IZDELAVA KORPUSA STAREJŠIH SLOVENSKIH BESEDIL V OKVIRU PROJEKTA IMPACT

Tomaž Erjavec
Institut »Jožef Stefan«, Ljubljana

Ines Jerele, Maša Kodrič
Narodna in univerzitetna knjižnica, Ljubljana

UDK 801.8=163.6"17/18":004.91

Institut »Jožef Stefan« in Narodna in univerzitetna knjižnica Ljubljana od leta 2010 sodelujeta pri evropskem projektu IMPACT (Improving Access to Text), katerega cilj je razviti tehnologije, ki bodo uporabniku in bralcu omogočale uspešnejši dostop do polnega besedila digitaliziranih starejših tiskanih besedil v slovenskem jeziku. V članku predstavimo postopek izdelave korpusa slovenskih besedil in izpostavimo težave, na katere smo naleteli med delom in ki med drugim izvirajo iz zgodovinskih in strukturnih značilnosti slovenskega jezika in črkopisa, ki smo ga uporabljali v preteklosti.

digitalizacija, OCR, starejša tiskana slovenska besedila, arhaični slovenski jezik, projekt IMPACT

Since 2010 the Jožef Stefan Institute and the National and University Library have been collaborating on the project EU IMPACT (Improving Access to Text), which has as its goal the development of technologies that will enable better full-text access to digitised printed historical Slovene texts. The paper presents the work-flow of the corpus compilation and highlights the problems that we have had to face, which stem also from the historical and structural characteristics of Slovene.

digitisation, OCR, historical printed Slovene texts, historical Slovene, IMPACT project

1 Uvod

Slovenska kulturna dediščina je že dobrih pet let dosegljiva tudi v digitalni obliki prek portala Digitalne knjižnice Slovenije (dLib)[1] ter med drugim obsega tudi starejše tiskane knjige in časopise. Digitalizirano tiskano gradivo je na portalu dosegljivo v obliki faksimilov v formatu PDF, kar uporabnikom omogoča, da si gradivo ogledajo ali shranijo na svojem računalniku. Za uporabnike in raziskovalce je veliko bolj prijazna in uporabna možnost HTML-predogleda polnega besedila v elektronski obliki. Taka »odprta« oblika omogoča vpogled v samo besedilo, saj nudi možnosti iskanja, urejanja in besedilnih analiz, obenem pa je tudi prostorsko bistveno manj potratna in primerna za predstavitev na raznovrstnih platformah, npr. na mobilnih telefonih. Uporaba polnega besedila je smiselna in večinoma uspešna pri digitaliziranem tiskanem gradivu iz obdobja po letu 1850. Težave se pojavijo pri polnem besedilu starejših tiskov, saj so besedila večinoma netočna, kar zmanjšuje njihovo dostopnost in uporabnost za nadaljnje raziskovanje. Sodobni programi za optično prepoznavo znakov (OCR)[2] namreč niso sposobni uspešno izvesti »branja« starih tiskov in besedil v zadovoljivi meri. Raznovrstna narava gradiva, tiskanega od Gutenbergovega izuma sredi 15. stoletja pa vse do industrijske izdelave sredi 19. stoletja, je še vedno prezahtevna, da bi se z njo

1 Seznam spletnih virov je naveden na koncu prispevka.

lahko uspešno spoprijela sodobna orodja OCR. Ročno popravljanje takega besedila je za vsa gradiva zelo drago, zamudno in povsem nekonkurenčno za potrebe masovne digitalizacije[3] v 21. stoletju.

Evropske knjižnice, muzeji in arhivi so ob množični digitalizaciji historičnih tekstov naleteli na mnoge težave. Kot največji sta se izkazali 1) neuspešna pretvorba digitaliziranih strani knjig in časopisov v polni tekst s pomočjo OCR in 2) historična oblika jezikov ter besedišče besedil, ki so bila izdana pred letom 1850. Historična oblika jezikov je namreč tuja obstoječi leksiki, ki je kot podpora vgrajena v današnja orodja OCR in izhaja iz sodobnega besedišča jezikov, obenem pa problematizira iskanje po polnem besedilu, saj se uporabniki praviloma ne zavedajo arhaičnih zapisov besed.

