Transkribus za namen optične prepoznave rokopisa: poročilo o uporabi in učenju besedilnega modela

Transkribus for handwritten text recognition: report on the use and training of a text model

Andreja Hari

Oddano: 10. 9. 2024 – Sprejeto: 31. 3. 2025

1.04 Strokovni članek
1.04 Professional article

UDK 004.85:[004.352.242:091]
DOI https://doi.org/10.55741/knj.69.2-3.5

Izvleček

Digitalizacija in prepoznavanje besedila sta ključna za omogočanje dostopa do zgodovinskih dokumentov, vključno z rokopisi. Članek predstavlja pregled obstoječih praks na področju optične prepoznave rokopisov (HTR) v Sloveniji in nekaj primerov tujih praks. Sledi poročilo o uporabi in učenju besedilnega modela za optično prepoznavo rokopisov z uporabo orodja Transkribus. Narodna in univerzitetna knjižnica (NUK) je v okviru projekta EODOPEN testirala platformo Transkribus na razmnoženih rokopisih semeniških predavanj Janeza Evangelista Kreka iz začetka 20. stoletja. Zapiske Krekovih predavanj so med letoma 1903 in 1907 nekateri bogoslovci stenografirali in po njegovem pregledu litografirali. Poročilo opisuje proces učenja besedilnega modela na enem delu in nato uporabo nastalega modela na treh dodatnih delih s podobno pisavo. Predstavljeni so tudi izzivi, s katerimi so se soočili, ter rezultati, ki so pokazali, da je uporaba Transkribusa učinkovita pri prepoznavanju besedil v slovenščini, čeprav sprva zahteva nekaj dodatnega ročnega dela. Članek se zaključuje s priporočili in idejami za nadaljnjo uporabo ter raziskovanje te tehnologije.

Ključne besede: digitalizacija, knjižnice, umetna inteligenca (UI), rokopisno gradivo, Optična prepoznava rokopisov (HTR)

Abstract

Digitization and text recognition are key to enabling access to historical documents, including manuscripts. This article provides an overview of existing practices in the field of Handwritten Text Recognition (HTR) and several examples of foreign practices. It continues with a report on the use and training of a text model for HTR using the Transkribus tool. The National and University Library (NUK), as part of the EODOPEN project, tested the Transkribus platform on the reproduced manuscripts of seminar lectures by Janez Evangelist Krek from the early 20th century. Some of his students stenographed Krek's lecture notes between 1903 and 1907, which were then lithographed after his review. The report describes the process of training a text model on one book and then applying the resulting model to three additional books with similar handwriting. It also presents the challenges encountered and the results, which demonstrated that the use of Transkribus is effective for recognizing texts in Slovene, although it initially requires some additional manual work. The article concludes with recommendations and ideas for further use and exploration of this technology.

Keywords: digitization, libraries, artificial intelligence (AI), manuscripts, Handwritten Text Recognition

1 Uvod in pregled področja

Začetki digitalizacije v knjižnicah segajo v devetdeseta leta prejšnjega stoletja, kmalu po pojavu svetovnega spleta. S pomočjo digitalizacije so knjižnice začele svojim uporabnikom ponujati nov način dostopa do gradiva v svojih zbirkah. Po letu 2000 so že začele nastajati nacionalne strategije, smernice za digitalizacijo, projekti digitalizacij ipd.,[1] zato smo v zadnjih 30 letih lahko videli, kako drastično se je področje razvijalo; od razvoja postopkov, ki so omogočali hitrejše

