SLOVENSKA PREVODNA KNJIŽEVNOST 1848–1918: DIGITALNA KNJIŽNICA IN KORPUS AHLIB

Tomaž Erjavec
Institut »Jožef Stefan«, Ljubljana

UDK 801.82=03=112.2=163.6"1848/1918":004.91

Digitalna zbirka AHLib vsebuje slovensko prevodno književnost v letih od 1848 do 1918, okoli sto zelo raznovrstnih knjig z nemškimi izvirniki. Zbirka je sestavljena iz skenogramov in čistopisov, zapisana v skladu z mednarodnimi standardi in priporočili, označena tudi strukturno in jezikoslovno ter prosto dostopna. Zato lahko zgodovinoslovcem služi kot digitalna knjižnica, jezikoslovcem pa kot diahroni korpus. V prispevku opišemo izdelavo in zvrstnost zbirke, postavitev zbirke kot spletne knjižnice in kot jezikoslovno označenega korpusa, ki vsebuje sodobne ustreznice, leme in oblikoskladenjske oznake besed ter je dostopen preko spletnega konkordančnika. Zaključimo s smernicami za nadaljnje delo.

zgodovinski jezik, digitalne knjižnice, računalniški korpusi, standardi zapisa

The digital collection AHLib contains around one hundred Slovene books translated from German in the period from 1848 to 1918. The collection is composed of facsimiles and their transcriptions, encoded in line with international standards and recommendations, structurally and linguistically annotated and freely available. It can thus serve as a digital library for historians and as a diachronic corpus for linguists. The paper discusses the making of the collection, its text typology and digital encoding, and its presentation as a digital library and as an annotated corpus, where each word is marked-up with its modern-day equivalent, lemma and morphosyntactic description, and is available via a Web concordancer. The paper concludes with suggestions for further research.

historical language, digital libraries, computer corpora, encoding standards

1 Uvod

Za sodobni slovenski jezik je bilo izdelanih že večje število korpusov, na katerih temeljijo raznovrstne raziskave, doktorske in magistrske disertacije in učbeniki (npr. Gorjanc 2005; Gorjanc, Fišer 2010). Na spletu so preko konkordančnika dostopni referenčni korpusi FidaPLUS (Arhar, Gorjanc 2007), Gigafida in Nova beseda (Jakopin, Bizjak 2007), na portalu SVEZ večjezični Evrokorpus (Željko 2008), na Institutu »Jožef Stefan« (IJS) pa specializirani iKorpus (Erjavec, Vintar 2008).

Za starejši slovenski jezik bi zaman iskali primerljive korpuse in donedavna celo digitalno dostopna besedila, vendar se, vsaj pri slednjih, stanje vedno hitreje spreminja. Starejša besedila so namreč za digitalizacijo in javno dostopnost primerna iz dveh razlogov: po eni strani so jim potekle avtorske pravice, ki so pogosto največja ovira za razširjanje jezikovih virov, po drugi strani pa njihovo digitalizacijo spodbuja financiranje z naslova ohranjanja kulturne dediščine. Evropski projekt EOD tako sofinancira digitalizacijo starejših besedil v okviru portala dLib.si (Krstulović, Šavnik 2011), Ministrstvo za kulturo podpira eBooks on Demand (Kavčič - Ćolić 2010) in dodajanje slovenske književnosti na WikiVir (Hladnik 2009), več raziskovalnih projektov Javne agencije za raziskovalno dejavnost pa izdelavo spletne knjižnice tekst-

no-kritičnih izdaj slovenskega slovstva eZISS (Ogrin, Erjavec 2009).[1]

Digitalizacija starejših besedil s sabo prinese tudi probleme, ki jih pri sodobnih besedilih ne srečamo. Prvi in največji je izdelava čistopisa, saj programi za prepoznavanje besedila iz skenogramov slabo delujejo na starih besedilih, zato je potrebna dolgotrajna in draga ročna redakcija, da dosežemo zadovoljivo kvaliteto besedila. Tudi avtomatsko jezikoslovno označevanje, kot je bilo narejeno npr. za korpus FidaPLUS ter za katerega obstajajo prosto dostopna orodja in označeni korpusi (Erjavec, Krek 2008), za starejšo slovenščino še ni bilo razvito. Korpusno jezikoslovje in jezikovne tehnologije za starejšo slovenščino so tako še v povojih, pri čemer je to zanimivo področje, saj bi lahko korpusne metode prenesli na diahrone študije slovenščine, obenem pa bi lahko izboljšali dostopnost starejših besedil v digitalnih knjižnicah.

