Jezikovni viri starejše slovenščine IMP: zbirka besedil, korpus, slovar

Tomaž Erjavec

Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, Ljubljana
tomaz.erjavec@ijs.si

Povzetek

V prispevku predstavimo tri jezikovne vire starejšega slovenskega jezika: zbirko besedil oz. digitalno knjižnico, referenčni jezikoslovno označeni korpus in slovar. Zbirka besedil vsebuje večinoma knjige, z redigirano transkripcijo besedila in faksimili, korpus sestavlja besedilo strani, vzorčenih iz zbirke, kjer je vsaki besedni pojavnici pripisana ročno pregledana sodobna ustreznica, njena lema in leksikalna oblikoskladenjska oznaka, slovar pa je bil zajet iz razširjenega ročno pregledanega korpusa. Vsi trije viri so zapisani skladno s smernicami za zapis besedil TEI (Text Encoding Initiative Guidelines) in dostopni na spletu za pregledovanje in preiskovanje, kot tudi za prenos pod licenco Creative Commons – priznanje avtorstva. Ti viri po eni strani predstavljajo podatkovno infrastrukturo za razvoj jezikovnih tehnologij za starejšo slovenščino, po drugi pa omogočajo empirično podprte diahrone jezikoslovne raziskave in približajo starejša besedila in leksiko sodobnemu bralcu. Viri IMP so dostopni na http://nl.ijs.si/imp/.

Language resources for historical Slovene

The paper presents three language resources for historical Slovene: the text collection or digital library, a linguistically annotated reference corpus and a dictionary or lexicon. The text collections contains mostly books with hand corrected transcriptions and facsimile, the corpus comprises sampled pages from this text collections, where each word token has been manually annotated with its contemporary equivalent, its lemma and part-of-speech tag, while the lexicon was automatically extracted from the (extended) annotated corpus. The three resources are encoded according to the Text Encoding Initiative Guidelines and are available on the web for browsing and searching, as well as for download under the Creative Commons – Attribution licence. The resources, on the one hand, constitute a data infrastructure for the development of language technologies for historical Slovene, and, on the other, enable corpus based diachronic studies of Slovene and bring old texts and vocabulary closed to today's readers. The IMP language resources are available from http://nl.ijs.si/imp/.

1. Uvod

V prispevku predstavimo tri medsebojno povezane digitalne vire starejšega slovenskega jezika, ki predstavljajo referenčne podatkovne vire za razvoj jezikovnih tehnologij za obravnavo starejše slovenščine, imajo pa tudi namen razširiti ponudbo spletnih priročnikov in postaviti osnovo za korpusno podprte diahrone jezikoslovne študije. Če se osredotočimo samo na jezikovnotehnološke vidike uporabe, imajo viri naslednji namen:

- **zbirka besedil** vsebuje čistopise besedil, s tem pa tudi besedišče starejše slovenščine, ki je uporabno za izboljšanje avtomatskega prepoznavanja besedil iz slik (OCR); zbirka služi tudi kot osnovna podatkovna množica, ki jo lahko izkoristimo za povečanje jezikoslovno označenega korpusa in leksikona;
- **jezikoslovno označeni korpus** omogoča šolanje oblikoskladenjskih označevalnikov za starejši jezik in razvoj programov za avtomatsko posodabljanje starih besedil, s čimer jih naredijo bolj razumljiva sodobnemu bralcu;
- **slovar** omogoča šolanje modulov za posodabljanje in lematizacijo besed in je uporaben za podporo programom za iskanje po celotnem besedilu digitalnih knjižnic.

S formalno validacijo in upoštevanjem standardov poskrbimo, da so jezikovni viri brez tehničnih napak in da je njihova struktura dobro dokumentirana. Viri IMP so skladno zapisani in medsebojno povezani, pri čemer upoštevajo dvoje specifikacij:

- **shema XML IMP** definira strukturo XML posameznih, sicer strukturno zelo raznovrstnih virov;
- **oblikoskladenjske specifikacije IMP** definirajo nabor oblikoskladenjskih oznak, ki se uporabljajo v korpusu oz. slovarju.