[1] S strani Narodne in univerzitetne knjižnice naj omenimo *Strategijo razvoja Digitalne knjižnice Slovenije – dLib.si 2007–2010* (2006), *Strategijo trajnega ohranjanja digitalnih virov v Narodni in univerzitetni knjižnici 2012–2020* (2012) ter *Strategijo trajnega ohranjanja in omogočanja dostopnosti do zapisov na nosilcih z omejeno obstojnostjo v Narodni in univerzitetni knjižnici* (2018). Narodna in univerzitetna knjižnica sledi tudi *Smernicam za zajem, dolgotrajno ohranjanje in dostop do kulturne dediščine v digitalni obliki* (2013) Ministrstva za kulturo, internim *Smernicam za digitalizacijo knjižničnega gradiva* (2010), internim *Smernicam za varovanje in trajno ohranjanje knjižničnega gradiva v Narodni in univerzitetni knjižnici* (2021) ter *Enotnim zahtevam in postopkovnemu modelu izvajanja interne digitalizacije knjižničnega gradiva v Narodni in univerzitetni knjižnici, različica 1.1* (2023). Digitalizacija gradiva je bila omenjena ali vključena tudi v *Strateškem načrtu NUK 2004–2008* (2004), *Strateškem načrtu Narodne in univerzitetne knjižnice za obdobje 2010–2013* (2010), *Strateškem načrtu Narodne in univerzitetne knjižnice za obdobje 2015–2019* (2014) ter *Strateškem načrtu Narodne in univerzitetne knjižnice za obdobje 2020–2024* (2019) in bo zagotovo tudi v bodoče. Narodna in univerzitetna knjižnica je bila vključena v nacionalne in mednarodne projekte na področju digitalizacije (npr. EU project eTEN DOD 2006–2008, EOD Culture project 2009–2014 in EODOPEN 2019–2024), trenutno pa poteka tudi najobširnejši *Nacionalni projekt digitalizacije slovenike in kulturnozgodovinsko pomembnega knjižnega gradiva (2021–2030)* (2020).

in kakovostnejše delo, ter masovne digitalizacije do razvoja in načinov dostopa prek nastalih digitalnih knjižnic. S pojavom avtomatskih postopkov in strojnega učenja se je, in se bo, področje digitalizacije še naprej razvijalo. Digitalne knjižnice z umetno inteligenco (UI) že danes omogočajo drugačno delo z besedilom kot do pred kratkim. Omogočajo npr. prevajanje, priprave povzetkov, glasno branje,[2] kar je za uporabnike nekaj novega in jim hkrati omogoča nov način upravljanja gradiva – tudi tistim, ki jim je bilo do sedaj zaradi npr. jezikovne omejenosti nedostopno.

Velik napredek na področju digitalizacije besedil je nastopil z razvojem optične prepoznave znakov (*Optical Character Recognition* – OCR). Slednja zajema avtomatske postopke, ki iz skenogramov prepoznajo strukturo strani, le-to razčlenijo in hkrati prepoznajo besedilo. Ta postopek je enostavnejši in zanesljivejši pri tiskanem gradivu, saj je prepoznava osnovana na pisavah oziroma obliki črk (tipografiji), ki se tudi med današnjimi sodobnimi tipskimi tipografijami pisav zelo malo razlikujejo. Nasprotno pa postopek optične prepoznave znakov ni uporaben za rokopisno gradivo, saj se pisave razlikujejo od pisca do pisca in ni mogoče ustvariti enotnega sistema za prepoznavo vseh pisav. V danih primerih je še vedno pogosta praksa, da se optična prepoznava znakov ne izvede oziroma se besedilo ročno prepisuje.

Z razvojem optične prepoznave znakov se je zdelo, da bodo podobni rezultati v roku nekaj let na voljo tudi za rokopisno gradivo, vendar tehnologija, ki temelji na ideji izoliranja posameznih znakov, nikoli ni bila sposobna prinašati dobrih rezultatov, razen nekaterih uspehov pri prepoznavanju lepo napisanih črk iz srednjega veka. Šele leta 2010 je uvedba nevronskih mrež in globokega učenja privedla do izrazitega in presenetljivega napredka pri optični prepoznavi rokopisov (*Handwritten Text Recognition* – HTR). (Hodel, 2022). Muehlberger idr. (2006) so bili že pred dejanskim razvojem optične prepoznave rokopisov mnenja, da bo uspešen razvoj prepoznave rokopisov izboljšal in povečal dostop do zbirk, kar bi uporabnikom omogočilo hitro in učinkovito iskanje določenih tem, besed, oseb, krajev ter dogodkov v dokumentih. Poleg tega pa bo spremenil razumevanje konteksta in povečal možnosti za raziskovanje. Nockels idr. (2024) menijo, da model optične prepoznave rokopisov močno olajša prepis virov iz predindustrijske in predmehanske dobe (oziroma vsaj iz obdobja pred tiskom in pisalnim strojem), s čimer povečuje njihovo iskalnost ter možnost spreminjanja njihove oblike. Hkrati optična prepoznava rokopisov omogoča strojno obdelavo