Zato se je pojavila potreba po boljši programski opremi OCR in tehnologijah, ki bi lahko uspešno in z veliko natančnostjo tudi starejša besedila pretvorile v obliko, ki bi bila prijazna do uporabnika in uporabna tudi za nadaljnje raziskovalne postopke. Da bi uspešno premestili to težavo, so v začetku leta 2008 v okviru projekta IMPACT (Improving Access to Text) združili svoje izkušnje, znanje in moči sedem evropskih knjižnic, šest raziskovalnih inštitutov in dve podjetji. Cilj projekta je izboljšati optično prepoznavo starejših besedil evropske tiskane dediščine, tako na ravni posameznih znakov kot na ravni identifikacije strukture besedil.

Delo na projektu, ki je v večini financiran s strani Evropske komisije, je v letu 2010 preraslo osnovne okvire, zato se je prvotnemu konzorciju pridružilo še enajst partnerjev iz petih evropskih držav, med njimi tudi Narodna in univerzitetna knjižnica (NUK) in Institut »Jožef Stefan« (IJS). Projekt se bo zaključil konec leta 2011.

V prispevku predstavljamo svoje delo na izgradnji referenčne podatkovne množice – korpusa v okviru projekta IMPACT, ki vključuje skenograme približno 5000 strani starejših slovenskih tiskanih besedil, skupaj s segmentacijo posameznih strani na (strukturna) območja in ročno pregledanimi transkripcijami. Transkripcije sledijo izvirniku z vsaj 99,95-odstotno natančnostjo.[4]

Korpus bo služil kot testna in učna množica za programe OCR, pa tudi za namene razvoja jezikovno-tehnološke podpore obdelavi starejše slovenščine (predvsem kot vir leksike, skupaj s primeri uporabe) in kot osnova za izdelavo jezikoslovno označenega korpusa. Izboljšanje tehnologij OCR in jezikovnih tehnologij za starejšo slovenščino bo po eni strani omogočilo izdelavo boljših avtomatskih transkripcij, po drugi pa oblikosklad enjsko označevanje, lematizacijo in posodabljanje besednih oblik, kar olajša tako iskanje po polnem besedilu kot tudi njegovo razumevanje ter odpira možnosti za empirično podprte diahrone jezikoslovne raziskave.

2 Gradivo in njegove posebnosti

Kriteriji za izbor besedil so bili predvsem jezikovni in zgodovinski, pri čemer smo si prizadevali za raznolikost besedil po zgradbi, namembnosti in besedišču. Zato v izboru najdemo različne oblike in tipe besedil: od romanov, pesmi, dramskih tekstov do slovnic,

2 Optična prepoznava znakov je »postopek pretvorbe bitne slike besedila v besedilo, ki ga je mogoče obdelovati v urejevalniku besedil« (iSlovar). Omogoča nam, da tiskane knjige in revije pretvorimo v elektronsko obliko, v kateri lahko besedilo poljubno urejamo. Programska oprema OCR se uporablja za različne namene in v različnih strokah ter je postala nepogrešljiva pri digitalizaciji tiskane dediščine.

3 Masovna digitalizacija je izraz, ki se uporablja v bibliotekarski stroki in označuje digitalizacijo večjih količin gradiva naenkrat.

4 99,95-odstotna natančnost na znakovni ravni pomeni, da mora v celotni podatkovni množici, ki jo izdelamo, 99,95 odstotkov znakov ustrezati izvirniku. 0,05-odstotno odstopanje je dovoljeno zaradi človeškega faktorja, saj tudi pri ročnem popravljanju dokumentov včasih spregledamo kakšno napako.

pridig, kuharskih knjig, pratik, ugank in seveda časopisov.