V prispevku predstavimo digitalno zbirko starejše slovenščine, ki se je najprej gradila v okviru projekta Deutsch-slowenische/kroatische Übersetzung 1848–1918 (Prunč 2007). Projekt se je začel leta 2004 pod pokroviteljstvom Avstrijske akademije znanosti, po letu 2008 nadaljeval v okviru Univerze v Gradcu (Karl-Franzens-Universität Graz), v zadnjem času pa je bil dodatno podprt s strani projekta EU IMPACT in Googlove nagrade za projekt Developing Models for Historical Slovenian (pri obeh projektih sodeluje IJS). Namen projekta je narediti večji korpus prevodne literature v letih 1848–1918, ki bi vseboval tako nemške izvirnike kot prevode v slovenski in hrvaški jezik. Projekt je prišel najdlje v izdelavi materialov za slovenske prevode, kjer so bili za okoli 200 knjig narejeni skenogrami, skupaj z avtomatsko generiranimi transkripcijami, nato pa so bile transkripcije za okoli sto knjig ročno pregledane in popravljene. Popravki transkripcij ter tudi strukturno in uredniško označevanje so bili izdelani s pomočjo urejevalnika, ki smo ga razvili na IJS (Erjavec 2007). Program omogoča tudi jezikoslovno označevanje, kjer je vsaki besedi mogoče pripisati njeno sodobno lemo.

V prispevku predstavimo zbirko AHLib (Austro-Hungarian Library), kot imenujemo podmnožico besedil, ki so bila digitalizirana v okviru avstrijskih projektov in vsebuje vse knjige, za katere je bila do zdaj končana ročna redakcija. Knjige so dostopne v izvornem zapisu XML (ki je skladen z najnovejšo različico široko uveljavljenih priporočil za zapis besedil Text Encoding Initiative, TEI P5; TEI 2007), na splet so za branje postavljene v formatu HTML, za jezikoslovne študije pa preko konkordančnika, kjer je bila zbirka tudi avtomatsko jezikoslovno označena.

2 Lastnosti zbirke AHLib

AHLib trenutno vsebuje 90 knjig (preko deset tisoč strani oz. dva milijona besed), ki so izšle v letih 1848–1918, torej v času avstrijskega cesarstva in, predvsem, Avstro-Ogrske. Vse knjige so bile prevedene v slovenščino iz nemščine, izbrane pa so bile kot najpomembnejše predstavnice tiste dobe, tako glede na svoj zgodovinski kot jezikovni vpliv. Največ knjig je leposlovnih, večinoma vzgojnih in religioznih (npr. *Zlata Vas*, *Hirlanda bretanjska vojvodinja ali zmaga čednosti in nedolžnosti*), vendar je po številu strani skoraj polovica neleposlovja, predvsem poučnih knjig (npr. *Kratki navod zreji domačih sviloprejk*, *Stelja in gnoj*) in učbenikov (npr. *Knjigovodstvo za dvorazredne trgovske šole*, *Nauk o serviranju*). Pri teh velja posebej izpostaviti obsežne učbenike Friedricha Karla Ludwiga Schödlerja s področja naravoslovja (*Astronomija*, *Fizika*, *Mineralogija*, *Kemija*, *Botanika*, *Zoologija*), ki so bili prevedeni v mnoge evropske jezike, so bogat vir terminologije tistega časa, obenem pa vsebujejo tudi obilo ilustracij.

Digitalna zbirka AHLib je sestavljena iz skenogramov in dokumentov XML. Skenogrami so zapisani v formatu JPEG, za večino knjig so binarizirani (črno-beli) in so na voljo

1 Spletne povezave do posameznih portalov so zbrane na koncu prispevka.

v več velikostih, pri čemer ima največja okoli 1,5 M točk. Ta velikost je primerna za zaslonsko uporabo, žal pa ni dovolj natančna za izdelavo kvalitetnih tiskanih faksimilov. Dokumenti XML obsegajo metapodatke in transkripcije, pri čemer vsaki knjigi ustreza en dokument oz. ena datoteka XML. Metapodatki so sestavljeni po eni strani iz bibliografskih podatkov o publikaciji, po drugi pa iz podatkov o sami digitalni izdaji (npr. o njeni dostopnosti in navedbi odgovornosti za izdajo, seznami uporabljenih oznak v besedilu itn.).