[2] Primer takšne digitalne knjižnice je skupni portal EODOPEN, na katerem so združena vsa dela petnajstih projektnih partnerjev in ki ob prijavi v sistem omogoča omenjene funkcije. https://diglib.eodopen.eu/

besedila zasebnih, intimnih dokumentov, vključno z osebnimi pismi, dnevniki, institucionalno korespondenco, rokopisi, knjigovodskimi zapisi, računi in uradnimi zapisi, kot so popisni materiali, do katerih je bila dostopnost do sedaj omejena oziroma so bili postopki zaradi ročnih prepisov dolgotrajni.

Za namen raziskovanja novih možnosti v postopkih digitalizacije in sistemov, ki do sedaj še niso bili v uporabi, se je Narodna in univerzitetna knjižnica oktobra 2023 v okviru projekta EODOPEN[3] pridružila združenju READ-COOP SCE.[4] Poslanstvo združenja je zagotavljati širok spekter orodij in storitev, ki raziskovalcem, institucijam ter posameznikom omogočajo skupno odkrivanje in raziskovanje bogate zgodovine, skrite v zgodovinskih dokumentih. (READ-COOP SCE, b. d. a) V združenju je trenutno več kot 150 članic iz več kot 30 držav. Narodna in univerzitetna knjižnica je trenutno edina pridružena organizacija iz Slovenije; iz sosednje Avstrije sodeluje 22 organizacij, Italije 5 organizacij, Madžarske 1 organizacija in Hrvaška trenutno nima partnerske organizacije.

Pod okriljem združenja deluje Transkribus[5] – platforma in orodje, ki omogoča samodejno prepoznavanje besedila, napisanega z roko, na enostaven način ter brezhibno urejanje skenogramov in prepoznanega besedila. Hkrati podpira enostavno sodelovanje, pri čemer lahko več oseb sodeluje na istih dokumentih, in skupno rabo že obstoječih modelov. Omogoča tudi učenje modelov z uporabo umetne inteligence, in sicer za digitalizacijo ter interpretacijo zgodovinskih dokumentov v katerikoli obliki. (READ-COOP SCE, b. d. b) Transkribus je zasnovala, lansirala in financirala Evropska komisija leta 2015 kot del projekta tranScriptorium,[6] nadaljeval se je pod okriljem projekta READ[7] ter bil nato prevzet z ustanovitvijo pravne osebe READ-COOP za ohranjanje in nadaljnji razvoj platforme (Muehlberger idr., 2006; READ-COOP SCE, b. d. c).

Platforma Transkribus omogoča učenje dveh vrst modelov: postavitvenega modela, ki razčleni postavitev vsebine na posamezni strani, in besedilnega modela za prepoznavo besedila na posamezni strani. Za obe vrsti modelov je za učenje

[3] EODOPEN ali eBooks-On-Demand-Network Opening Publications for European Netizens (E-knjige po naročilu – odpiranje dostopa do gradiva za evropske uporabnike spleta) je projekt, ki ga sofinancira Evropska komisija v okviru programa Ustvarjalna Evropa. Države partnerice projekta so Avstrija, Češka, Estonija, Litva, Madžarska, Nemčija, Poljska, Portugalska, Slovaška, Slovenija in Švedska. Skupna vrednost projekta je približno 4 milijone evrov. Trajanje projekta: od oktobra 2019 do oktobra 2024. http://www.eodopen.eu/

[4] https://readcoop.eu/

[5] https://www.transkribus.org/

[6] Projekt je potekal v obdobju 2013–2015.