Gradivo je sestavljeno iz treh sklopov. V prvem je 15 knjig (okoli 2250 strani) iz dLib, posebno pozornost smo namenili izbiri pomembnih besedil slovenske kulturne dediščine iz poznega 18. in začetka 19. stoletja, tj. iz obdobja slovenskega razsvetljenstva, ko so se oblikovali temelji sodobnega slovenskega jezika. Med njimi najdemo besedila, kot so Pohlinove *Kratkozhasne uganke*, *Branja, inu evangeliumi*, *Abecedika* ali *Plateltof*, *Glossarium slavicum* in Japljeve *Pridige*, Vodnikove *Kuharske bukve* in Linhartova *Županova Micka*.

Drugi sklop, ravno tako iz dLib, obsega vzorce iz 47 letnikov (1843–1890) izdaj časopisa *Kmetijske in rokodelske novice*[5] (590 strani), ki med drugim vsebujejo besedila z obdobja črkarske pravde in romantike. Pri tem nismo vključevali celotnih letnikov časopisa, saj bi jih bilo preveč, pač pa smo vzorčili vsak posamezen letnik po izdajah, tako da smo vzeli celotni prvi letnik časopisa, ki je zapisan v bohoričici, nato pa vsako leto manj izdaj.

Tretji sklop obsega 15 knjig (1900 strani) iz druge polovice 19. stoletja, ki so prevedene iz nemškega jezika in so del digitalne zbirke AHLib[6] ter še niso bile deležne korektur OCR. Ta sklop vsebuje npr. takrat zelo popularne zgodbe Christopha von Schmida, prvi prevod romana *Stric Tomaž* Harriet Beecher - Stowe, pa tudi neleposlovna dela, kot je *Rudninoslovje* Sigmunda Fellöckerja.

Na podlagi smernic IMPACT smo tako zbrali skupno 4983 digitaliziranih strani, ki ustrezajo vsem tehničnim zahtevam in določilom ter tvorijo izbor ročno pregledanega slovenskega gradiva.

Slovenska historična besedila zaznamuje kar nekaj posebnosti na ravni strukture posameznih strani, besedila, tiska (historični fonti in tipografije), črkopisa in seveda jezika. V 19. stoletju so bili v uporabi štirje črkopisi (bohoričica, metelčica, dajnčica in sodobna gajica), pri čemer smo v korpus zajeli besedila glavnih dveh, tj. bohoričice, ki se je uporabljala približno do leta 1845,[7] in gajice, ki jo uporabljamo še danes.

Ker so uporabljeno gradivo skenirali različni izvajalci digitalizacijskih storitev v različnih obdobjih, so skenogrami, ki smo jih uporabili v postopku transkripcije, po kvaliteti in tehničnih značilnostih zelo različni. Razlikujejo se predvsem v barvni globini (barvni, sivinski, binarni), nekateri so enostranski, drugi dvostranski, starejši so večinoma slabe kvalitete (narejeni v nizki resoluciji), razlikujejo pa se tudi v formatih (JPEG, PDF in TIFF). Ker je veliko digitaliziranega gradiva, ki smo ga vključili v izbor, v slabem fizičnem stanju (to še posebej velja za časopise in nekatere knjige), digitalizacije nismo ponavljali, temveč smo uporabili že obstoječe, čeprav slabo kakovostne skenograme. Posledica slabe kvalitete skenogramov je bil tudi slab OCR in več ročnega popravljanja. Težav pri delu pa nista povzročili le različnost in kvaliteta skenogramov, temveč tudi fizično stanje gradiva, ki je poslabšalo digitalizacijske rezultate (slaba kvaliteta papirja, transparentnost papirja, neenakomerno porazdeljeno črnilo/tisk, prepognjene/zapognjene strani, madeži na papirju zaradi staranja). Težave so povzročali tudi tipografija (latinica in gotica, spreminjajoča se velikost fonta), nepravilni razmiki med črkami, besedami in vrsticami, pogosta uporaba tabel in stolpcev, struktura časopisnih strani, uporaba

5 Časopis je začel izhajati 1843 in se je najprej imenoval *Kmetijske in rokodelske novice*, v letih 1849–1852 *Novice kmetijskih, rokodelnih in narodskih reči*, nato pa *Novice gospodarskih, obertnijskih in narodskih stvari*. V prispevku ga ne glede na letnico izida imenujemo kar s prvim imenom.