Izdelava predstavljenih virov je bil razmeroma dolgotrajen in drag proces, zato je smiselno poskrbeti, da so ne samo čim širše uporabni, temveč tudi uporabljani. Vsi viri so dostopni neposredno na spletu skozi izvedene oblike (HTML), obenem pa tudi za prenos, pod licenco Creative Commons – Priznanje avtorstva.

2. Računalniški zapis

Glede na možnost prenosa celotnih virov je format, v katerem so zapisani, pomemben tudi za druge uporabnike. Viri IMP so zapisani v skladu s smernicami za zapis besedil TEI »Text Encoding Initiative Guidelines«, (TEI, 2007). Smernice temeljijo na standardu XML (Extensible Markup Language) in so namenjene za zapis besedil v znanstvene namene. Uporabljajo se za večino kompleksnejših izdaj v digitalnih knjižnicah kot tudi za zapis jezikoslovno označenih korpusov in slovarjev. Smernice TEI definirajo in dokumentirajo nabor oznak (elemente in atribute XML) za zapis strukturnih in konceptualnih lastnosti besedil. Smernice so izražene kot modularna in razširljiva shema XML, ki ji je pridružena podrobna dokumentacija, dostopne pa so pod odprtokodno licenco. Prva je iz leta 1994, zadnja izdaja smernic, ki se sicer sproti dopolnjuje in popravlja, pa je TEI P5 iz leta 2007 in je usklajena z ustreznimi smernicami W3C in ISO, ki jih s tem upoštevajo tudi naše izdaje.

V kontekstu starejše slovenščine velja posebej opozoriti na kodo, ki jo uporabljamo za besedila, zapisana v bohoričici, in sicer sl-bohoric, ki smo jo tudi prijavili na IANA (Internet Assigned Numbers Authority), skupaj s kodama za metelčico (sl-metelko) in dajnčico (sl-dajnko), čeprav slednjih dveh pisav vir IMP zaenkrat ne vsebujejo.

Kanonična oblika posameznega vira IMP je torej dokument XML, ki je veljaven (validiran) glede na shemo XML IMP, ta pa je narejena v skladu s smernicami TEI. S tem je zagotovljeno, da neodvisno od naših datotek obstaja podrobna dokumentacija, v kateri je struktura virov dokumentirana tako s proznim opisom (smernice) kakor s shemo XML, s pomočjo katere je mogoče pravilnost strukture teh virov tudi formalno preveriti.

Zapis TEI je namenjen izmenjavi, primeren je za raznovrstne uporabe, neodvisen od računalniške platforme, in, kolikor je to glede na hiter razvoj računalniških tehnologij sploh mogoče, odporen na zastaranje. Ta format s skriptami XSLT nato pretvorimo v formate za uporabo v konkretnem orodju (npr. konkordančniku) ali za prikaz na spletu.

3. Zbirka besedil

Zbirka besedil IMP je zasnovana kot digitalna knjižnica in vsebuje večinoma celotna dela, v obliki faksimilov in pregledanih transkripcij besedil. Trenutno vsebuje 158 del (13.000 strani oz. 2 milijona besed), sestavljena pa je iz zbirk AHLib (Erjavec, 2011) in zbirk besedil, označenih v NUK ter na ZRC SAZU (Erjavec in dr., 2011). Dela so večinoma celotne knjige, 38 enot pa so izdaje časopisa *Kmetijske in rokodelske novice*.

Stopnja označevanja se razlikuje glede na izvor posameznega dela, v vseh primerih pa vsebuje prelome strani s kazalci na faksimile, naslove, odstavke in oznake za posebne dele besedila, kot so številke strani, tiskarska znamenja itd., kot je razvidno iz primera na sliki 1.

Na spletu je zbirka predstavljena kot digitalna knjižnica, kjer je vsaka enota stavljena kot svoj HTML, ti pa so povezani s statičnimi kazali glede na različne metapodatke: naslov, avtor, leto, signatura.