[7] READ ali Recognition and Enrichment of Archival Documents (Prepoznava in obogatitev arhivskih dokumentov). Projekt je potekal v obdobju 2016–2019.

najprej treba ustvariti osnovno bazo podatkov in uporabiti že obstoječi postavitveni oziroma besedilni model. Nekateri že obstoječi modeli so javno dostopni za vse člane združenja, medtem ko so modeli lahko tudi zasebni ali omejeni na vrsto članstva v združenju. Vse to podpira sodelovanje, saj lahko nekdo uporabi javni model, ki ga je razvil že nekdo drug; hkrati na vsakem projektu znotraj organizacije lahko sodeluje več oseb naenkrat. Glede modelov lahko še omenimo, da je bilo članom združenja 12. julija 2024 na voljo trinajst javno dostopnih postavitvenih modelov in 204 javno dostopni besedilni modeli.[8] Hkrati READ-COOP SCE (b. d. b) navaja, da Transkribus uporablja prek 150.000 registriranih uporabnikov, da je bilo ustvarjenih že več kot 20.000 modelov optične prepoznave rokopisov z umetno inteligenco in obdelanih že prek 50 milijonov strani.

Milioni (2020) je izvedla raziskavo rabe Transkribusa med evropskimi knjižnicami in arhivi, pri čemer je od 44 posredovanih vprašalnikov prejela petnajst odgovorov na temo digitalizacije in optične prepoznave rokopisov. Rezultati so pokazali, da deset organizacij (66,7 %) ne ponuja prepoznanih besedil rokopisov svojim uporabnikom, medtem ko preostalih pet (33,3 %) to naredi le občasno. Med rezultati navaja, da ročne prepise rokopisov običajno pripravijo arhivisti ali specialisti ter da občasno to delo opravijo tudi sami raziskovalci in zaključene prepise darujejo inštituciji. Institucije so navajale, da še niso našle sistematske rešitve, ki bi omogočila dovolj kakovostno besedilo za objavo poleg slikovne reprodukcije, zato prepis besedil pogosto poteka ročno, kar je lahko zelo časovno zamuden postopek. Nekatere institucije so tudi navajale prihodnje načrte rabe Transkribusa za namen prepisa besedila oziroma naklonjenost orodju v primeru, da bi se izkazalo za uporabno. Prebor (2024) kot primer dobre prakse navaja uporabo Transkribusa za prepoznavo hebrejskih rokopisov iz 15. stoletja in povzema, da čeprav je še vedno potrebna ročna korekcija, prinaša Transkribus sorazmerno dobre rezultate, ki znatno zmanjšujejo čas in stroške, povezane s prepisovanjem. Dodaja, da integracija tehnologije v procese prepisovanja ponuja obetavne priložnosti za znanstvenike in raziskovalce, ki delajo s starejšimi rokopisi, ter da to povečuje učinkovitost, dostopnost in ohranjanje dragocene kulturne dediščine.

Kot primer dobre prakse rabe Transkribusa za optično prepoznavo rokopisov naj navedemo tudi mednarodni projekt *Peter Handke: Notizbücher (HNB-DE)*,[9] katerega prvo obdobje je trajalo od 15. februarja 2021 do 14. junija 2024. V projektu sodelujejo avstrijske in nemške organizacije s ciljem, da bi bili prepisi

[8] Pri tem gre za članstvo *Individual*. V primeru višje stopnje članstva, ki se imenuje *Scholar*, bi imeli na voljo tudi supermodele, prepoznavo polj in preglednic, pametno iskanje, Transkribus Sites za ustvarjanje spletnih strani in večjo hitrost obdelave.

