Posodabljanje starejše slovenščine

Tomaž Erjavec, Institut Jožef Stefan, Odsek za tehnologije znanja, Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Izvleček

V prispevku obravnavamo metodo za posodabljanje besed v starejših slovenskih besedilih, ki vključuje posodabljanje besednih oblik s pomočjo računalniških leksikonov in pravil za transkripcijo, oblikoskladenjsko označevanje in lematizacijo. Posodabljanje je koristno predvsem pri iskanju po polnem besedilu digitalnih knjižnic naše kulturne dediščine, pa tudi kot način, da starejša besedila približamo sodobnemu bralcu. Program za posodabljanje uporablja jezikovne vire starejše slovenščine IMP, ki vključujejo ročno označeni korpus besedil in leksikon starejše slovenščine, za oblikoskladenjsko označevanje in lematizacijo pa modele, naučene na virih sodobne slovenščine, razvitih v okviru projekta Sporazumevanje v slovenskem jeziku. Prispevek predstavi uporabljene vire, program za jezikoslovno označevanje ToTrTaLe, evalvacijo natančnosti programa in smernice za nadaljnje delo.

Ključne besede: starejša slovenščina, jezikovne tehnologije, jezikovni viri za slovenski jezik.

Abstract

Modernizing Historical Slovene

The paper presents a method for modernising words in historical Slovene texts, which includes modernising word-forms with the help of computational lexicons and transcription rules, morphosyntactic tagging, and lemmatisation. Modernisation is useful for full-text search in cultural heritage digital libraries as well as a way to make older texts more accessible to today's readers. The program for modernisation uses the IMP language resources for historical Slovene, which include a hand-annotated text corpus and a lexicon of historical Slovene, while morphosyntactic tagging and lemmatization rely on models trained on resources for contemporary Slovene, which were developed in the scope of the "Communication in Slovene" project. The paper introduces the language resources, the ToTrTaLe program for linguistic annotation, an evaluation of the accuracy of the program and directions for future research.

Key words: historical Slovene, language technologies, language resources for Slovene.

1 UVOD

V zadnjih letih smo priča hitremu razmahu digitalnih knjižnic, pri čemer je veliko dostopnih besedil starejšega datuma, saj ni ovir za njihovo razširjanje, ker so jim potekle avtorske pravice, ob tem pa so taka besedila zanimiva za seznanjanje in preučevanje kulturne dediščine posameznih narodov. Za slovenski jezik sta največji digitalni knjižnici dLib.si (Krstulović in Šetinc, 2005) in projekt Googlovih knjig. Ta dela so tipično dostopna predvsem kot faksimili, v najboljšem primeru s predogledom avtomatsko razpoznanega besedila, v katerem pa je zaradi poškodb papirja, starega tiska in uporabe bohoričice veliko napak. Besedila tudi niso strukturno označena, kar onemogoča npr. generiranje kazala in stavljenje besedila. Obstaja tudi več manjših, a zato bolj natančno obdelanih digitalnih knjižnic slovenske pisne kulturne dediščine,[1] na prvem mestu projekt »Slovenska leposlovna klasika« na Wikiviru, kot tudi portal Sistory (Šorn in Hadalin, 2010), knjižnica eZISS (Ogrin in Erjavec, 2009) in veliko projektov posameznih knjižnic.

Za iskanje po polnem besedilu digitalnih knjižnic je, vsaj za jezike z bogato morfologijo, kot je slovenščina, zelo koristno besedila predhodno lematizirati, torej vsaki besedi pripisati njeno osnovno obliko, npr. »ljubezen« za besedne oblike »ljubezni«, »ljubeznijo« itd. Šele tako bo namreč poizvedba za »ljubezen« vrnila tudi besedila s katero koli pregibno obliko te besede. Za sodobno standardno slovenščino je bilo razvitih že več lematizatorjev, tudi prosto dostopnih (Erjavec in Džeroski, 2004; Juršič idr., 2010; Logar Berginc idr., 2012), pri čemer bolj kakovostni najprej opravijo oblikoskladenjsko označevanje, pri čemer vsaki besedni pojavnici pripišejo njene oblikoskladenjske lastnosti, npr. »občni samostalnik moškega spola v orodniku ednine«, saj je v splošnem šele s to informacijo mogoče neko besedno obliko tudi pravil-

* Delo, objavljeno v tem članku, sta podprla projekt EU IP IMPACT *Improving Access to Text* in nagrada Google *Developing Language Models of Historical Slovene* ter raziskovalni program P2-0103 Tehnologije znanja.

[1] Podroben, čeprav že rahlo zastarel pregled je podan v Hladnik (2009).

no lematizirati. Tako je npr. za besedno obliko hotela treba vedeti, ali je glagol ali samostalnik, da ji lahko pripišemo bodisi lemo hoteti bodisi lemo hotel. Za pravilno lematizacijo neznanih besed pa je oblikoskladenjska oznaka še posebno potrebna.

Sodobni lematizatorji in oblikoskladenjski označevalniki se modela jezika naučijo samodejno na podlagi vnaprej pripravljenih jezikovnih virov. Za razliko od ročno napisanih pravil imajo induktivno naučeni modeli prednost, da so bolj robustni in lahko (razmeroma) uspešno obdelajo tudi neznane besede, zato pa potrebujejo za učne množice ročno izdelane jezikovne vire, tj. leksikone za lematizatorje in označene korpuse za oblikoskladenjske označevalnike. Izdelava dovolj natančnih, obsežnih in raznovrstnih jezikovnih virov za posamezen jezik je drag in dolgotrajen postopek, vendar je za slovenščino v zadnjem času postalo dostopnih večje število takšnih virov, predvsem v okviru projektov Jezikoslovno označevanje slovenskega jezika (JOS) in Sporazumevanje v slovenskem jeziku (SSJ), tako da izdelava induktivnih orodij ni več nepremostljiva težava; kot omenjeno, sedaj obstajajo tudi že vnaprej naučeni prostodostopni lematizatorji in oblikoskladenjski označevalniki za sodobno standardno slovenščino.