4. Jezikoslovno označeni korpus

Ročno pregledani referenčni korpus goo300k (Erjavec, 2012) obsega 1.000 vzorčenih strani iz besedilne zbirke IMP oz. nekaj manj kot 300.000 besednih pojavnic. Postopek vzorčenja je potekal v dveh fazah. Najprej smo iz zbirke IMP izbrali besedila izdana pred letom 1900, ki so čim bolj raznovrstna, obenem pa ne preveč izstopajoča po zapisu, s čimer smo dobili 81 enot. Iz teh besedil smo nato naključno izbrali posamezne strani, z nastavljenimi maksimumi za število strani po besedilu in časovnem obdobju. Prednost smo dali besedilom, izdanim med letoma1850 in 1875, saj smo po eni strani imeli največ besedil iz tega obdobja, po drugi pa je jezik tu že zadosti različen od sodobne slovenščine, da bi bila jezikovnotehnološka podpora že upravičena, vendar hkrati še ni preveč različen, da bi bilo posodabljanje posameznih besed neuporabno. Vseeno pa korpus vsebuje tudi besedila, ki segajo do leta 1750 in celo par starejših vzorcev.

Korpus je bil najprej avtomatsko označen, nato pa ročno pregledan s strani skupine študentov, ki so pri svojem delu uporabljali urejevalnik CoBaLT (Kenter in dr. 2012). Poleg pregledovanja oznak je bilo popravljeno tudi besedilo, saj čistopis ni bil brez napak.

Korpus je zapisan v 1.001 datoteki, ena krovna, ostale pa za posamezne strani. Krovna datoteka vsebuje element teiCorpus, ki je sestavljen iz kolofona TEI (teiHeader) in serije elementov TEI, od katerih vsak vsebuje po eno enoto korpusa. Element TEI ima nato svoj kolofon, temu sledijo podatki o faksimilu (za vsako stran njen identifikator, URL-je grafičnih datotek na strežniku in ustrezne strani v digitalni knjižnici) in XInclude kazalce na posamezne strani, ki so vključene v korpus.

Kot je ilustrirano na sliki 2, ima posamezna stran kazalko na svoj faksimile in označeno besedilo, ki je najprej razdeljeno na bloke (naslove, odstavke itd.), ti na stavke oz. povedi, ti pa na besedne pojavnice, ločila in presledke.

```
<ab type="p" corresp="NUKR10214-1790/00422751.xml#r49">Goſpa. O ! jeſt niſim vezh tvoja <lb n="3"/>
Rosalka , katęro ſi ſizer lubil ! jeſt ſim<lb/>
ena oboga shéna, ena ſirota — <lb n="3"/>
nimam moshá!<lb/>
</ab>
<ab type="fw" subtype="catch" corresp="NUKR10214-1790/00422751.xml#r4">Ba-</ab>
<ab type="fw" subtype="sig" corresp="NUKR10214-1790/00422751.xml#r5">c 5</ab>
<pb xml:id="pb.095" n="95" facs="#NUKR10214-1790-00422752"/>
<ab type="fw" subtype="pageNum" corresp="NUKR10214-1790/00422752.xml#r1">58</ab>
<ab type="p" corresp="NUKR10214-1790/00422752.xml#r6">Baron. Vſmili ſe!<lb/></ab>
```