[9] https://edition.onb.ac.at/fedora/objects/context:hnb/methods/sdef:Context/get

75 rokopisnih beležnic Petra Handkeja javno dostopni vsakomur. Ob zaključku prvega obdobja projekta je na spletu že na voljo 21 beležnic, pri čemer lahko uporabnik istočasno pregleduje izvirnik in prepisano besedilo. V naslednjem obdobju načrtujejo pretvorbo še 26 beležnic. V postopku so najprej uporabili postavitveni model za prepoznavo odsekov in vrstic, nato so ročno dopolnjevali postavitev. V nadaljevanju so ročno prepisovali besedilo, ga uporabili za učenje besedilnega modela in nastali model nato uporabili za optično branje rokopisov na drugih beležnicah. (Peter Handke Notizbücher, 2024)

Ob pregledu besedilnih modelov Terras (2022) navaja, da je Transkribus uspešno naučen prepoznati besedila različnih jezikov, vključno z angleščino, italijanščino, nizozemščino, latinščino, švedščino, finščino, danščino, staro nemščino, poljščino, bengalščino, hebrejščino, cerkveno slovanščino in arabščino, pri čemer so bili ustvarjeni različni besedilni modeli za različna časovna obdobja. Najboljši rezultati iz Transkribusa dosegajo nizek odstotek napačno prepoznanih znakov (*Character Error Rate* – CER): pod 5 % za rokopisno gradivo in pod 1 % za tiskano gradivo. Terras (2022) prav tako omenja, da so v letu 2018 opravili raziskavo med uporabniki orodja Transkribus, pri čemer je zanimivo, da so takrat na platformi prevladovali dokumenti v nemškem, latinskem, angleškem, francoskem, italijanskem, nizozemskem, grškem in španskem jeziku. Mnogo manj so bili zastopani danski, madžarski, poljski, katalonski, norveški, portugalski, kitajski jezik ipd. Slovenski jezik v dani raziskavi še ni bil prisoten.

Dosedanja manjša testiranja, ki smo jih izvedli v Transkribusu, so pokazala, da so obstoječi postavitveni modeli primerni za slovenske monografije oziroma dela z enostavno postavitveno strukturo; če pride do napak v avtomatski prepoznavi, sistem omogoča ročno odpravo oziroma ureditev morebitnih napak. Kar se tiče besedilnih modelov, so trenutno za slovenščino na voljo le trije, ki pa vsebujejo tudi druge jezike, torej niso specifični le za slovenščino. Od tega sta imela dva besedilna modela za osnovno bazo podatkov tiskano gradivo (*Transkribus Print M1* in *Glagolitic printings PyLaia*) ter le eden rokopisno gradivo iz 18. stoletja (*Slovenian 18th century manuscript*); slednjega je leta 2023 ustvarila Slovenska akademija znanosti in umetnosti (SAZU). Zaradi pomanjkanja besedilnih modelov za slovenski jezik bi za rabo orodja za rokopise potrebovali učiti nove modele, predvsem ko gre za novejše rokopise, ki so pisani v slovenščini. Drugačna situacija je s srednjeveškimi rokopisi, ki so pisani večinoma v latinščini in katerih področje bi bilo treba dodatno raziskati, saj morda zanje že obstaja besedilni model, ki bi ga lahko uporabili. Težave pa ne predstavljajo le trenutno obstoječi besedilni modeli, ampak tudi že omenjeno dejstvo, da se pisave različnih piscev med seboj razlikujejo, zato en model ne zadošča za vse rokopise.