Stanje pa je drugačno za računalniško obravnavo starejše slovenščine, saj se ta razlikuje od sodobnega jezika, zaradi česar z obstoječimi programi zanjo dobimo zelo slabe rezultate. Besede so se včasih pisale drugače, njihov zapis se je skozi zgodovino tudi spreminjal, ob tem pa pisni jezik ni bil standardiziran, tako da lahko za isto besedo tudi v istem časovnem obdobju najdemo več zapisov. Če k temu prištejemo še bohoričico, ki so jo uporabljali do srede devetnajstega stoletja, ima lahko posamezna lema zelo veliko število oblik, ki so težko predvidljive vnaprej. Tako za lemo ljubezen v korpusu starejših besedil poleg sodobnih oblik ljubezen, ljubezni in ljubeznijo najdemo še ljubesni, ljubesin, lubeſn, lubesen, lubesni, ljubesen, lubeſne, lubeſni, ljubesnijo, ljubezin, lubesnio, lubesne, lubesn, lubiesn in lubiesen. Dodaten problem so besede, ki jih ne uporabljamo več, kot npr. »bukvovez«, ki je danes knjigovez, saj od uporabnika, ki bi rad iskal po besedilih digitalne knjižnice, težko pričakujemo, da se bo zavedal vseh zastarelih ustreznic sodobnim besedam.

V prispevku predstavimo program, ki starejše slovenske besede posodobi, jih oblikoskladenjsko označi in lematizira. V drugem razdelku najprej predstavimo jezikovne vire, ki so omogočili izdelavo programa, v tretjem razdelku opišemo delovanje programa, v četrtem ocenimo njegovo točnost in v petem razdelku podamo sklepe in smernice za nadaljnje delo.

2 UPORABLJENI JEZIKOVNI VIRI

Za označevanje starejših besed uporabljamo več jezikovnih virov, bodisi neposredno ali pa za učenje modelov za posamezne ravni jezikoslovne analize. V tem razdelku opišemo te vire, ki so uporabni tudi zunaj konteksta posodabljanja starejših besedil. Vsi so zapisani po mednarodnih standardih in priporočilih in prosto dostopni pod eno od licenc Creative Commons, tako da so čim bolj odprti (Erjavec, 2009) in lahko v največji meri spodbujajo napredek jezikovnih tehnologij za slovenski jezik.

Večina predstavljenih virov je zapisana skladno s smernicami za zapis besedil TEI, Text Encoding Initiative Guidelines (TEI, 2007). Smernice temeljijo na XML, opredeljujejo formalni zapis besedil za znanstvene namene in se uporabljajo za večino kompleksnejših izdaj v digitalnih knjižnicah, za jezikoslovno označene korpuse, za računalniške slovarje itd. Smernice TEI in s tem spodaj našteti viri so usklajeni z ustreznimi standardi W3C, ISO in IANA, npr. pri kodah za označevanje časov in jezikov. Kot primer izpostavimo oznako za bohoričico, ki do sedaj ni imela svoje standardizirane kode. V postopku izdelave virov starejše slovenščine smo na IANA (Internet Assigned Numbers Authority) prijavili kodo za podjezik »sl-bohoric«, ki je namenjena za označevanje slovenskih besedil, zapisanih v bohoričici, in – čeprav naši viri ne vsebujejo teh pisav[2] – še za »sl-metelko« in »sl-dajnko«.

2.1 Zbirka starejših slovenskih besedil IMP

Podlaga za izdelavo vseh drugih jezikovnih virov starejše slovenščine (Erjavec, 2012a) je zbirka besedil, imenovana IMP, ki je zasnovana kot digitalna knjižnica. Zbirka vsebuje tiskana besedila, večinoma celotne knjige, ki so predstavljene tako s faksimili kot z ročno pregledanimi in označenimi prepisi besedil. IMP trenutno vsebuje 658 del oz. okoli 46.000 strani ali 14 milijonov besed. S par izjemami obsegajo dela obdobje od konca 18. stoletja do leta 1918, večina pa jih je iz druge polovice 19. stoletja.

[2] Zbirka IMP sicer vsebuje knjigo Čelarstvo (Čebelarstvo) Petra Dajnka (1831), ki je zapisana v dajnčici, vendar v prepisu uporabljamo gajico, saj za dajnčico ne obstajajo znaki unikod niti ustrezni fonti za prikaz.

Stopnja označevanja TEI se razlikuje glede na digitalni vir posameznega dela, v vseh primerih pa vsebuje metapodatke (kolofon TEI), prelome strani s kazalci na faksimile, naslove razdelkov in odstavke, tipično pa tudi oznake za posebne dele besedila, kot so verzi, opombe, tiskarska znamenja, uredniški popravki, tuje besede itd. Na spletu je zbirka dostopna v obliki digitalne knjižnice z več kazali, pri čemer je vsaka enota svoja datoteka HTML, samodejno prevedena s stili TEI XSLT iz izvornega zapisa zbirke v XML/TEI.

Na sliki 1 ilustriramo iztržek ene od knjig iz zbirke IMP v zapisu TEI, pri čemer element <pb> pomeni prelom strani, nato se začne razdelek besedila ( ), ki vsebuje naslov (<head>) in začetek prve kitice (<lg>); ta je nato sestavljena iz vrstic (<l>), ki lahko vsebujejo tudi opombe (<note>). Elementi imajo tudi atribute, ki vsebujejo npr. identifikator (@xml:id), preko katerega je mogoče kazati na določen element, opis prikaza elementa (@rend), kazalko na faksimile (pb/@facs) ali dejstvo, da je opomba avtorjeva (note[@type=«authorial«]) in ne uredniška.

```
<pb facs="#WIKI00009-019" n="19"
xml:id="pb.019"/>
<div xml:id="wv-1._Dershi_ali_vmirajozha_.
C5.BFkopo.C5.BFt.">
    <head rend="centered italic">1. Dershi
    ali vmirajozha ſkopoſt.</head>
    <lg>
        <l>Dershi<note xml:id="ref1"
        type="authorial">Dershi, Paſko:
        peſje iména, <hi
        rend="gothic">Hundsnahmen mie
        z. B. Phylar.</hi>
        </note>, ker je v' neki nozhi</l>
        <l>Ne satiſnil ſvojih ózhi,</l>
        <l>Da je sveſti varih bil;</l>
        …
```