Transkripcija posamezne knjige vsebuje poleg ročno pregledanega besedila tudi razmeroma bogat nabor oznak, tako za strukturo knjige (prelomi strani, razdelki in naslovi, opombe itn.) kot tudi za osnovne jezikoslovne in uredniške oznake (npr. oznake za tujejezične dele besedila in uredniške popravke tiskarskih škratov v izvirniku). Ravno tako so označena mesta slik, s čimer je možno enostavno identificirati vse skenograme, ki vsebujejo slike. Kot poseben tip slike (»ornament«) so označena tudi okrasna znamenja na straneh (npr. secesijske vinjete).

Shema XML poda gramatiko oznak, ki se uporabljajo v zbirki, in je skladna s priporočili za označevanje besedil TEI P5 (TEI 2007) in je sestavni del zbirke. TEI P5 definira pomensko usmerjene oznake XML in se uporablja pri večini digitalnih knjižnic, ki vsebujejo bogato strukturirana besedila. Priporočila TEI so javno dostopna, dobro dokumentirana, obstaja pa tudi programska oprema, ki omogoča obdelavo dokumentov na različne načine. V zapisu TEI je vsaka knjiga znotraj zbirke AHLib predstavljena kot svoj dokument; sestavljen je iz kolofona TEI, ki vsebuje metapodatke, razdelka s kazalkami na skenograme in besedila s transkripcijo. Besedilo je strukturirano v hierarhične razdelke, znotraj njih na naslove, odstavke, opise slik itn. Transkripcija vsebuje še nadaljnje oznake (ilustracije itn.); transkripcije AHLib skupaj vsebujejo preko 30 različnih oznak. Primer zapisa transkripcije v TEI P5 je podan na Sliki 1.

```
<pb n="2" facs="#FPG00279-005" xml:id="pb.006"/>
<figure xml:id="figure.1">
 <figDesc>Doprsni portret cesarja Maksimilijana I.</figDesc>
 <head>Pod. 1. Cesar Maksimilijan I.</head>
</figure>
 <p xml:id="p.16">Kaj rad je zahajal cesar k ljudskim veselicam, zakaj povsod so ga ljubili in visoko
čislali. Ko se je nekoč mudil v Nürnbergu, so mu mestni svetovavci skrili črevlje in ostroge, hoteč ga
                                                    nekako
    <pb n="3" facs="#FPG00279-006" xml:id="pb.007"/>
    prisiliti, da bi ostal še dalje pri njih. Ta šala je Maksimilijanu prijala, in ostal je še nekoliko v
gostoljubnem
...
izboljšal je strelila in osnoval v Inomostu, Gradcu, Gorici in na Dunaju orožnice. Topovom je nadeval
imena, kakor <foreign>«Weck' auf»</foreign> ali <foreign>«Burlebaus».</foreign> Osnoval je tudi
pehoto najemnikov (Pod. 2.) ter jo opremil z oklopi, kopji, dolgimi meči in semtertja tudi s
puškami.</p>
          <figure xml:id="figure.2">
            <figDesc>Pehota cesarjeve vojske</figDesc>
            <head>Pod. 2. Pešci.</head>
          </figure>
```

Slika 1: AHLib v izvornem formatu XML TEI P5[2]

2 Element <pb> zaznamuje prelom strani in vsebuje kazalko na skenograme, slike imajo opis in izvirni naslov, besedilo je razdeljeno po odstavkih, označeni so tudi tujejezični vrivki.