Narodna in univerzitetna knjižnica za namen digitalizacije uporablja *Smernice za digitalizacijo knjižničnega gradiva* (Smernice za digitalizacijo knjižničnega gradiva, 2010), osnovane na mednarodnih priporočilih, ter interno razvito orodje Digitisation Manager (DM), ki je namenjeno urejanju in obdelavi skenogramov, zajemu in pripravi metapodatkov ter izvozu datotek za objavo na Digitalni knjižnici Slovenije.[10] Pri obdelavi skenogramov se opravijo poravnava, obrez in izenačevanje velikosti skenogramov ter občasno tudi drugi postopki, kot sta poravnava ukrivljenih vrstic besedila in prilagoditev kontrasta ali svetlosti. Za optično prepoznavo znakov se v Narodni in univerzitetni knjižnici uporablja strežniška verzija programa Abbyy FineReader,[11] pri čemer se na letni ravni zakupi izbrana količina strani za pisavi: gotica (nemška kurenta) in latinica. Pri latinici je vključen tudi jezik slovenščina, če gre za dela v slovenskem jeziku. Optična prepoznava znakov se opravi le pri gradivu, kjer je besedilo natisnjeno. (Klasinc idr., 2023) Vsi omenjeni postopki so standardni pri tiskanem gradivu, težave nastanejo, ko je v digitalizaciji rokopisno gradivo. Izvajanje optične prepoznave znakov takega gradiva je prek serverja Abbyy FineReader na voljo, vendar je nesmiselno, saj ne dobimo primernih oziroma berljivih rezultatov. Objavljeni in javnosti dostopni digitalizirani rokopisi so tako samo slikovne reprodukcije v formatu PDF, ki niso optično prepoznane. Podoben primer internega orodja imajo tudi v Univerzitetni knjižnici v Greifswaldu. Tam se s Transkribusom ukvarjajo že od leta 2015, dosegajo uspešne rezultate in je implementiran v njihovo orodje za digitalizacijo, Goobi (READ-COOP SCE, b. d. č).

Zaradi nedostopnosti samega besedila rokopisov smo se v okviru projekta EODOPEN odločili za prvo testiranje Transkribusa na primeru rokopisov enega pisca, in sicer pod pogojem, da so dela izšla v časovnem okviru, ki ga projekt pokriva – 20. in 21. stoletje. Za prvo fazo učenja besedilnega modela je bilo izbrano delo *Psihologija*, razmnožen rokopis zapiskov semeniških predavanj Janeza Evangelista Kreka, iz leta 1905, ki je v javni domeni. Zapiske Krekovih predavanj so namreč nekateri bogoslovci med letoma 1903 in 1907 stenografirali – pisali v tesnopisu za hitrejše zapisovanje – in jih po Krekovem pregledu litografirali (tehnika razmnoževanja) pri Blasniku. (Slovenska biografija, 2013) Po prvi fazi smo testiranje razširili še na tri dela, prav tako Krekovih semeniških predavanj, in sicer podobnih rokopisnih pisav, saj smo besedilni model že imeli in učenje slednjega pri teh delih ni bilo potrebno. Celoten postopek je bil izveden v juniju in juliju 2024 ter je podrobneje opisan v nadaljevanju.

[10] https://www.dlib.si/
[11] https://www.abbyy.com/

2 Pomembni postopki za uporabo Transkribusa

Glede na izvedbo in rezultate testiranj priporočamo, da se pri odločitvi o uporabi Transkribusa za namen optične prepoznave rokopisov najprej vprašamo:

- Ali je oseba pisala dovolj berljivo, da znamo sami brati besedilo? Za pripravo osnovne baze podatkov besedila, ki bo namenjena učenju modela, je treba ročno prepisati del besedila, zato je to vprašanje prvo in ključno.
- Ali se bo Transkribus lahko naučil razpoznati pisavo na relativno majhnem začetnem številu besed? Za učenje besedilnega modela je treba presoditi smiselnost učenja – ali bo hitreje kratka besedila pretipkati na roke ali pa gre za večji obseg gradiva in bo z učenjem besedilnega modela postopek hitrejši.
- Ali imamo dovolj kreditov? Krediti so predhodno zakupljeno plačilno sredstvo za izvajanje obdelave v Transkribusu. Število kreditov, ki jih imamo na posameznem uporabniškem računu, se določi ob registraciji, in sicer se cene razlikujejo tudi glede na vrsto članstva (Individual, Scholar, Team). Obstaja tudi možnost dodatnega zakupa kreditov, če presodimo, da jih na mesečni ravni potrebujemo več. Hkrati je ob vključitvi organizacije v združenje en uporabnik upravičen do 1000 kreditov na leto. Transkribus za izvajanje postavitvenega in besedilnega modela zahteva kredite. Prvi model zahteva okvirno 0,25 kredita na stran in drugi en kredit na stran. Za sto strani knjige okvirno nanese 125 kreditov.