Slika 1: **Zapis TEI iztržka besedila iz zbirke besedil IMP**

2.2 Ročno označeni korpus starejših slovenskih besedil goo300k

Iz zbirke IMP smo vzorčili 1.100 strani iz 90 enot in vsako besedno pojavnico (nekaj manj kot 300.000) ročno označili z več jezikoslovnimi lastnostmi, s čimer smo dobili referenčni korpus starejše slovenščine po imenu goo300k (Erjavec, 2012b). Označene jezikoslovne lastnosti so:

1. sodobna ustreznica, torej besedna oblika, kot se piše danes, napisana z malimi črkami, pri čemer za zastarele (izumrle) besedne oblike upoštevamo pravila sodobnega pravopisa;
2. lema oz. osnovna oblika sodobne ustreznice;
3. najbližje sodobne ustreznice oz. kratka razlaga pomena (samo za zastarele besede);
4. leksikalni del oblikoskladenjske oznake JOS (razloženo v nadaljevanju).

Zapis korpusa ponazarja slika 2 z besedilom, ki se glasi: »*Pri vkvartirjanju ni drugači*.« Ta stavek (<s>) ima označene besede (<w>), ločila (<pc>) in presledke (<c>), besede pa nosijo informacijo o lemi (w/@lemma) in oblikoskladenjski oznaki (w/@ana). V primerih, ko se posodobljena beseda (ki je vedno napisana z malimi črkami) razlikuje od besedne oblike iz korpusa, se lahko odločimo (<choice>), ali želimo upoštevati izvorno (<orig>) ali posodobljeno obliko (<reg>). Pri zastarelih besedah je dodan opis (<desc>), sestavljen iz sodobne ustreznice oz. razlage (<gloss>) in vira te razlage (<bibl>); v podanem primeru je bil to kar (širši) kontekst, v katerem se je pojavila beseda »*vkvartirjanju*«. Zapis je bolj kompleksen, kot se zdi potrebno, vendar mora zajeti tudi primere, ko je ena zgodovinska beseda pisana kot več sodobnih ali obratno, npr. »po noči« proti »ponoči«.

```
<s>
    <choice>
        <orig><w>Pri</w></orig>
        <reg><w lemma=«pri« na=«#S«>pri</
        w></reg>
    </choice>
    <c> </c>
     <choice>
        <orig><w>vkvartirjanju</w></orig>
        <reg><w lemma=«ukvartiranje«
        ana=«#Ncn«>ukvartiranju</w>
            <desc><gloss>prenočevanje</
            gloss><bibl>kontekst</bibl></desc>
        </reg>
    </choice>
    <c> </c>
    <w lemma=«biti« ana=«#Va«>ni</w>
    <c> </c>
    <choice>
        <orig><w>drugači</w></orig>
        <reg><w lemma=«drugače«
        ana=«#Rgp«>drugače</w></reg>
     </choice>
    <pc>.</pc>
</s>
```

Slika 2: **Primer iz ročno označenega korpusa goo300k**

Čeprav je natančna definicija vsake od oznak kompleksna, saj v jeziku vedno srečujemo mejne primere, je osnovni pomen vsake od njih vseeno intuitivno jasen. Izjema so oblikoskladenjske oznake, zato jih podrobneje opišemo v nadaljevanju.

Oblikoskladenjske oznake JOS so kratki nizi (npr. »Ggdn«), ki jih lahko pripišemo posamezni besedni pojavnici v korpusu (ali besedni obliki v leksikonu) in kodirajo oblikoskladenjske lastnosti (npr. »glagol, vrsta=glavni, vid=dovršni, oblika=nedoločnik«). Nabor teh oznak (preko 1.900) za slovenski jezik in njihova preslikava v lastnosti so definirane v oblikoskladenjskih specifikacijah JOS (Erjavec in Krek, 2008). Na spletu so dostopne celotne specifikacije tako v izvornem zapisu TEI kot v izvedenem HTML, na voljo pa so tudi tabele, ki oznake preslikajo v lastnosti oz. iz slovenskega v angleški jezik (npr. »Ggdn« ≡ »Vmen« ≡ »Verb, Type=main, Aspect=perfective, VForm=infinitive«). Oznake JOS uporabljajo raznovrstni viri sodobne slovenščine, med drugim v nadaljevanju opisana računalniški leksikon Sloleks in učni korpus ssj500k.

Pri izdelavi jezikovnih virov starejše slovenščine je bil poudarek na ročnem označevanju sodobne oblike in leme, ne pa oblikoskladenjskih lastnosti, kar je zelo zamudno delo. Vseeno smo želeli imeti ročno preverjene vsaj leksikalne lastnosti posameznih lem, zato smo kompleksen nabor vseh oznak JOS reducirali s skoraj dva tisoč na 32. V naboru JOS je tako npr. za besedno pojavnico »ni« zapisano, da je »glagol vrsta=pomožni oblika=sedanjik oseba=tretja število=ednina nikalnost=zanikani«, v goo300k pa samo »glagol vrsta=pomožni« oz. »Va«, ker uporabljamo angleške oznake. Specifikacije oblikoslovnih lastnosti in oznak IMP so, tako kot JOS, tudi formalno zapisane in dostopne na spletu.

2.3 Leksikon starejše slovenščine IMP

Leksikon vsebuje zajete podatke iz korpusa, sestavljen pa je iz gesel, pri čemer posamezno geslo vsebuje lemo, njene oblikoskladenjske lastnosti in (za zastarele besede) sodobne ustreznice, nato seznam sodobnih besednih oblik, za vsako od teh njene zgodovinske ustreznice in nekaj primerov (konkordanc) iz besedil. Leksikon je pretvorjen iz korpusa goo300k, poleg tega pa dopolnjen z ročno obdelanimi pogostejšimi besedami iz večje podmnožice zbirke IMP. Ker leksikon izvira iz označenih korpusnih primerov, so v njem zajete samo dejansko izpričane oblike oz. njihove oznake, zato leksikon tipično ne vsebuje celotnih pregibnih paradigem (tj. vseh besednih oblik) posameznih lem.