6 AHLib zbirka digitaliziranih strani in transkripcij je nastala v okviru projektnega sodelovanja Avstrijske akademije znanosti in IJS (Prunč 2004; Erjavec 2007).

7 Še posebej zanimiv primer so *Kmetijske in rokodelske novice*, kjer najdemo na isti strani članke v gajici in bohoričici (leto 1844, št. 33, 14. veliki serpan).

dveh črkopisov in danes arhaičnega jezika (bohoričica, gajica, posebni znaki, naglasi in ligature ter slaba prepoznava besed).

3 Delovni postopek izdelave čistopisov in segmentacija strani

V prvem koraku priprave korpusa smo zbrali skenograme ter vse pretvorili v enotni zapis TIFF, dvostranske skenograme smo pretvorili v enostranske in poenotili poimenovanje datotek. Nato so bili skenogrami shranjeni na skupnem repozitoriju projekta IMPACT, kjer je bila vsakemu pripisana tudi identifikacijska številka. Naslednji korak je bilo segmentiranje strani in izdelava transkripcij.

Naše prepričanje, da je za dober OCR odločilno poznavanje jezika, v katerem je zapisano besedilo, so podprli tudi rezultati in analiza testne serije besedil. Rezultati testne serije, ki smo jo izdelali v NUK, so se izkazali za nadpovprečne, saj so bili zelo dobri v primerjavi z rezultati drugih partnerjev v projektu. Zato smo se v NUK odločili za razvoj lastnega delovnega postopka, ki se razlikuje od siceršnje prakse partnerjev IMPACT. Večini partnerjev na osnovi vhodnih zapisov, ki jih priskrbi Univerza v Salfordu (USAL), izdelajo končne zapise zunanji izvajalci, izvedbeno kakovost pa zagotavljajo partnerji projekta, zadolženi za evalvacijo in pregled primarne podatkovne množice. V NUK pa celoten postopek do evalvacije opravimo sami, in sicer vzporedno izvajamo izdelavo vhodnih (OCR) in končnih (ročno pregledanih) zapisov.[8] Pri tem uporabljamo programsko opremo OCR (Abbyy FineReader), ki je prilagojena slovenskemu jeziku.[9] Dosedanji rezultati izpolnjujejo zelo visoko raven natančnosti, naš delovni postopek pa se vse bolj izkazuje kot primer dobre prakse in je bil kot tak tudi že predstavljen v okviru projektne skupine IMPACT.

Cilj delovnega postopka je bil izdelati 5000 strani transkripcij v obliki PAGE XML (Pletschacher, Antonacopoulos 2010), ki je nadgradnja standardnega formata za analizo oblike in besedilnih elementov skenogramov ALTO (Analyzed Layout and Text Object).

Proces priprave zapisov za primarno podatkovno množico je potekal sočasno na dveh ravneh:

S programom Abbyy FineReader je bil za vsak posamezen skenogram (tj. za vsako posamezno stran knjige ali časopisa posebej) narejen OCR in shranjen kot golo besedilo (format txt) ali v programu Microsoft Word (format doc). Besedilo smo v programu Microsoft Word ročno popravili, kolikor je bilo to mogoče.

