodobi samo ortografsko, npr. *gnjilec* → *gnilec (vino iz nagnitega grozdja)*;
– *oblikoskladenjska oznaka IMP*, ki določi besedno vrsto besede v kontekstu, vsebuje pa lahko tudi leksikalne oblikoskladenjske lastnosti (ločimo 32 različnih oznak);
– *sodobna lema* oz. osnovna oblika besede, kot izhaja iz posodobljene oblike; v avtomatsko označenem korpusu je lema odvisna tudi od oblikoskladenjske oznake (npr. *hotela* → *hoteti* oz. → *hotel*, glede na to, ali ima pripisano glagolsko ali samostalniško oznako), ni pa pomensko razdvoumljena (npr. *sedel* → *sedeti* oz. → *sesti*), pač pa ji je v takih primerih pripisana ena, lahko tudi napačna lema.
– *razlaga*, ki je dodana samo pri zastarelih besedah in tipično poda eno ali več sodobnih ustreznic lemi besedne oblike.

Slika 1 podaja zapis označenega niza »*bi se žnjim*«, s katerim ilustriramo zapis večbesedne posodobitve izvorne besedne pojavnice kot tudi oznake za leme in oblikoskladenjske lastnosti, ki so privzeto v angleškem jeziku (npr. P = Pronoun), vendar jih je mogoče avtomatsko prevesti v slovenske oznake in razstaviti v posamezne lastnosti (Va = Verb auxiliary = Gp = glagol, pomožni).

```
<w lemma="biti" ana="#Va">bi</w>
<c> </c>
<w lemma="se" ana="#P">se</w>
<c> </c>
<choice>
      <orig>
            <w>žnjim</w>
      </orig>
      <reg>
            <w lemma="z" ana="#S">z</w>
            <c> </c>
            <w lemma="on" ana="#P">njim</w>
      </reg>
</choice>
```

Slika 1: Zapis jezikoslovno označenega korpusa v XML TEI

2.3 Analiza korpusa

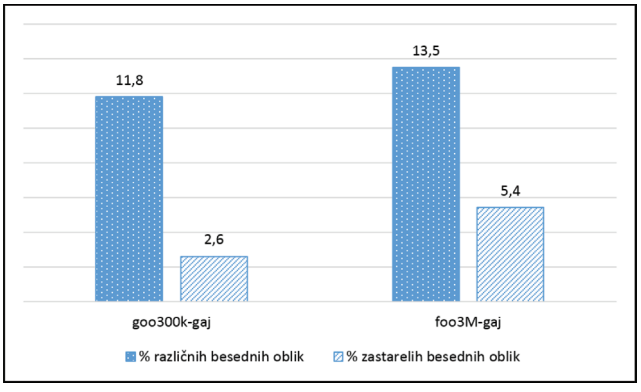
Tabela 1 poda sumarne podatke za gajične korpuse, začenši z obdobjem besedilne produkcije, ki ga zajemajo. Drugi stolpec poda število besedil, ki so vključena v posamezne korpuse (prva dva korpusa vsebujeta samo vzorce izbranih besedil). Število pojavnic poda tako število besed kot ločil v korpusih, naslednji stolpec pa samo besednih pojavnic. Število starinskih besednih pojavnic je število oblik, ki se razlikujejo od posodobljene ustreznice, število zastarelih pa pove, koliko besednim pojavnicam je pripisana razlaga; te bi bile npr. kandidati za vključitev v slovar. V zadnjem stolpcu je število besednih pojavnic, pri katerih so bile oznake ročno preverjene. Ker je korpus imp15M polno avtomatsko označen, je to število zanj 0, ravno tako pa nima pripisanih razlag za zastarele besede.