Leksikon vsebuje več kot 80.000 zgodovinskih oblik, 58.000 sodobnih besednih oblik in 28.000 lem. Štete so tudi »besede«, kot so cifre, zatipkane in tuje besede, pa tudi besede, ki so enake tistim v sodobni slovenščini. Če štejemo samo vnose, ki imajo vsaj eno besedno obliko različno od sodobne, dobimo okoli 36.000 zgodovinskih oblik, 25.000 sodobnih oblik in 12.000 lem, med katerimi je 4.000 lem zastarelih, zato imajo tudi dodano razlago. Leksikon je dostopen na spletu v formatu HTML, ki je s posebej zato napisanim slogom XSLT pretvorjen iz izvornega TEI/XML.

2.4 Oblikoslovni leksikon sodobne slovenščine Sloleks

Za posodabljanje in lematizacijo potrebujemo tudi leksikon sodobne slovenščine, pri čemer uporabljamo oblikoskladenjski leksikon sodobne slovenščine Sloleks (Arhar, 2009), ki vsebuje okoli 100.000 lem, vse njihove pregibne oblike z oblikoskladenjskimi lastnostmi in s številom pojavitev v korpusu Gigafida, vsega skupaj skoraj 2,800.000 oblik. Leksikon za razliko od drugih naštetih virov ni zapisan v shemi TEI, temveč po XML, ki sledi LMF (Lexicon Markup Framework), standardu ISO 24613:2008 za predstavitev računalniških leksikonov. Ker je struktura LMF razmeroma zahtevna za uporabo, vsebuje pa tudi podatke, ki jih mnoge aplikacije ne potrebujejo, smo leksikon pretvorili še v preprost tabelarični format, v katerem je vsak vnos (vrstica) sestavljen iz besedne oblike, leme, oblikoskladenjske oznake in frekvence tega trojčka na milijon besed. Kot primer podamo v sliki 3 paradigmo samostalnika »skopost«, pri čemer frekvenca nič pomeni, da tega trojčka program ni identificiral v korpusu.

skopostih	skopost	Ncfdl	0.000000
skopostih	skopost	Ncfpl	0.000000
skopostim	skopost	Ncfpd	0.000000
skoposti	skopost	Ncfdn	0.000000
skoposti	skopost	Ncfdg	0.000000
skoposti	skopost	Ncfda	0.000000
skoposti	skopost	Ncfsd	0.000010
skoposti	skopost	Ncfsl	0.000088
skoposti	skopost	Ncfsg	0.000131
skoposti	skopost	Ncfpn	0.000001
skoposti	skopost	Ncfpg	0.000003
skoposti	skopost	Ncfpa	0.000004
skopostjo	skopost	Ncfsi	0.000037
skopostma	skopost	Ncfdd	0.000000
skopostma	skopost	Ncfdi	0.000000
skopostmi	skopost	Ncfpi	0.000000
skopost	skopost	Ncfsn	0.000179
skopost	skopost	Ncfsa	0.000092

Slika 3: **Paradigme ene besede iz leksikona Sloleks v tabelaričnem formatu**

2.5 Učni korpus sodobne slovenščine ssj500k

Za oblikoskladenjsko označevanje potrebujemo učni korpus, za kar uporabimo korpus sodobne slovenščine ssj500k (Arhar, 2009). Korpus vsebuje 500.000 besednih pojavnic; vsaka je ročno označena z oblikoskladenjsko lastnostjo in lemo. Korpus je tudi delno označen s skladenjskimi analizami in imenskimi entitetami, vendar tu ne uporabljamo teh informacij. Zapis je podoben kot za korpus starejše slovenščine, vendar preprostejši, saj ne vsebuje posodabljanja besed.

3 PROGRAM ToTrTaLe

Program za jezikoslovno označevanje starejših besedil ToTrTaLe, katerega prva različica je predstavljena v Erjavec (2011), implementira cevovod, ki iz vhodnega dokumenta TEI izlušči besedilo, nato pa nad njim enega za drugim pokliče posamezne module za označevanje. Za osnovo mu služi program ToTaLe (Erjavec idr., 2005), ki razdeli besedilo na pojavnice (tokenizacija), te oblikoskladenjsko označi (tagiranje) in jim pripiše osnovno obliko (lematizacija). Program, ki ga predstavljamo, doda prepis starinskih oblik v sodobne (transkripcija) takoj za tokenizacijo in se zato imenuje ToTrTaLe. Program na izhod izpiše dokument TEI, v katerem so vhodnim oznakam TEI dodane jezikoslovne oznake, kot so bile prikazane na primeru ročno označenega korpusa goo300k na sliki 2; izhod iz programa na delčku besedila iz slike 1 je prikazan na sliki 3.

```
<div xml:id="wv-1._Dershi_ali_vmirajozha_.
C5.BFkopo.C5.BFt.">
   <head rend="centered italic">
      <s>
         <w lemma="1." ana="Mdo">1.</w>
         <c> </c>
         <choice>
            <orig><w>Dershi</w></orig>
            <reg><w lemma="držati"
            ana="Vmpr3s">drži</w></reg>
         </choice>
         <c> </c>
         <w lemma="ali" ana="Cc">ali</w>
         <c> </c>
         <choice>
            <orig><w>vmirajozha</w></orig>
            <reg type="pattern"
            n="[u←v+č←zh]">
               <w lemma="umirajoč"
               ana="Agpfsn">umirajoča</w>
            </reg>
         </choice>
         <c> </c>
         <choice>
            <orig><w>ſkopoſt</w></orig>
            <reg type="pattern"
            n="[s←ſ+s←ſ]">
               <w lemma=«skopost«
               ana=«Ncfsn«>skopost</w>
            </reg>
         </choice>
         <pc>.</pc>
      </s>
```

Slika 4: **Primer besedila, označenega s ToTrTaLe**

Program je v glavnem jezikovno neodvisen, saj uporablja zunanja pravila in modele, ki jih je mogoče napisati oz. se jih induktivno naučiti za večino evropskih jezikov, čeprav je mišljen predvsem za jezike z bogato morfologijo, kot je slovenščina. Program je napisan v programskem jeziku Perl, vendar je glavni program v resnici samo ovojnica, ki kliče druge programe in nato kombinira njihove rezultate. V nadaljevanju razdelka predstavimo posamezne module ToTrTaLe, pri čemer se najbolj posvetimo specifikam obdelave starejše slovenščine.