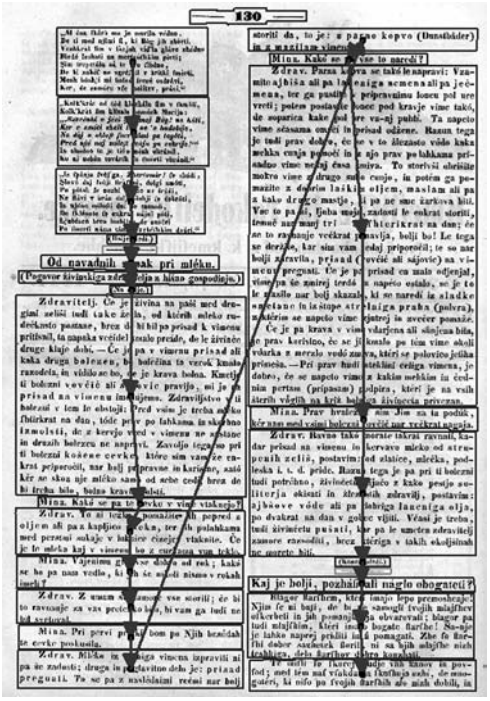
V programu Aletheia[10] smo pripravili vsak posamezen PAGE XML posebej, tako da smo ročno označili strukture posamezne strani, določili parametre za vse pojavne oblike/elemente na strani, ročno popravili besedilo in določili vrstni red segmentov. Na Sliki 1 je prikaz strani s končano segmentacijo na območja in izdelanim vrstnim redom območij na strani.

Postopek dela v programu Aletheia je zelo zamuden in zahteva veliko mero natančnosti. S programskimi orodji lahko sicer zelo natančno določimo posamezne strukture (območja) na strani, vendar se območja med seboj ne smejo prekrivati, kar je še posebej težko pri besedilih, kjer so razmiki med vrsticami izjemno majhni. Za vsako območje je treba določiti tip (ali gre za besedilno,

8 Medtem ko poteka optična prepoznava besedila z že uveljavljeno tehnologijo Abbyy FineReader (verzije 7–10), izvajamo popravke istih zapisov v programu Aletheia in tako izdelamo končno popravljeno različico zapisa, ki čim natančneje ustreza izvirniku.

9 USAL, ki je odgovorna za izdelavo vhodnih zapisov za ostale partnerje, ima programsko opremo OCR, prilagojeno za prepoznavanje nemškega, francoskega in angleškega, ne pa slovenskega jezika. Zato je smiselno, da NUK sam izdela svoje vhodne zapise, saj so tako že v izhodišču bolj natančni.

10 Program Aletheia je grafični urejevalnik segmentacije strani in korekcije transkripcij, ki je bil razvit v okviru projekta IMPACT.



Slika 1: Urejanje zaporedja besedilnih območij v Alethei

grafično, slikovno območje itn.), znotraj posameznih tipov pa še podrobnejše značilnosti (npr. pri besedilnem območju se določi, ali gre za odstavek, naslov, inicialko, številko strani, opombo itn., barva podlage in tiska, pisava, jezik; pri grafičnem območju se določi tip grafičnega znaka itn.). Namen označevanja območij in njihovega določanja je, da bi nova programska oprema OCR znala »prebrati« tudi zgradbo strani in v polnem besedilu ohraniti to strukturo. Pri prepoznavanju in določanju besedilnih območij se je izkazalo še posebej odločilno znanje slovenskega jezika, saj smo lahko le tako pravilno določili, ali gre za naslov ali le npr. za refren molitve, ki spada v kategorijo odstavka, a je grafično in vizualno postavljen in oblikovan tako, da bi lahko nosil tudi funkcijo naslova.

V naslednjem koraku se v posebno tekstovno okno vnese besedilo vsakega tekstovnega območja posebej (vsak naslov, odstavek itn.). Besedilo smo običajno v program prenesli iz predhodno narejenega zapisa OCR, ki je bil do določene mere že popravljen v Wordu. V določenih primerih, kjer je bil OCR izjemno slab, smo, da bi prihranili čas, besedilo ročno pretipkali. Besedilo v tekstovnem oknu je moralo tako na ravni znaka kot grafično popolnoma ustrezati tekstu na digitalizirani strani – upoštevani so bili torej vsi znaki, ostrivci, krativci, pri bohoričici dolgi s (ſ), ligature ſi, ſh, ſt, ſſ in struktura vrstic.