Slika 1. Zapis zbirke besedil.
Element ab zaznamuje »anonimni blok«, ki je opredeljen z vrednosti atributa @type: odstavek, naslov, tiskarsko znamenje, itd. Element lb poda prelom vrstice, pb pa strani. Slednji ima tudi kazalko na faksimile.

```
<?xml version="1.0" encoding="utf-8"?>
<div xmlns="http://www.tei-c.org/ns/1.0" type="pb" xml:lang="sl-bohoric" xml:id="goo18B-NUKR10214-1790.pb.095"
   n="NUKR10214-1790.pb.095_Shupanova_Mizka" facs="../goo300k.xml#NUKR10214-1790-00422752">
  <ab type="p" corresp="NUKR10214-1790/00422752.xml#r6" part="F">
    <s>
      <w nform="baron" mform="baron" lemma="baron" ctag="Ncm">Baron</w><pc ctag=".">.</pc>
    </s>
    <c> </c>
    <s>
      <w nform="vſmili" mform="usmili" lemma="usmiliti" ctag="Vme">Vſmili</w><c> </c>
      <w nform="ſe" mform="se" lemma="se" ctag="P">ſe</w><pc ctag="!">!</pc>
    </s>
  </ab>
  …
</div>
```

Slika 2. Zapis označenega korpusa.

Element div/@type="pb" vsebuje eno stran označenega besedila. Atributi določijo, da element pripada imenskemu prostoru TEI, da je »jezik« na tej strani bohoričica, mu podajo identifikator in labelo, ter kazalko na faksimile. Bloki so razdeljeni na stavke, ti pa na besede, ločila in presledke. Besede imajo pripisano normalizirano obliko, posodobljeno obliko, lemo in oblikoskladenjsko oznako.

4.1. Jezikoslovne oznake

Vsaki besedi v korpusu je pripisana normalizirana oblika, sodobna oblika, lema in oblikoskladenjska oznaka IMP (gl. sliko 2). Normalizirane oblike so zapisane z malimi črkami, poleg tega pa so odstranjena naglasna znamenja nad samoglasniki, saj se ta v sodobni slovenščini ne uporabljajo več.

Sodobne oblike besed so najbolj zanimive, saj takih oznak ne najdemo v korpusih sodobnega jezika. Razdelimo jih lahko v štiri skupine:

1. besedna oblika iz korpusa je enaka sodobni obliki, kot je to pri prvi besedi na sliki 2;
2. razlika je samo v zapisu posamezne besede, kot je to pri drugi besedi na sliki 2;
3. razlika je pri pisanju skupaj – narazen (npr. *nar bolj*, sedaj *najbolj* ali obratno *namoresh*, sedaj *ne moreš*); take besede so problematične s stališča zapisa jezikoslovnega označevanja, saj so posodobljene oblike sicer enostavno pripisane kot atribut posamezni besedi, tu pa je potrebno vzpostaviti relacijo med več besednimi pojavnicami in eno analizo oz. eno pojavnico in nizom analiz;
4. zastarele besede, torej tiste, ki nimajo sodobne ustreznice, ali pa so se jim spremenile skladenjske lastnosti, kot npr. spol; v takih primerih je kot posodobljena oblika vzeta kar zastarela beseda, vendar napisana v skladu s sodobnim pravopisom (npr. *ajfram* posodobimo v *ajfrom*), je pa takim besedam pripisana tudi najbližja sodobna ustreznica oz. ustreznice (v tem primeru *gorečnost*).

Pri zastarelih besedah velja še opomba, da kot zastarelih zaenkrat nismo šteli tistih, ki se pojavljajo v SSKJ, četudi imajo pripisano oznako *zastarelo* oz. *starinsko*. Glavni razlog je bil, da so te besede že obdelane v SSKJ in je sodobne ustreznice oz. razlage možno zajeti iz tega vira.

Besedam je nadalje pripisana lema oz. osnovna oblika besede, ki je tudi posodobljena in izhaja iz posodobljene besedne oblike, npr. *sonce*, *ne moči*, *ajfer*.

Tretji jezikoslovni podatek, ki je pripisan besednim pojavnicam, je njihova kontekstno razdvoumljena oblikoskladenjska oznaka – za namene projekta smo razvili nov nabor oznak oz. specifikacije zanje, kar je podrobneje opisano v naslednjem razdelku.

Korpus je dostopen za iskanje preko spletnega vmesnika, ki omogoča izpis konkordanc in frekvenčnih leksikonov, tudi po regularnih izrazih, iskanje in prikaz vseh oznak po besedah, filtriranje in prikaz bibliografskih podatkov, izračun kolokacij, itd.