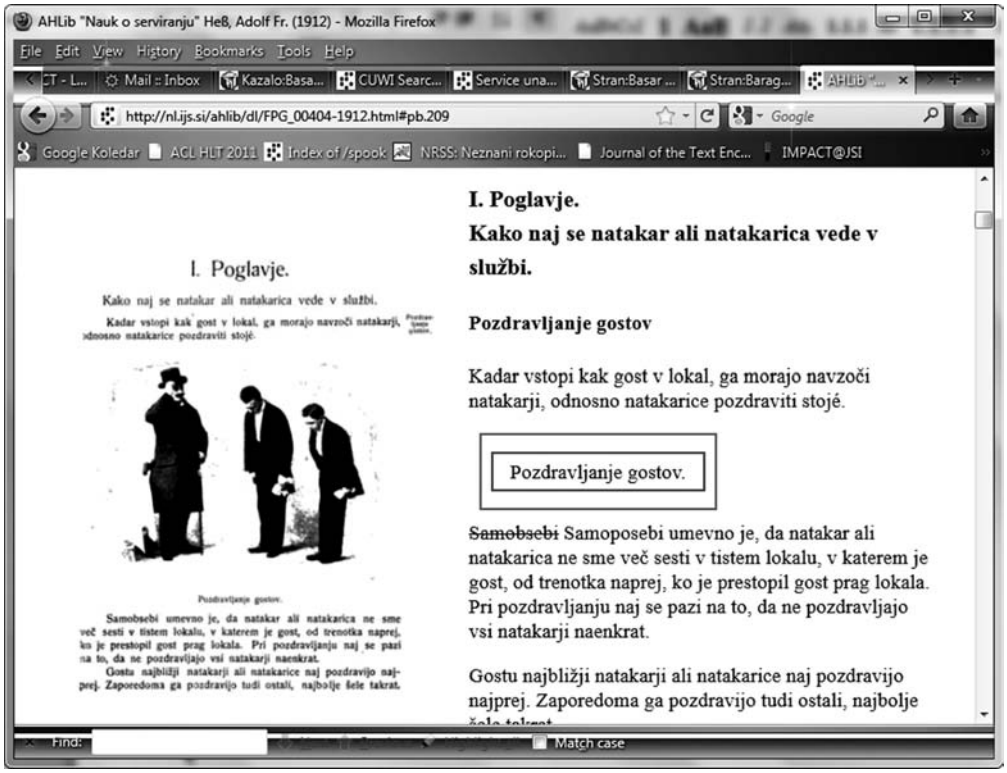
3 AHLib kot digitalna knjižnica

Iz kanoničnega zapisa zbirke v XML je z ustrezno programsko opremo mogoče avtomatsko izvesti poljuben predstavitveni format, prirejen za tisk ali splet. Trenutno je na voljo pretvorba v statični HTML, ki vzporedno prikaže faksimile in prepis. Spletna oblika vsake knjige vsebuje kazalo po straneh in po poglavjih ter kazalki na kolofon TEI, kjer so imena elementov kolofona bodisi v slovenskem bodisi v angleškem jeziku. Pretvorba v HTML je narejena s pomočjo standardnega TEI-paketa XSLT-transformacij, ki so bile za zbirko AHLib nadgrajene, da npr. prikažejo kolofon TEI, vzporedni faksimile in transkripcijo. Primer spletne predstavitve je podan na Sliki 2.

V trenutni izdaji so tako vse knjige zbirke na voljo za branje, bodisi kot faksimile bodisi kot stavljena besedila. Seveda bi bilo mogoče spletno postavitev še bistveno izboljšati, npr. omogočiti iskanje po polnem besedilu skozi poizvedovalni sistem, kot je Fedora/Solr, kar je bilo že implementirano za *Slovenski biografski leksikon* (Ogrin idr. 2010), oz. omogočiti samo izpis besedila, kar je predvsem primerno za platforme z omejenimi možnostmi prikaza in kjer bi radi minimizirali količino prenesenih podatkov, npr. za mobilne telefone.

4 AHLib kot korpus

Izvorni namen digitalne zbirke je bil, da služi kot (vzporedni) korpus za prevodoslovne



Slika 2: AHLib v spletni predstavitvi[3]

3 Predstavitev v HTML vzporedno prikaže faksimile in transkripcijo; transkripcija kaže na mesto ilustracij, pa tudi na uredniške posege v besedilu.

```
<pc ctag="&#34;">„</pc>
<w nform="še" lemma="še" ctag="Q">Še</w><c> </c>
<w lemma="en" ctag="Mlpfsn">ena</w><c> </c>
<w lemma="drug" ctag="Mlpfsn">druga</w><c> </c>
<w lemma="žalost" ctag="Ncfsn">žalost</w><c> </c>
<w lemma="ti" ctag="Pp2-sd--y">ti</w><c> </c>
<w lemma="biti" ctag="Va-r3s-n">je</w><c> </c>
<w subtype="pattern" pattern="[ri←er+lj←l]" mform="pripravljena" lemma="pripravljen"
    ctag="Appfsn">perpravlena</w>
<pc ctag=".">.</pc>
</s>
<c> </c>
<s>
<w nform="per" lemma="per" ctag="Xf">Per</w><c> </c>
<w subtype="lexicon" mform="našemu" lemma="naš" ctag="Ps1msdp">našimu</w><c> </c>
<w lemma="sosed" ctag="Ncmsd">sosedu</w>
...
```