3.1 Tokenizacija

Za razdelitev besedila na stavke, besede, ločila in presledke uporabljamo večjezični tokenizator mlToken, ki je del paketa To(Tr)TaLe. Program jezikovno odvisne podatke hrani v ločenih datotekah, predvsem seznam okrajšav (besede, ki se končajo s piko in ne končajo nujno stavka), seznam večbesednih enot (pojavnice, ki so sestavljene iz več s presledki ločenih besed) in seznam levih ali desnih naslonk (besed, ki jih je treba obravnavati kot del neke pojavnice). V kontekstu posodabljanja starejše slovenščine sta posebno zanimiva seznama večbesednih enot in naslonk, saj se precej besed, ki so se včasih pisale skupaj, sedaj piše narazen oz. obratno, npr. »nemore« proti »ne more« oz. »še le« proti »šele«. Te besede so dodane v ustrezen seznam, tako da že mlToken poskrbi za njihovo tokenizacijo v skladu s sodobno normo. Potrebni seznami za tokenizacijo starejše slovenščine za ToTrTaLe niso napisani posebej za to orodje, pač pa so zajeti neposredno iz leksikona IMP.

Trenutni pristop k reševanju teh posebnih pojavnic ima dve slabosti.

- Tokenizator pozna samo tiste posebne pojavnice, ki so v leksikonu, in torej ne obravnava pravilno novih, neznanih okrajšav, večbesednih enot oz. naslonk. Problem je posebno opazen pri presežniku pridevnikov, ki so se včasih pisali narazen (npr. »nar večji«), saj bomo s leksikonom težko zajeli vse oblike vseh stopnjevanih pridevnikov.
- Kot pri vseh drugih jezikoslovnih analizah se tudi pri posebnih pojavnicah srečamo s problemom dvoumnosti, pri čemer je klasifikacija neke pojavnice ali kombinacije pojavnic odvisna od sobesedila, npr. »Vesoljni potop je *po tem* vso deželo potopil«, kjer mora biti sodobna oblika »potem«, in »To se vidi tudi *po tem*, da vse tuje bolj ceni«, kjer pa mora biti »po tem«. Da vsaj deloma rešimo ta problem, v leksikon vedno vključimo oba primera, torej ne samo, ko se starinski »po tem« piše sodobno »potem«, temveč tudi ko so piše »po tem«. V tokenizator nato dodamo posebne primere samo tam, kjer je njihova frekvenca višja od navadnih, torej nezdruženih oz. nerazdeljenih pojavnic.

3.3 Transkripcija

Transkripcija zgodovinskih besednih oblik v sodobne je ključni modul za procesiranje starejšega jezika. Pri posodabljanju besednih oblik so le-te najprej normalizirane, tj. zapisane z malimi črkami, odstranjena pa so tudi naglasna znamenja nad samoglasniki; naglase so namreč pogosto, a neenotno uporabljali predvsem v 19. stoletju, v sodobni normi pa jih skoraj ni zaslediti.

V procesu iskanja sodobne ustreznice program najprej išče normalizirano zgodovinsko besedno obliko v leksikonu IMP; če jo najde, je s tem našel tudi sodobno ustreznico, če ne, pa besedno obliko išče v Sloleksu. Če nobeden od leksikonov ne vsebuje iskane oblike, program njen sodobni zapis skuša najti s pomočjo t. i. transkripcijskih vzorcev.

Veliko sprememb v pisavi lahko namreč izrazimo v obliki pravil, ki podajo vzorec, v katerem se sodobna beseda razlikuje od zgodovinske, npr. »r → er« za pare kot je »brž → berž«, »srce → serce«, pri čemer je na levi sodobni in na desni zgodovinski zapis. Pri uporabljenem pristopu v leksikonu sodobnih oblik Sloleks skušamo najti tiste, ki jih je mogoče izpeljati iz zgodovinske oblike z uporabo enega ali več takih pravil. Ta pristop je tipičen za posodabljanje starejših besedil (Pilz idr., 2008; Gotscharek idr., 2009; Bennett idr., 2010; Sánchez-Marco idr., 2010), se pa pristopi razlikujejo v tehnologiji, ki jo uporabljajo za preverjanje ujemanja zgodovinske oblike s sodobnimi oblikami s pomočjo takšnih vzorcev.

V paketu ToTrTaLe ujemanje prek transkripcijskih vzorcev implementira knjižnica Vaam, Variant aware approximate matching (Gotscharek idr., 2009; Reffle, 2011), ki jih modelira kot (razširjene) končne avtomate, zaradi česar je prostorsko, predvsem pa časovno nezahtevna. Seznam sodobnih kandidatov, ki ga vrne za posamezno zgodovinsko besedo, je urejen glede na število vzorcev, ki jih je bilo treba uporabiti. Proces določanja sodobnih ustreznic je torej nedeterminističen, je pa v danem kontekstu seveda pravilna samo ena posodobitev. Trenutno modul za transkripcijo izbere tistega kandidata, ki ima v leksikonu Sloleks najvišjo frekvenco, vendar so mogoči tudi kompleksnejši modeli, ki bi odložili izbiro najboljšega kandidata, dokler nista opravljena še oblikoskladenjsko označevanje in lematizacija vseh (variant) pojavnic, saj bi s tem imeli več informacij za pravilno odločitev.

Za posodabljanje trenutno uporabljamo okoli sto vzorcev, ki smo jih določili s pomočjo ročno označenega korpusa goo300k; razdeljeni so na vzorce za starejša besedila v gajici (torej sodobni abecedi) in na vzorce za bohoričico. Razlog za dve množici ni samo

razlika v pisavah, temveč so v besedilih izpred leta 1850 vzorci pogosto drugačni.