Kvaliteta besedil, ki smo jih vnesli v program in potem ročno popravljali, je bila odvisna od kvalitete predhodnega OCR. Programi OCR še ne prepoznavajo znakov, značilnih za bohoričico, zato je bil OCR pri besedilih v bohoričici zelo slab, ročno je bilo treba popraviti vse ligature, dolgi s, pogosto je prišlo do zamenjave črke ę z g ali ç itn.

V zadnjem koraku delovnega postopka se določi vrstni red posameznih območij, kakor naj bi si sledila v knjigi ali časopisu in kakor naj bi jih bral posamezen bralec. Določitev vrstnega reda pri knjigah ni povzročala težav, zato pa je bilo toliko težje določiti vrstni red elementov/branja pri časopisih, ki s svojo strukturo bralcu ponujajo možnost izbire. Program namreč omogoča povezavo elementov v urejeno skupino (ki je predvsem uporabna pri knjigah, kjer si odstavki sledijo eden za drugim) ali neurejeno skupino (pogosto uporabljena pri sodobnih časopisih, kjer so članki neodvisno razporejeni na strani). Vendar so strani *Kmetijskih in rokodelskih novic* po zgradbi bližje knjigam kot sodobnim časopisom, saj so časopisni članki razporejeni v dva stolpca in si sledijo zaporedno, eden za drugim. Zato smo se odločili, da jih razporedimo v urejeno skupino, kar olajša tudi izdelavo polnega besedila korpusa (glej poglavje 4).

Preden je bila transkripcija besedila shranjena v končnem zapisu, je bilo nujno programsko preveriti opravljeno delo. Določene napake, nastale med delovnim procesom (npr. prekrižana območja, manjkajoča območja v vrstnem redu, napačno klasificirana območja ali pomanjkanje besedila v določenih

besedilnih območjih) program avtomatsko zazna in javi napako. Vendar ne prepozna tipkarskih napak v tekstu ali napačno določenih območij, zato je ocenjena 99,95-odstotna natančnost transkripcij.

Da bi zagotovili kar se da visoko kvaliteto opravljenega dela, se je sočasno izvajalo preverjanje opravljenega dela na več ravneh:

1. vsak končni dokument je bil preverjen z validacijskim ukazom v programu,
2. naključni dokumenti so bili pregledani s strani partnerjev samih,
3. na ravni koordinatorja (Univerza v Innsbrucku) in
4. s strani IJS, z namensko napisanimi programi.

4 Izdelava besedilnega korpusa

Zbirka transkripcij, kjer je vsaka stran shranjena v svoji datoteki PAGE XML, je sicer primerna za raziskave OCR, manj pa za besedilni korpus, saj bi želeli v njem imeti celotno in zvezno besedilo posamezne enote korpusa (npr. knjige) skupaj z metapodatki (npr. avtor in leto izida), ki v PAGE XML niso vključeni. Zato smo napisali pretvorbeni program, ki kot vhod vzame tabelo z metapodatki (katere datoteke PAGE XML tvorijo posamezno publikacijo in kateri so njeni metapodatki) in datoteke PAGE XML ter iz njih naredi zaključena besedila v zapisu XML, ki je skladen s smernicami Text Encoding Initiative P5 (TEI 2007).