Slika 3: Jezikoslovno označen korpus AHLib v XML TEI P5[4]

raziskave, ki bi ga bilo treba zato tudi jezikoslovno označiti. Že spletni vmesnik za izdelavo transkripcij podpira osnovno jezikoslovno analizo na ravni besed in združevanje ročno popravljenih lematizacij nazaj v korpus. Na ta način je bilo preko sodelavcev projekta Avstrijske akademije znanosti do zdaj lematiziranih 14 knjig (1200 strani, 250.000 besed), v postopku lematizacije pa je bil vzporedno grajen tudi leksikon parov historična besedna oblika – sodobna lema, ki vsebuje čez 50.000 vnosov.

V zadnjem času smo implementirali program ToTrTaLe (Erjavec 2011) za jezikoslovno označevanje starejše slovenščine. Program skuša najprej posodobiti starinske besede, nato pa uporabi standardne jezikovne modele za sodobno slovenščino, s katerimi oblikoskladenjsko označi besede in jih lematizira. Pri koraku posodabljanja je odvisen od leksikona starejšega jezika, ki vsebuje sodobne ustreznice, in vzorcev za transkripcijo besed iz starih v sodobne oblike; oboje je še v gradnji, kljub temu pa so njegovi rezultati že zdaj razmeroma uporabni (odvisno od starosti besedila). S pomočjo programa ToTrTaLe smo avtomatsko označili celotno zbirko AHLib. Na Sliki 3 podamo primer jezikoslovno označenega besedila v formatu XML TEI P5.

5 Dostopnost zbirke

Dostopnost jezikovnih virov ima več dimenzij (Erjavec 2009); idealno so besedila dostopna tako za pregledovanje kot za prenos, zapisana po mednarodnih standardih in priporočilih, pogoji uporabe pa dovoljujejo nadaljnje predelave in razširjanje.

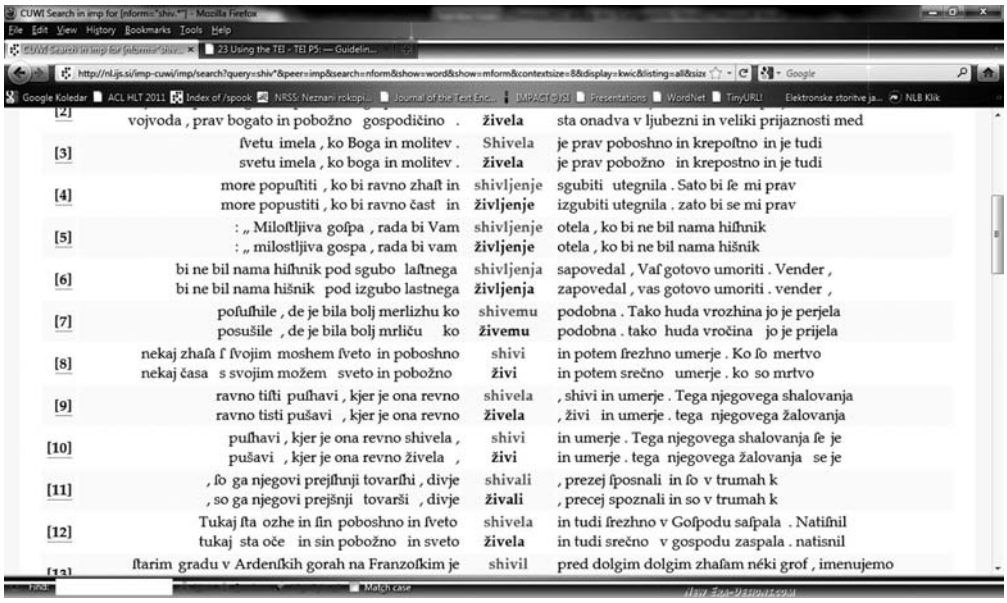
Zbirka AHLib je dostopna v okviru spletne knjižnice TraDok na Oddelku za prevodoslovje Univerze v Gradcu. Tam so na voljo natančni metapodatki o vsaki knjigi, PDF izvornih skenogramov ter (za večino del, kjer je bila korekcija že opravljena) tudi