4.2. Oblikoskladenjske specifikacije

V korpusih sodobnega jezika, kot sta FidaPLUS (Arhar in Gorjanc, 2007) in JOS[1] (Erjavec in Krek, 2008), uporabljamo oblikoskladenjske oznake, ki zajemajo tako leksikalne (npr. obči samostalnik srednjega spola) kot pregibne lastnosti (npr. rodilnik ednine) posameznih besed. V korpusu in slovarju smo ta sistem, ki zajema skoraj 2.000 različnih oznak, poenostavili in besedam pripisali samo leksikalne lastnosti, tako da število oznak pade na 32.

Oznake IMP so definirane, tako kot oblikoskladenjske oznake JOS, v dokumentu TEI, kjer so definirane besedne vrste, vsaki pripisane njene oblikoskladenjske lastnosti, množice teh pa pripisane posameznim oblikoskladenjskim oznakam, pri čemer so imena lastnosti in oznak definirana tako v slovenščini kot angleščini. V korpusu uporabljamo oznake v angleščini, vendar je te preko izvedenih tabel enostavno prevesti v slovenščino in jih tudi razstaviti v posamezne lastnosti. Tako imamo npr. korpusno oznako

[1] http://nl.ijs.si/jos/

Vmp, ki pomeni Verb Type = main, Aspect = progressive, ki je ekvivalentna slovenski oznaki Ggn oz. glagol vrsta = glavni, vid = nedovršni.

Razlog za uvedbo poenostavljenega nabora oznak je predvsem v tem, da je bil poudarek pri ročnem označevanju na posodobljenih oblikah besed, pri tem pa je natančno označevanje oblikoskladnje zelo zamudno – zato smo raje označili več besedila, vendar z bolj grobimi oznakami. Oznake so vseeno koristne, saj nam omogočijo, da v korpusu npr. iščemo vse kombinacije pridevnikov z neko besedo, po drugi strani pa jih lahko uporabimo za učenje modelov avtomatskega oblikoskladenjskega označevanja starejših besedil.

5. Slovar IMP

Tretji jezikovni vir starejše slovenščine je slovar oz. besedišče. Slovar je bil avtomatsko izluščen iz korpusa goo300k, poleg tega pa so mu bile dodane tiste besedne oblike iz zbirke besedil IMP, ki se ne pojavljajo v goo300k, v zbirki besedil pa vsaj dvakrat. Tudi tem besednim oblikam smo oznake najprej pripisali avtomatsko, nato pa ročno pregledali v CoBaLTu, ravno tako pa je bil naknadno pregledan še celoten slovar. Iz takšnega postopka izdelave seveda sledi, da slovar vsebuje samo korpusno izpričane oblike.

Celoten slovar vsebuje preko 25.000 lem, 50.000 besednih in 70.000 zgodovinskih besednih oblik, vendar to zajema vse pregledane besedne pojavnice iz korpusa oz. zbirke besedil, torej tudi številke, simbole, tujejezične in zatipkane besede in besedne oblike, ki so enake sodobnim. Če se omejimo samo na »prave« besede, pade število lem na nekaj pod 20.000, če samo na tiste leme, ki imajo vsaj eno besedno obliko drugačno, kot je sodobna, na 11.000, če samo na zastarele besede, pa na 2.000.

Slovar, kot vsak dokument TEI, vsebuje najprej kolofon, ki mu sledijo geselski članki. Kot je ilustrirano na sliki 3, vsebuje vsak zaglavje in korpusno izpričane besedne oblike. Zaglavje je sestavljeno iz geselske iztočnice, torej leme, njene oblikoskladenjske oznake oz. lastnosti in, za zastarele besede, sodobne ustreznice kot tudi vira, na osnovi katerega so bile te ustreznice določene. Posamezen geselski sestavek tako definira njegovo zaglavje: enake leme se kot homonimi pojavljajo v več geselskih člankih, če se ti razlikujejo glede na oblikoskladenjsko oznako ali sodobne ustreznice.