3.4 Oblikoskladenjsko označevanje

V naslednji stopnji označevanja program pripiše vsaki besedni pojavnici njeno (od konteksta odvisno) oblikoskladenjsko oznako JOS. Sodobni oblikoskladenjski označevalniki se modela jezika naučijo iz ročno označenega korpusa, vendar pa je razvoj dovolj velikega korpusa dolgotrajen in drag proces, ki bi ga težko ponovili za zgodovinski jezik. Ker so bile besedne oblike v predhodnem koraku posodobljene, lahko označevalniku kot vhod ponudimo posodobljeno besedilo in nato uporabimo model, naučen na sodobnem jeziku. Seveda model še vedno deluje slabše kot nad sodobnim jezikom, saj so zgodovinska besedila drugačna ne samo v pisavi posameznih besed, temveč tudi na skladenjski ravni, pa tudi nekatere besedne oblike, kot npr. deležja na -vši, so bila v preteklosti bistveno bolj pogosta, kot so danes.

Za oblikoskladenjsko označevanje uporabljamo program TnT, Tri-grams and tags (Brants, 2000), ki je robusten in hiter trigramski označevalnik, označevati pa zna tudi neznane besede, čeprav je tu natančnost manjša kot za znane. Model označevanja je bil naučen na učnem korpusu sodobne slovenščine ssj500k, pri čemer je kot zaledni leksikon uporabljen Sloleks.

3.5 Lematizacija

Zadnja stopnja jezikoslovne obdelave je pripis osnovne oblike vsaki besedni pojavnici. Kot pri oblikoslovnem označevanju se tudi pri tem večina sodobnih lematizatorjev nauči modela jezika iz vnaprej pripravljenih jezikovnih virov, v tem primeru iz leksikona sodobnih besednih oblik, v našem primeru Sloleksa. Seveda bi lahko leme besednih oblik, vsebovanih v leksikonu Sloleks, preprosto prepisali iz leksikona, vendar imajo lematizatorji to prednost, da znajo lematizirati tudi neznane besede. Če je beseda pravilno posodobljena in ji je pripisana pravilna oblikoskladenjska oznaka, deluje lematizator s precej visoko stopnjo natančnosti.

Kot lematizator uporabljamo CLOG (Erjavec in Džeroski, 2004), ki se na podlagi vhodnih primerov (parov besedna oblika – lema, pri čemer je model za vsako oblikoskladenjsko oznako obravnavan posebej) nauči odločitvene sezname prvega reda, pri čemer je definirana operacija povezovanje nizov. Naučene strukture so predikati v programskem jeziku Prolog, vendar jih za lažjo povezljivost s ToTrTaLe prevedemo v Perl.

Zanimiva lastnost lematizatorja CLOG je, da mu ne uspe lematizirati poljubnega para oblika – oblikoskladenjska oznaka. Pri starejših besedilih so taki primeri skoraj vedno zastarele besede, ki niso bile pravilno posodobljene, tako da so nelematizirane besede dobri kandidati za dodajanje v leksikon IMP.

3.6 Izhod TEI

Zadnja stopnja obdelave je zapis označenega besedila v dokument TEI, kar dosežemo s kombinacijo obdelave v jeziku Perl s skriptami XSLT, čemur sledi še validacija dobljenega dokumenta XML glede na shemo TEI, pri čemer je ta izražena v Relax NG (ISO/IEC 19757-2). Če pride pri validaciji do napak, je to indikator, da vhodni dokument krši (mogoče implicitne) predpostavke označevanja; v tem primeru je treba bodisi popraviti oznake v vhodnem dokumentu ali pa – če je bilo uporabljeno označevanje smiselno – dopolniti program ToTrTaLe, da bo zajel tudi takšne primere. Označevanje v dokumentih TEI je namreč lahko zelo kompleksno, zato je v splošnem težko zagotoviti, da vstavljanje novih (jezikoslovnih) oznak v tak dokument ne privede do nepravilnih struktur. Vendar je ToTrTaLe razmeroma robusten, saj označi vseh 658 del iz zbirke IMP tako, da je izhod pravilen TEI.

4 EVALVACIJA OZNAČEVANJA

V tem razdelku poskusimo odgovoriti na vprašanje, kako dobro ToTrTaLe posodablja, lematizira in oblikoskladenjsko označuje neznane besedne oblike glede na časovno obdobje, v katerem je nastalo besedilo.

Kot je bilo omenjeno v razdelku 2.3, je leksikon zgodovinskih besednih oblik IMP sestavljen iz:

1. vseh besednih oblik iz korpusa goo300k,
2. besednih oblik z ročno preverjenimi oznakami iz vzorca celotne zbirke besedil IMP; ta vzorec tu poimenujemo korpus IMPtest.

Za eksperiment smo programu ToTrTaLe dali na voljo samo prvi leksikon, drugi leksikon pa smo uporabili kot testno množico. Povedano bolj natančno, korpus IMPtest smo najprej razdelili v tri podkorpuse, vsakega za eno časovno obdobje, in sicer za drugo polovico 18. stoletja (18B), prvo polovico 19. stoletja (19A) in drugo polovico 19. stoletja (19B). Nato smo vsakega od podkorpusov označili s ToTrTaLe in

iz njih izločili leksikon ročno pregledanih besednih oblik, skupaj z njihovimi ročnimi ter avtomatskimi oznakami za posodobljeno obliko, lemo in oblikoskladenjsko oznako.

V tabeli 1 podamo nekaj kvantitativnih podatkov o tem testnem leksikonu. V tabeli posebej izpostavimo zgodovinske in sodobne oblike, pri čemer kot sodobne štejemo tiste, v katerih je besedna oblika iz besedila enaka kot sodobna, četudi s transliteracijo iz bohoričice v gajico, kot zgodovinske pa vse ostale. Tako kot sodobno štejemo npr. *bojiš* → *bojiš* kot tudi *bojiſh* → *bojiš*, za zgodovinsko pa npr. *boh* → *bog*. Za zgodovinske, sodobne in vse oblike podamo število vseh vnosov v leksikonu, število različnih besednih oblik, število različnih posodobljenih besed in število različnih lem.