Tabela 1: Velikost korpusov in število zvrsti pojavnic

Korpus	Od	Do	Besedil	Pojavnic	Besed	Starinskih	Zastarelih	Preverjenih
goo300k	1845	1899	70	256.410	209.468	76.643	5.491	209.468
foo3M	1845	1899	294	2.785.330	2.289.153	659.044	3.627	67.041
imp15M	1846	1918	619	16.247.831	13.291.721	3.569.433	0	0

Podrobneje nas je zanimal delež starinskih in zastarelih besed v besedilih, torej koliko sta se zapis in leksika v opazovanem obdobju že poenotila ter kako se je ta delež spreminjal skozi čas. Da bi dobili čim bolj realno sliko, smo upoštevali samo ročno pregledane besedne oblike.

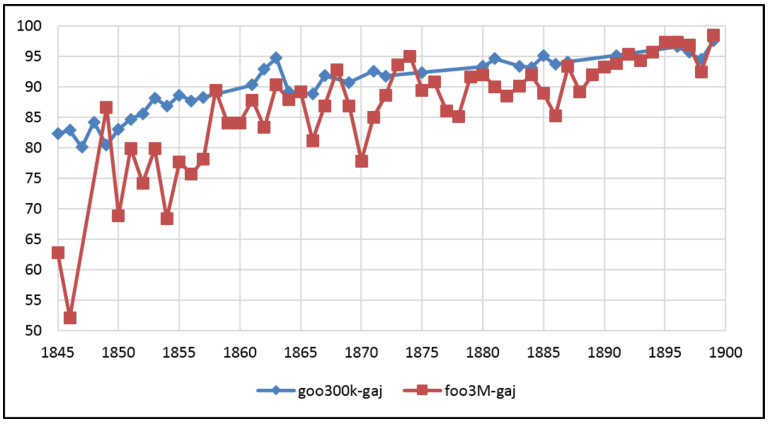
Kot prikazuje Slika 2, je v goo300k-gaj malo manj kot 12 % besednih oblik takšnih, pri katerih se izvirna in posodobljena oblika razlikujeta v zapisu,[4] medtem ko je v foo3M-gaj ta delež nekoliko večji. Delež zastarelih oblik je v obeh korpusih precej manjši od deleža posodobljenih, kot je razvidno iz Slike 2.



Slika 2: Delež posodobljenih in zastarelih besednih oblik (v odstotkih)

4 Razlika samo v naglasu ni bila upoštevana.

V skupni statistiki goo300k-gaj vsebuje 88,2 % besednih oblik, pri katerih se je zapis že poenotil in ustalil s stališča sodobne pisne norme, v ročno pregledanem foo3M-gaj pa je takšnih oblik 86,5 %. Da bi pokazali, kako se je ta delež spreminjal skozi čas, smo pogledali, kako so se razmerja spreminjala po letih, v katerih so besedila nastala. Kot je razvidno iz Slike 3, se delež ortografsko standardiziranih oblik (pričakovano) v obeh korpusih postopno povečuje, kljub določenim odstopanjem. Nihanja so večja v foo3M-gaj, kar je delno mogoče pripisati dejstvu, da je teh besednih oblik v primerjavi z goo300k-gaj precej manj, pomemben pa je tudi podatek, da vsebuje samo oblike, ki jih goo300k ne, tj. manj pogoste besedne oblike oz. njihove zapise.



Slika 3: Spreminjanje deleža besednih oblik s sodobnim zapisom (v odstotkih)

Časovni razpon, ki ga pokrivata podkorpusa, je pomembno obdobje v razvoju sodobne pisne norme. Podkorpusa dokumentirata čas po uveljavitvi sodobnega črkopisa (1845), prek »pomladi narodov« in zahteve po »enakih pravicah [za slovenski jezik], kakor jih ima nemški jezik« (1848), novih oblik (1851), Janežičeve slovnice (1854), Levstikovih *Napak slovenskega pisanja* (1858), Wolf-Cigaletovega slovarja (1860), Cigaletove *Znanstvene terminologije* (1880) do Pleteršnikovega slovarja (1894/95) in Levčevega pravopisa (1898), če omenimo samo nekaj pomembnih prispevkov k utrjevanju pisne knjižnojezikovne norme (Pogorelec 2011). Slika 3 prikazuje ta postopni premik k enotni pravopisni normi, ki v zadnjem letu 19. stoletja doseže že 97,5- (goo) oz. 98,4-odstotno (foo) ujemanje zapisa besednih oblik s sodobnim zapisom.