Zaglavju sledi seznam vseh sodobnih besednih oblik, vsaka od teh pa ima seznam svojih zgodovinskih različic. Vsaka zgodovinska različica je pospremljena s primeri uporabe iz korpusa, ki so jim pripisani bibliografski podatki. V slovarju je vključenih samo nekaj primerov za vsako zgodovinsko obliko, saj bi bilo vseh, posebej za visokofrekvenčne funkcijske besede, preveč.

Tako kot zbirka besedil je tudi slovar dostopen na spletu za pregledovanje, v več različicah, od polnega slovarja do tistega, ki vsebuje samo zastarele besede. Vsake različica slovarja je razdeljena na večje število strani v HTML, ki so med sabo povezane preko kazala po geselskih iztočnicah. Vsako geslo je opremljeno s kazalci v SSKJ in Pleteršnikov slovar na ZRC SAZU, v konkordančnik in v digitalno knjižnico.

```
<entry xml:id="lex.19e01750cca5a43b17fb31078b279905" n="anati-Vmp_izogibati_se">
  <form type="lemma">
    <orth type="hypothetical">anati</orth>
    <gramGrp norm="Vmp">
      <gram type="msd">Ggn</gram>
      <pos>glagol</pos> <gram type="vrsta">glavni</gram> <gram type="vid">nedovršni</gram>
    </gramGrp>
    <gloss>izogibati se</gloss> <bibl>Pleteršnik</bibl>
  </form>
  <form type="wordform">
    <orth type="hypothetical">anaj</orth>
    <form type="historical">
      <orth type="normalised">anej</orth>
      <cit>
        <quote xml:lang="sl-bohoric">drugemi pohujshanje, ke je njim ſavol myrnoſte tega serza
        navoshliv. <oVar>Anej</oVar> se slehernega kraja, kjer myru ni, inu vogibej se tajſteh </quote>
```

Slika 3. Zapis slovarja.

Geslo (entry) vsebuje podatke o lemi (form/@type="lemma") in njenih posodobljenih besednih oblikah (form/@type="wordform"), ti pa vsebujejo vse atestirane besedne oblike (form/@type="historical") skupaj s primeri uporabe.

6. Dostopnost virov

Če je ena plat dostopnosti virov njihov zapis, je druga dejanska možnost dostopa do njih. Dostopnost jezikoslovnih virov je v Sloveniji vse prevečkrat omejena na spletno pregledovanje, kar sicer v večini primerov zadošča za njihovo neposredno (jezikoslovno) uporabo, ne omogoča pa uporabe za razvoj jezikovnih tehnologij ali za bolj poglobljene, celostne jezikoslovne študije, kjer potrebujemo možnost prenosa celotnega vira na lastni računalnik. Razlogi za takšno zapiranje so v nekaterih primerih sicer legitimni (npr. avtorska zaščita izvornih besedil), v večini primerov pa so bolj želja institucije, ki je vir razvila, da ohrani monopol nad njim, in to kljub temu, da je bil izdelan z javnimi sredstvi (Erjavec, 2009) – najbolj znan primer zapiranja nacionalno pomembnega jezikovnega vira je seveda SSKJ.

Za vse vire IMP velja, da so dostopni ne samo za pregledovanje preko spleta, temveč tudi za prenos v kanonični obliki TEI po licenci *Creative Commons, priznanje avtorstva*. Licenca omogoča prenos virov za uporabo v raziskovalne namene ali komercialno, kot tudi predelavo virov in nadaljnjo distribucijo virov. Edini pogoj, ki je postavljen, je, da se vir IMP, ki se ga uporablja, tudi primerno citira.

7. Zaključki

V članku smo predstavili tri uniformno zapisane, medsebojno povezane in prosto dostopne referenčne vire starejšega slovenskega jezika, dosegljive na http://nl.ijs.si/imp/.