4 Element <pc> zaznamuje ločilo, <c> presledek in <w> besedo. Besede so označene z normalizirano in posodobljeno obliko (kjer se razlikujeta od besede) ter lemo in oblikoskladenjsko oznako. Besede so lahko posodobljene preko leksikona ali vzorcev, če slednje, so ji pripisani še vzorci. Program dela tudi napake, npr. beseda »Per« je napačno identificirana kot tuja beseda.

transkripcij RTF, XML in HTML. Trenutno so dokumenti v formatu XML na voljo samo v stari različici TEI, TEI P4, saj to inačico podpira spletni program za korekcije transkripcij in strukturno označevanje. Dostop do zbirke AHLib smo omogočili tudi na spletnih straneh IJS, kjer je zbirka dostopna preko spletišča, posebej posvečenega tej zbirki. Za to različico smo popravili določene napake v transkripcijah, dokumenti so bili iz TEI P4 pretvorjeni v TEI P5, poleg tega pa smo jim dodali vse metapodatke iz knjižnice TraDok. Tudi pretvorba v HTML je bila posodobljena, tako da bolje prikaže poravnavo faksimila s transkripcijo. Knjižnica AHLib je dostopna tako v formatu HTML za branje kot tudi v izvornem formatu XML TEI P5 za prenos in nadaljnje obdelave oz. analize. Dostopnost zbirke ureja licenca Creative Commons, Priznanje avtorstva 3.0, kar pomeni, da je vsakomur dovoljeno kopirati, predelati in razširjati zbirko v poljubne namene, pod pogojem, da navede avtorje izvorne zbirke. V znanstvenih publikacijah to pomeni citiranje dela ali del, ki opisujejo zbirko AHLib.

Na Univerzi v Gradcu je korpus dostopen tudi preko konkordančnika, postavljenega v okviru sistema GAMS, ravno tako pa preko konkordančnika na spletišču IJS. Ta omogoča iskanje po izvornem besedilu oz. njegov izpis (normalizirane in posodobljene različice besedila) ali pa po lemah oz. oblikoskladenjskih oznakah. Podpira več vrst izpisa (npr. konkordanc in frekvenčnega slovarja) in sortiranje po zadetkih oz. levem ali desnem kontekstu, pri čemer je za vsak zadetek mogoče priklicati skenogram, kjer se pojavlja. Za ilustracijo prikažemo enega od možnih izpisov konkordančnika na Sliki 4.

Jezikoslovno označen korpus AHLib za zdaj še ni dostopen v izvorni obliki XML, predvsem zaradi dejstva, da je treba avtomatsko označevanje še izboljšati. Ko bo program ToTrTaLe (predvsem pa jezikovni viri, ki jih uporablja) razvit do te mere, da bo dovolj kvalitetno označeval starejšo