3 Besedje IMP-gaj

Iz ročno označenih besednih pojavnic v korpusih smo avtomatsko generirali besedje, katerega izvorni namen je olajšati iskanje po digitalnih knjižnicah starejših besedil, kot že rečeno, pa smo ga ponudili tudi na spletu za pregledovanje. Iz zasnove označenega korpusa tudi izvira struktura besedja, ki ima za gesla (leme) samo posamezne besede, z izjemo besed, ki se zdaj pišejo narazen, v besedilih pa so se pisale

skupaj. Geselski članki tako ne vsebujejo večbesednih enot, kot so npr. frazemi ali termini, ravno tako pa nimajo pomenske ločitve, razen ko imata dve gesli enako iztočnico, vendar je eno zastarelo, drugo pa ne. Posamezni geselski sestavek je namreč enolično določen kot njegova lema, njene oblikoskladenjske lastnosti in, za zastarele besede, razlaga v sodobni slovenščini. Geslo ima nato naštete vse sodobne besedne oblike, ki se pojavijo med ročno označenimi besednimi pojavnicami v goo300k(-gaj) in foo3M(-gaj), te pa so nadrejene vsem zgodovinskim besednim oblikam iz korpusa.

Besedje, izluščeno iz gajičnega dela korpusa IMP, obsega 23.000 gesel. Če izločimo slovarsko nezanimive pojavnice, kot so tuje, zatipkane in napačno tokenizirane besede ter arabske in rimske številke, dobimo 20.818 gesel, ob dodatni izločitvi lastnih imen pa 18.556. Med njimi je 16 % besed takšnih, ki so označene kot zastarele. Pri skoraj dveh tretjinah zastarelih besed je bilo razlago za geselski članek mogoče črpati iz SSKJ (1.339 besed) oz. iz Pleteršnikovega slovarja (606 besed), za preostalih 1.008 pa je bilo treba razlago poiskati v drugih slovarskih virih oz. jo pri večini (80,5 %) besed določiti z analizo sobesedila. Razlage zastarelih besed so razmeroma kratke, v večini primerov samo ena ali dve sodobni ustreznici oz. zelo kratka razlaga; pomembno je tudi, da so gesla vedno posamezne besede, zato so razlage samo delno ustrezne, če je geslo del stalne besedne zveze, na kar opozarja tudi Ahačič (2014).

Da zastarelih besed iz te skupine ni ne v SSKJ ne v Pleteršniku, je mogoče pripisati dejstvu, da je veliko besed terminoloških, iz strokovnih priročnikov, nekaj pa je tudi takšnih, ki so značilne samo za določen prevod (pri tuji literaturi) ali določenega avtorja.

4 Zaključek

V prispevku smo predstavili gajični del korpusov in besedja starejše slovenščine IMP, da bi dobili vpogled, kako bi lahko vsebovano besedje pripomoglo k boljši obvestilnosti spletnih slovarjev, ki vsebujejo tudi nekoliko starejše besedje slovenskega jezika. Analiza pokaže, da korpusi in besedje vsebujejo veliko besed, ki so se včasih pisale drugače ali pa so zastarele in so zato večini sodobnih bralcev neznane. Tako je npr. v besedju IMP več kot 1.300 gesel zastarelih besed, ki so že zajete v SSKJ, zato bi bilo slovar, vsaj za enoumne besede, razmeroma enostavno nadgraditi s povezavami v besedje oz. korpus (konkordančnik). V nasprotju z veliko večino slovarskih (pa tudi korpusnih) virov imajo korpusi in besedje IMP to prednost, da so prosto in odprto dostopni po licenci Creative Commons s priznanjem avtorstva, zato ni nikakršnih lastniških ali pravnih ovir za takšno povezovanje ali celo vključevanje.