V nadaljnjem delu bi želeli razširiti ročno označeni korpus, predvsem z besedili iz druge polovice 18. in prve polovice 19. stoletja, ravno tako pa zbrati še dodatna pregledana besedila in s pomočjo teh povečati tudi slovar.

Zaželeno bi bilo tudi obogatiti metapodatke virov IMP. Tako ima npr. trenutno vsaka publikacija en naslov, kot smo ga dobili s strani izdelovalcev digitalne predloge. Vendar pa so, posebej pri starejših besedilih, naslovi v različnih jezikih ali pisavah, zato bi bilo koristno imeti naslov v več variantah, tudi v sodobni slovenščini.

Glavna bodoča naloga pa je uporaba razvitih virov, tako v lastnih raziskavah, ki se bodo osredotočile na posodabljanje starejših besedil, kot tudi spodbujanje drugih, da vire uporabijo pri svojih raziskavah in razvoju. Trenutno se slovar že uporablja pri podpori iskanja po starejših besedil v digitalni knjižnici dLib.si, imajo pa izdelani viri še mnogo širše potenciale.

Zahvala

Avtor se zahvaljuje anonimnima recenzentoma za koristne pripombe in nasvete. Pri delu, ki je opisano v prispevku, so sodelovali Kozma Ahačič, Tina Benčina, Katja Cingerle, Metod Čepar, Darja Fišer, Alenka Jelovšek, Urška Kamenšek, Alenka Kavčič Ćolić, Maša Kodrič, Nina Mikulin, Matija Ogrin, Daša Pokorn, Erich Prunč, Zala Šmid, Ines Vodopivec in Maja Žorga Dulmin. Delo sta podprla projekt EU IP IMPACT *Improving Access to Text* in nagrada Google *Developing Language Models of Historical Slovene*

Literatura

Špela Arhar, Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. Jezik in slovstvo, 52 (2). URL: http://www.jezikinslovstvo.com/pdf/2007-02-Razprave-SpelaArharInVojkoGorjanc.pdf

Tomaž Erjavec, Simon Krek. 2008. Oblikoskladenjske specifikacije in označeni korpusi JOS. Zbornik Šeste konference Jezikovne tehnologije, Ljubljana. URL: http://nl.ijs.si/jos/bib/jos_isltc08.pdf

Tomaž Erjavec. 2009. Odprtost jezikovnih virov za slovenščino. V zborniku Obdobja, Simpozij »Infrastruktura slovenščine in slovenistike«. str. 115–121. URL: http://www.centerslo.net/files/file/simpozij/simp28/Erjavec.pdf

Tomaž Erjavec. 2011. Slovenska prevodna književnost 1848–1918 : digitalna knjižnica in korpus AHLIB. V zborniku Meddisciplinarnost v slovenistiki, (Obdobja, Simpozij, = Symposium, 30). Ljubljana: Znanstvena založba Filozofske fakultete, str. 33–40. URL: http://www.centerslo.net/files/file/simpozij/simp30/Zbornik/Erjavec.pdf

Tomaž Erjavec, Ines Jerele, Maša Kodrič. 2011. Izdelava korpusa starejših slovenskih besedil v okviru projekta IMPACT. V zborniku Obdobja, Simpozij »Meddisciplinarnost v slovenistiki«. Ljubljana: Znanstvena založba Filozofske fakultete, str. 41–47. URL: http://www.centerslo.net/files/file/simpozij/simp30/Zbornik/Erjavec_Jerel_Kodric.pdf

Tomaž Erjavec. 2012. The goo300k corpus of historical Slovene. V zborniku Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul. European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2012/summaries/445.html

Tom Kenter, Tomaž Erjavec, Maja Žorga Dulmin, Darja Fišer. 2012. Lexicon construction and corpus annotation of historical language with the CoBaLT editor. V zborniku EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Avignon, France, April. ACL.

TEI Consortium (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange. URL: http://www.tei-c.org/Guidelines/P5/