[2]	vojvoda , prav bogato in pobožno gospodičino .	**živela**	sta onadva v ljubezni in veliki prijaznosti med
[3]	ſvetu imela , ko Boga in molitev .	Shivela	je prav poboshno in krepoſtno in je tudi
	svetu imela , ko boga in molitev .	**živela**	je prav pobožno in krepostno in je tudi
[4]	more popuſtiti , ko bi ravno zhaſt in	shivljenje	sgubiti utegnila . Sato bi ſe mi prav
	more popustiti , ko bi ravno čast in	**življenje**	izgubiti utegnila . zato bi se mi prav
[5]	: „ Miloſtljiva goſpa , rada bi Vam	shivljenje	otela , ko bi ne bil nama hiſhnik
	: „ milostljiva gospa , rada bi vam	**življenje**	otela , ko bi ne bil nama hišnik
[6]	bi ne bil nama hiſhnik pod sgubo laſtnega	shivljenja	sapovedal , Vaſ gotovo umoriti . Vender ,
	bi ne bil nama hišnik pod izgubo lastnega	**življenja**	zapovedal , vas gotovo umoriti . vender ,
[7]	poſuſhile , de je bila bolj merlizhu ko	shivemu	podobna . Tako huda vrozhina jo je perjela
	posušile , de je bila bolj mrliču ko	**živemu**	podobna . tako huda vročina jo je prijela
[8]	nekaj zhaſa ſ ſvojim moshem ſveto in poboshno	shivi	in potem frezhno umerje . Ko ſo mertvo
	nekaj časa s svojim možem sveto in pobožno	**živi**	in potem srečno umerje . ko so mrtvo
[9]	ravno tiſti puſhavi , kjer je ona revno	shivela	, shivi in umerje . Tega njegovega shalovanja
	ravno tisti pušavi , kjer je ona revno	**živela**	, živi in umerje . tega njegovega žalovanja
[10]	puſhavi , kjer je ona revno shivela ,	shivi	in umerje . Tega njegovega shalovanja ſe je
	pušavi , kjer je ona revno živela ,	**živi**	in umerje . tega njegovega žalovanja se je
[11]	, ſo ga njegovi prejſhnji tovarſhi , divje	shivali	, prezej ſposnali in ſo v trumah k
	, so ga njegovi prejšnji tovarši , divje	**živali**	, precej spoznali in so v trumah k
[12]	Tukaj ſta ozhe in ſin poboshno in ſveto	shivela	in tudi frezhno v Goſpodu saſpala . Natiſnil
	tukaj sta oče in sin pobožno in sveto	**živela**	in tudi srečno v gospodu zaspala . natisnil
[13]	ſtarim gradu v Ardenſkih gorah na Franzoſkim je	shivil	pred dolgim dolgim zhaſam néki grof , imenujemo

Slika 4: Izpis konkordančnika[5]

5 Konkordance so iz edine knjige v zbirki AHLib, ki je zapisana v bohoričici. Izpisano je izvorno besedilo in posodobljene oblike besed.

slovenščino, bo postal tudi ta korpus javno dostopen.

6 Zaključki

V prispevku smo predstavili digitalno zbirko AHLib, ki vsebuje prevedene knjige iz obdobja Avstro-Ogrske in je prva večja, tematsko usmerjena, ročno pregledana, jezikoslovno označena in prosto dostopna zbirka starejšega slovenskega slovstva. Zbirka lahko služi kot digitalna knjižnica za zgodovinoslovje in kot diahroni korpus za jezikoslovne študije.

V nadaljnjem delu bomo zbirko razširili z dodatnimi knjigami, ki so sicer že bile skenirane, vendar transkripcije še niso bile ročno popravljene. Transkripcije nastajajo deloma na Univerzi v Gradcu, deloma pa v okviru projekta EU IMPACT, znotraj katerega Narodna in univerzitetna knjižnica izdeluje čistopise 5000 strani starejših besedil, med njimi tudi še nepregledanih knjig iz fonda AHLib (Erjavec, Kodrič 2011). Kot je bilo že omenjeno, načrtujemo tudi izboljšanje programa ToTrTaLe, predvsem skozi izboljšanje jezikovnih virov, na katerih temelji. Prvi tak vir je leksikon starejše slovenščine, ki nastaja na IJS, ravno tako v okviru projekta IMPACT (Erjavec idr. 2010), drugi pa jezikoslovno označen in ročno pregledan korpus starejše slovenščine, ki nastaja v sodelovanju z ZRC SAZU v okviru projekta Jezikoslovni modeli za starejšo slovenščino, ki ga je finančno podprl Google.

Zahvala

Delo na projektu so podprli projekt Deutsch-slowenische/kroatische Übersetzung 1848–1918 Avstrijske akademije znanosti, projekt EU 7. OP Improving Access to Text in nagrada Google za raziskave s področja digitalne humanistike za projekt Developing Models for Historical Slovenian.

Literatura

ARHAR, Špela, GORJANC, Vojko, 2007: Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2. 95–110.

ERJAVEC, Tomaž, 2007: Architecture for Editing Complex Digital Documents. Sanja Seljan, Hrvoje Stančić (ur.): *Zbornik konference INFuture 2007: Digital Information and Heritage*. Zagreb. 105–114.

ERJAVEC, Tomaž, 2009: Odprtost jezikovnih virov za slovenščino. Marko Stabej (ur.): *Infrastruktura slovenščine in slovenistike. Obdobja 28.* Ljubljana: Znanstvena založba Filozofske fakultete. 115–121.