V prihodnjem delu nameravamo odpravljati preostale napake v trenutnih različicah korpusov IMP(-gaj), hkrati pa povečati ročno označene dele korpusa, predvsem za besedila 1900–1918, ki še niso ročno pregledana. Na tehnični ravni se bomo osredotočili na izboljšanje metod za posodabljanje besed ter se lotili vprašanj besedotvorja in skladnje starejših slovenskih besedil.

Zahvala

Avtorja se zahvaljujeta anonimnim recenzentom za koristne pripombe. Raziskava, opisana v prispevku, je bila opravljena v okviru raziskovalnega programa Tehnologije znanja P2-0103 in programa Mladi raziskovalci, ki ju financira ARRS.

Literatura

AHAČIČ, Kozma, 2014: *Zgodovinski podatki v slovarju sodobne slovenščine. Posvet o novem slovarju slovenskega jezika*. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/11-_Kozma_Ahacic-clanek.pdf

ERJAVEC, Tomaž, 2012: Jezikoslovni viri starejše slovenščine. *Knjižnica* 56/3. 205–221.

ERJAVEC, Tomaž, FIŠER, Darja, 2014: Recepcija virov starejše slovenščine IMP. Alenka Žbogar (ur.): *Recepcija slovenske književnosti*. *Obdobja 33*. Ljubljana: Znanstvena založba Filozofske fakultete. 119–127. http://www.centerslo.net/files/file/simpozij/simp33/Zbornik/ErjavecFiser.pdf

ERJAVEC, Tomaž, JERELE, Ines, KODRIČ, Maša, 2011: Izdelava korpusa starejših slovenskih besedil v okviru projekta IMPACT. Simona Kranjc (ur.): *Meddisciplinarnost v slovenistiki*. *Obdobja 30*. Ljubljana: Znanstvena založba Filozofske fakultete. 121–127. http://www.centerslo.net/files/file/simpozij/simp30/Zbornik/Erjavec_Jerel_Kodric.pdf

GLIHA KOMAC, Nataša, JAKOP, Nataša, JEŽOVNIK, Janoš, KLEMENČIČ, Simona, KRVINA, Domen, LEDINEK, Nina, MIRTIČ, Tanja, PERDIH, Andrej, PETRIC, Špela, SNOJ, Marko, ŽELE, Andreja, 2015: Osnutek koncepta novega slovarja slovenskega knjižnega jezika. http://www.fran.si/novi-sskj

HLADNIK, Miran, 2009: Infrastruktura slovenistične literarne vede. Marko Stabej (ur.): *Infrastruktura slovenščine in slovenistike. Obdobja 28*. Ljubljana: Znanstvena založba Filozofske fakultete. 161–169.

KREK, Simon, KOSEM, Iztok, GANTAR, Polona, 2013: *Predlog za izdelavo Slovarja sodobnega slovenskega jezika*. http://www.sssj.si

MERŠE, Majda, 2009: Slovensko zgodovinsko slovaropisje s konceptualno-razvojnega vidika. Marko Stabej (ur.): *Infrastruktura slovenščine in slovenistike*. *Obdobja 28*. Ljubljana: Znanstvena založba Filozofske fakultete. 251–255.

PRUNČ, Erich, 2007: Deutsch-slowenische/kroatische Übersetzung 1848–1918. Ein Werkstättenbericht. [Nemško-slovensko/hrvaški prevodi, 1848–1918. Poročilo z delavnice.]. *Wiener Slavistisches Jahrbuch* 53/2007. Dunaj: Austrian Academy of Sciences Press. 163–176.