Pretvorba je v splošnem neproblematična, zaradi lastnosti PAGE XML pa se je vseeno pojavilo nekaj problemov. Največji problem so predstavljale neurejene skupine, saj pri njih ni mogoče določiti zaporedja besedila na strani; v našem primeru je bila rešitev v opustitvi takšnih skupin pri časopisnih člankih v *Kmetijskih in rokodelskih novicah*. Naslednji problem so bile ligature, v Alethei predstavljene kot posebni znaki (Unicode Private Use Area), ki niso prenosljivi med aplikacijami. Zato smo vse ligature iz PAGE XML pretvorili v ločene črke, ki so del standardnega unikoda. Tretji, resda redek problem pa so bili segmenti, ki vsebujejo spuščeno inicialko, ki se pojavi npr. na začetku poglavja. Inicialka tipično predstavlja prvo črko neke besede, katere ostanek neposredno začenja naslednji segment; v teh primerih jo enostavno združimo z naslednjim segmentom. Vendar pa so inicialke včasih tudi samostojne besede, pri čemer v PAGE XML ni načina, da bi ločili med obema primeroma. Problem rešujemo tako, da v PAGE XML za inicialke, ki so samostojne besede, dodamo presledek, pred pošiljanjem datotek v projektno validacijo pa te presledke avtomatsko odstranimo iz besedila.

Kot rezultat pretvorbe smo dobili dokumente TEI P5 za 79 enot korpusa, pri čemer vsaka enota vsebuje metapodatke, strukturne elemente, kot so poglavje, naslov, paginacija itn., znotraj njih pa zvezno besedilo, ki ohranja razdelitev po vrsticah. Besedilni korpus vsebuje skupaj okoli milijon besednih pojavnic.

5 Zaključek

V prispevku smo predstavili delo in rezultate sodelovanja NUK in IJS pri projektu IMPACT. Končni izdelek je korpus starejšega slovenskega jezika, ki vsebuje skenograme in pridružene datoteke PAGE XML s segmentacijo strani in čistopisom besedila, ter iz te osnove narejen besedilni korpus, zapisan v TEI P5. Takšen »zlati standard« naj bi omogočil izdelavo tehnologij, ki bodo sposobne uspešno in natančno »prebrati« starejša tiskana digitalizirana besedila in jih ponuditi uporabniku v polnem besedilu v elektronski obliki. Programska oprema bo zmožna prepoznati tudi besedila v bohoričici in v starejših tipografijah, opremljena bo z leksiko starejše slovenščine, zato bo besedilo pravilnejše in bolj čisto kot dosedanji poskusi. S tem se bodo uporabnikom in raziskovalcem odprla vrata do tiskanega izročila vsaj še ene duhovnozgodovinske dobe (od konca 18. stoletja do 1850), ki je bila do zdaj zaprta ter je ohranjala svojo eksistenco le na papirni

podlagi in kot slika na naših računalniških zaslonih. Besedila bodo strukturirana in dosegljiva v formatih, ki bodo uporabniku prijaznejši in bodo omogočali nadaljnje raziskovanje na področjih jezikoslovja, jezikovnih tehnologij, zgodovine, literarne zgodovine itn.

Literatura

BALK, Hildelis, PlLOEGER, Lieke, 2009: IMPACT: Working together to address the challenges involving mass digitization of historical printed text. *OCLC System and services: International digital library perspectives* 25/4. 233–248.

ERJAVEC, Tomaž, 2007: Architecture for Editing Complex Digital Documents. *Zbornik konference Digital information and heritage*. Zagreb: Odsjek za informacijske znanosti, Filozofski fakultet.

PLETSCHACHER, Stefan, ANTONACOPOULOS, Apostolos, 2010: The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. *20th International Conference on Pattern Recognition (ICPR2010). Istanbul, 23.–26. avgust 2010*. IEEE-CS Press. 257–260.

PRUNČ, Erich, 2007: Deutsch-slowenische/kroatische Übersetzung 1848–1918 (Ein Werkstättenbericht). *Wiener Slavistisches Jahrbuch* 53/2007. 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

Viri

Projekt IMPACT: www.impact-project.eu

Digitalna knjižnica Slovenije dLib: www.dLib.si

Digitalna zbirka AHLib: http://nl.ijs.si/ahlib

Spletni terminološki slovar informatike iSlovar: www.islovar.org

Standardni format za analizo oblike in besedilnih elementov skenogramov ALTO: www.loc.gov/standards/alto

Programski paket za OCR Abbyy FineReader: http://finereader.abbyy.com

Mednarodni konzorcij za zapis besedil TEI: www.tei-c.org.