ERJAVEC, Tomaž idr., 2010: Towards a lexicon of XIXth Century Slovene. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Sedme konference Jezikovne tehnologije, 14.–15. oktober 2010*. Ljubljana: Institut »Jožef Stefan«. 68–73.

ERJAVEC, Tomaž, 2011: Automatic Linguistic Annotation of Historical Language: ToTrTaLe and XIX Century Slovene. Kalliopi Zervanou, Piroska Lendvai (ur.): *Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities: Proceedings of the Workshop*. Portland, Oregon: The Association for Computational Linguistics.

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.

ERJAVEC, Tomaž, KREK, Simon, 2008: Oblikoskladenjske specifikacije in označeni korpusi JOS. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Šeste konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 49–53.

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

GORJANC, Vojko, FIŠER, Darja, 2010: *Korpusna analiza*. Ljubljana: Znanstvena založba Filozofske fakultete.

HLADNIK, Miran, 2009: Infrastruktura slovenistične literarne vede. Marko Stabej (ur.): *Infrastruktura slovenščine in slovenistike. Obdobja 28*. Ljubljana: Znanstvena založba Filozofske fakultete. 161–69.

JAKOPIN, Primož, BIZJAK, Aleksandra, 1997: O oblikoslovnem označevanju slovenskega besedila. *Slavistična revija* 45/3–4. 513–532.

KAVČIČ - ĆOLIĆ, Alenka, 2010: *Achieving Library Development Through European Projects: the Case of the National and University Library of Slovenia*. www.hkdrustvo.hr/hr

KRSTULOVIĆ, Zoran, ŠAVNIK, Mojca, 2011 (v tisku): Library Material Digitisation Project Coordination and e-Content National Aggregator in the Field of Culture. *Zbornik šeste konference SEEDI »Digitization of Cultural and Scientific Heritage«*. Zagreb.

OGRIN, Matija, ERJAVEC, Tomaž, 2009: Ekdotika in tehnologija: elektronske znanstvenokritične izdaje slovenskega slovstva. *Jezik in slovstvo* 54/6. 57–72.

PRUNČ, Erich, 2007: Deutsch-slowenische/kroatische Übersetzung 1848–1918. Ein Werkstättenbericht. *Wiener Slavistisches Jahrbuch* 53/2007. Dunaj: Austrian Academy of Sciences Press. 163–176.

TEI CONSORTIUM, 2007: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. www.tei-c.org/release/doc/tei-p5-doc/en/html

VIDE OGRIN, Petra idr., 2010: Slovenian Biographical Lexicon – the Nucleus of a Biographical Hub. *Personen – Daten – Repositorien*. Berlin.

VINTAR, Špela, ERJAVEC, Tomaž, 2008: iKorpus in luščenje izrazja za Islovar. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Šeste konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 65–69.

ŽELJKO, Miran, 2009: Pomnilniki prevodov EU. *Mostovi* 2009/10.

Spletni viri

AHLib – domača stran spletne zbirke: http://nl.ijs.si/ahlib

TraDok – digitalna knjižnica Translation-Dokumentation: http://itat2.uni-graz.at/pub/tradok

GAMS – konkordančnik AHLib, Unverza v Gradcu: www-gewi.uni-graz.at/cocoon/ahlib

TEI – konzorcij za označevanje besedil Text Encoding Initiative: www.tei-c.org

CC – Creative Commons: http://creativecommons.org

JOS – rezultati projekta Jezikoslovno označevanje slovenščine: http://nl.ijs.si/jos

FidaPLUS – referenčni korpus sodobnega slovenskega jezika: www.fidaplus.net

Gigafida – referenčni korpus sodobnega slovenskega jezika: http://demo.gigafida.net

Nova beseda – besedilni korpus na ZRC SAZU: http://bos.zrc-sazu.si/s_beseda.html

iKorpus – korpus informatike in računalništva: http://nl2.ijs.si/dsi.html

Evrokorpus – večjezični vzporedni korpus pravne terminologije EU: http://evrokorpus.gov.si

*dLib.si – digitalna knjižnica Slovenij*e: www. dlib.si

EOD – eBooks on Demand: www.books2ebooks.eu

eZISS – Elektronske znanstvenokritične izdaje slovenskega slovstva: http://nl.ijs.si/e-zrc

WikiVir – Slovenska leposlovna klasika na spletu: http://sl.wikisource.org/wiki/Wikivir:Slovenska_leposlovna_klasika