Tabela 1: **Velikost testnega leksikona**

Obdobje	**Zgodovinske oblike**				**Sodobne oblike**				**Vse oblike**			
	Vnosov	Oblik	Poso.	Lem	Vnosov	Oblik	Poso.	Lem	Vnosov	Oblik	Poso.	Lem
18B	3.400	3.224	2.843	1.885	1.105	1.090	1.090	902	4.505	4.270	3.841	2.535
19A	3.484	3.366	3.168	2.228	3.385	3.326	3.298	2.483	6.820	6.572	6.245	4.166
19B	2.104	2.040	2.012	1.581	10.668	10.320	10.320	7.677	12.745	12.220	12.078	8.596
Σ	8.790	8.407	7.209	4.629	14.677	14.239	13.932	9.660	23.341	22.270	20.050	12.288

Kot je razvidno iz tabele, ima leksikon nekaj čez 23.000 vnosov oz. 22.000 besednih oblik, 20.000 posodobljenih oblik in 12.000 lem. Od tega je v 18B sodobnih okoli 25 odstotkov besednih oblik, v 19A jih je 50 odstotkov, v 19B pa 85 oz. 64 odstotkov, ne glede na časovno obdobje; tolikšna bi bila torej tudi natančnost identifikacije sodobnih besednih oblik sistema, ki ne bi opravljal posodabljanja.

V tabeli 2 podamo točnost ToTrTaLe z leksikonom goo300k nad leksikonom neznanih besednih oblik iz tabele 1. Točnost posodabljanja čez vsa obdobja je okoli 70 odstotkov, kar vključuje tako sodobne kot zgodovinske besede. Samo za zgodovinske je točnost pod 30 odstotki za besedne oblike in nekoliko večja za lematizacijo. Zanimivo je, da je točnost posodabljanja največja pri najstarejših besedilih, pri katerih je nekaj manj kot 35-odstotna. Obratno, kar je tudi pričakovano, pa točnost oblikoskladenjskega označevanja pada s starostjo besedil, od skoraj 70 pri 19B do 57 odstotkov pri 19B.

Za sodobne oblike morda preseneča, da je točnost »posodabljanja« manjša od sto odstotkov, za 18B je napaka celo štiriodstotna. Te napake so posledica dejstva, da sodobni leksikon Sloleks ne vsebuje posodobitev vseh besed, ki jih najdemo v testnem leksikonu – v takih primerih sistem poskusi posodobiti neznano (sodobno) besedo, pri čemer mu to v nekaterih primerih tudi uspe, vendar dobimo kot rezultat napačno obliko.

Tabela 2: **Točnost posodabljanja, lematizacije in oblikoskladenjskega označevanja testnega leksikona**

Obdobje	**Zgodovinske oblike**			**Sodobne oblike**			**Vse oblike**		
	Poso.	Lem.	Oblikoskl.	Poso.	Lem.	Oblikoskl.	Poso.	Lem.	Oblikoskl.
18B	34,7 %	38,5 %	56,8 %	96,2 %	87,7 %	79,3 %	49,8 %	50,6 %	62,3 %
19A	26,5 %	31,1 %	57,8 %	97,3 %	90,8 %	84,4 %	61,3 %	60,5 %	70,8 %
19B	24,2 %	32,2 %	68,6 %	99,3 %	93,0 %	85,1 %	86,9 %	82,9 %	82,4 %
Σ	28,8 %	33,9 %	59,9 %	98,6 %	92,0 %	84,3 %	72,4 %	70,2 %	75,1 %

Kot je razvidno iz rezultatov, je točnost sistema trenutno razmeroma slaba, vendar se je treba zavedati, da je posodabljanje kompleksen proces, pa tudi da sistem v praksi deluje bolje, kot nakazujejo številke. Predstavili smo namreč rezultate na neznanih besedah, ne na vseh, pri tem pa ima produkcijski ToTrTaLe na voljo ves leksikon, vključno s testnim, ki smo ga tu izločili, zaradi česar je njegova točnost na

vseh besedah bistveno boljša. Predstavljeni rezultati so slabši tudi zaradi tega, ker testni leksikon vsebuje besede, ki jih – vsaj trenutno – program ne more najti v Sloleksu, tj. zastarele besede, tujke in zatipkane besede, ki skupaj sestavljajo več kot deset odstotkov vnosov v testnem leksikonu.

5 SKLEP

V prispevku smo predstavili metodologijo, jezikovne vire in program za posodabljanje, lematizacijo in oblikoskladenjsko označevanje starejših besedil ter izvedli poskus, s katerim smo ocenili točnost programa na neznanih besedah. Rezultati kažejo, da je točnost mogoče še zelo povečati, kar lahko dosežemo na več načinov, ki ostajajo za nadaljnje delo. Najbolj preprosto (pa tudi najbolj zamudno oz. drago) bi bilo dodajati nove besede in njihove posodobitve v leksikon IMP, v katerem so nato neposredno dostopne. Zelo koristno, vendar prav tako zamudno, bi bilo dodajati nove besede tudi v leksikon sodobnih besed, saj analiza napak posodabljanja pokaže, da bi vzorci včasih pravilno predvideli sodobno obliko, a te ni Sloleksu. Ravno tako bi bilo dobro v leksikon sodobnih oblik dodati tudi (najpogostejše) tuje besede, predvsem v latinščini, nemščini in francoščini. Več dela bi lahko vložili tudi v transkripcijske vzorce, saj nismo pokrili vseh regularnih sprememb. Vendar se ob tem pojavi problem lažnih ustreznic, saj s preveč pravili hitro najdemo neko sodobno besedo za skoraj poljubno zgodovinsko obliko, zaradi česar je treba nove vzorce dodajati s sprotnim testiranjem njihovega učinka na večji testni množici.

Zadnja od možnosti za izboljšavo sistema bi bila uporaba povsem drugačnega načina posodabljanja, ki je že dalo spodbudne rezultate (Scherrer in Erjavec, 2013), pri katerem učno množico (leksikon iz goo300k) izkoristimo za učenje statističnega strojnega prevajanja na ravni posameznih črk v besedi. Princip strojnega prevajanja bi lahko razširili tudi na prevajanje celotnih besedil, pri čemer bi za učno množico potrebovali izvorno besedilo (ali besedilo, posodobljeno na ravni posameznih besed), ki je poravnano s »prevodom« tega besedila v sodobno slovenščino. S takim pristopom bi lahko zajeli tudi spremembe na skladenjski ravni, vendar je pri tem pristopu največja težava pridobivanje zadosti velike in splošne učne množice.

LITERATURA

[1] Arhar, Š. (2009). Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo, 54(3–4)*, str. 43–56. URL: http://www.jezikinslovstvo.com/pdf/2009-03-04-Razprave-Spela-Arhar.pdf.

[2] Bennett, P., Durrell, D., Scheible, S., Whitt, R. J. (2010). Annotating a historical corpus of German: A case study. *Proceedings of the LREC 2010 workshop on Language Resources and Language Technology Standards*. Valletta, Malta, 18 May 2010. str. 64–68.

[3] Erjavec, T. (2009). Odprtost jezikovnih virov za slovenščino. V: Infrastruktura slovenščine in slovenistike (Obdobja, Simpozij, = Symposium, 28). Ljubljana: Znanstvena založba Filozofske fakultete, str. 115–121. URL: http://www.centerslo.net/files/file/simpozij/simp28/Erjavec.pdf.

[4] Erjavec, T. (2011). Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. V: LaTeCH 2011: The 5th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Portland, ZDA. Portland: Association for Computational Linguistics, str. 33–38. URL: http://aclweb.org/anthology-new/W/W11/W11-1505.pdf.

[5] Erjavec, T. (2012a). Jezikoslovni viri starejše slovenščine. *Knjižnica, 56(3)*, str. 205–221.

[6] Erjavec, T. (2012b). The goo300k corpus of historical Slovene. V: 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 .

[7] Erjavec, T., Džeroski, S. (2004). Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence*, 18(1):17–41.

[8] Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R. (2005). Massive Multi-Lingual Corpus Compilation: Acquis Communautaire and ToTaLe. V: *Proceedings of the 2nd Language & Technology Conference*, April 21-23, 2005, Poznan, Poljska. str. 32–36.

[9] Erjavec, T. in Krek. S. (2008). Oblikoskladenjske specifikacije in označeni korpusi JOS. V: *Zbornik Šeste konference Jezikovne tehnologije*. Ljubljana, Inštitut Jožef Stefan. URL: http://nl.ijs.si/jos/bib/jos_isltc08.pdf.

[10] Gotscharek, A., Neumann, A., Reffle, U., Ringlstetter, C., Schulz, K. U. (2009). Enabling Information Retrieval on Historical Document Collections – the Role of Matching Procedures and Special Lexica. *Proceedings of the ACM SIGIR 2009 Workshop on Analytics for Noisy Unstructured Text Data (AND09),* Barcelona.

[11] Hladnik, M. (2009). Infrastruktura slovenistične literarne vede. V: *Obdobja 28 – Infrastruktura slovenščine in slovenistike*, str. 161–169. URL: http://www.centerslo.net/files/file/simpozij/simp28/Hladnik.pdf.

[12] Juršič, M., Mozetič, I., Erjavec, T., Lavrač, N. (2010) LemmaGen: multilingual lemmatisation with induced Ripple-Down rules. *Journal of universal computing science*. 16/9, str. 1190–1214.

[13] Krstulović, Z. in Šetinc, L. (2005). Digitalna knjižnica Slovenije – dLib.si. *Informatika kot temelj povezovanja: zbornik posvetovanja*, str. 683–689.

[14] Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., Krek, S. (2012) *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES : gradnja, vsebina, uporaba*. (Zbirka Sporazumevanje). Ljubljana: Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede, 2012.

[15] Pilz, T. Ernst-Gerlach, A. Kempken, S., Rayson P., Archer, D. (2008). The Identification of Spelling Variants in English and German Historical Texts: Manual or Automatic? *Literary and Linguistic Computing*, 23/1, str. 65–72.
[16] Ogrin, M., Erjavec, T. (2009). Ekdotika in tehnologija: elektronske znanstvenokritične izdaje slovenskega slovstva. *Jezik in slovstvo, 54/6*, str. 57–72.
[17] Reffle, U. (2011). Efficiently generating correction suggestions for garbled tokens of historical language, Journal of Natural Language Engineering, Special Issue on Finite State Methods and Models in Natural Language Processing.
[18] Sánchez-Marco, C., Boleda, G., Maria Fontana, J., Domingo, J. (2010). Annotation and Representation of a Diachronic Corpus of Spanish. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. ELRA, Pariz.
[19] Scherrer, Y., Erjavec, T. (2013). Modernising historical Slovene words with character-based SMT. *Proceedings of the* ACL Workshop on *Balto-Slavic Natural Language Processing, BSNLP 2013. Sofija, Bulgarija.*
[20] Šorn, M. in Hadalin, J. (2010). Spletni portal SIstory: prost dostop do dosežkov slovenskega zgodovinopisja. *Zbornik prispevkov 4. skupnega posvetovanja Sekcije za specialne knjižnice in Sekcije za visokošolske knjižnice Zveze bibliotekarskih društev Slovenije*, Ljubljana, 27. in 28. oktober 2010, str. 103–107.
[21] TEI (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange. URL http://www.tei-c.org/Guidelines/P5/.

▪

Tomaž Erjavec je višji raziskovalni sodelavec na Odseku za tehnologije znanja na Institutu Jožef Stefan. Področja njegovega raziskovanja so jezikovne tehnologije in digitalna humanistika s poudarkom na izdelavi in označevanju ter predstavitvi jezikovnih virov slovenskega jezika. Na področjih jezikovnih tehnologij in korpusnega jezikoslovja je poučeval na univerzah v Novi Gorici in v Gradcu ter na mednarodni podiplomski šoli Jožefa Stefana. Je član uredniških odborov revij Journal for Language Resources and Evaluation, Journal of Corpus Linguistics in Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave; bil je ustanovni predsednik slovenskega Društva za jezikovne tehnologije, član svetov European Chapter of the Association for Computational Linguistics in Text Encoding Initiative Consortium ter sodeluje pri izdelavi standardov za zapis jezikovnih virov pri SIST in ISO TC 37.