Ko sprejmemo odločitev na podlagi prejšnjih vprašanj, izvedemo naslednje korake, ki so ključni za uspešen potek učenja novega besedilnega modela in izvedbo optične prepoznave rokopisa:

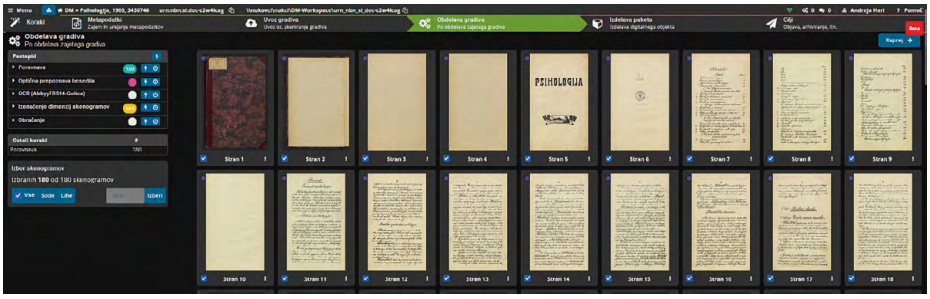
- Za uvoz v Transkribus priporočamo format PDF z že obdelanimi skenogrami, vendar brez izvedbe optične prepoznave znakov z Abbyy FineReaderjem. Druga možna formata uvoza sta: JPEG/JPG in PNG.
- V Transkribusu izvedemo prepoznavo postavitve z obstoječim postavitvenim modelom, ki prepozna vrstice in postavitev besedila na vsaki strani.
- Preverimo kakovost postavitvenega modela in popravimo vrstice, kjer je to potrebno.
- Ustvarimo osnovno bazo besedila tako, da ročno vnesemo vsaj 5000 besed ali 20 strani besedila, kar je minimalna vrednost za doseganje zadovoljivih rezultatov.
- Zaženemo učenje novega besedilnega modela na podlagi osnovne baze besedila in na podlagi obstoječega besedilnega modela (priporočljivo iz iste jezikovne družine).
- Preveriti je treba odstotek napačno prepoznanih znakov (CER) ustvarjenega besedilnega modela, ki naj bo čim nižji (vsaj pod 8 % ali pod 5 %).

– Če z odstotkom nismo zadovoljni, je treba povečati osnovno bazo podatkov besedila in ponavljati prejšnji dve alineji, dokler odstotek napačno prepoznanih znakov ni zadovoljiv.
– Zaženemo model optične prepoznave rokopisov na preostalem besedilu, in sicer z ustvarjenim besedilnim modelom.
– Pridobljeno besedilo izvozimo iz Transkribusa v formatih, ki jih potrebujemo za objavo. V primeru Narodne in univerzitetne knjižnice sta trenutno to PDF in TXT za objavo na portalu Digitalne knjižnice Slovenije.

3 Postopek digitalizacije in uvoz v Transkribus

3.1 Digitalizacija

Kot že omenjeno, smo za prvo fazo testiranja uporabili knjigo *Psihologija*, ki vsebuje zapiske predavanj Janeza Evangelista Kreka. Knjiga je sicer rokopis, vendar delo v COBISS-u (ID = 3430746) ni inventarizirano kot rokopisno gradivo, ker gre za reprodukcijo. Glede na smernice digitalizacije v Narodni in univerzitetni knjižnici velja, da sta v tem primeru digitalizacija in obdelava skenogramov potekali enako kot postopki za monografsko gradivo: skeniranje na 300 DPI, poravnava skenogramov, obrezano glede na notranji rob, razen prva dva in zadnja dva skenograma, ki sta obrezana nekoliko čez rob knjige. Kot zadnji postopek obdelave je bil izveden proces izenačevanja dimenzij skenogramov. Postopek optične prepoznave znakov, ki je običajno pri tiskanem gradivu tudi izveden v tej fazi, tokrat ni bil izveden, saj bodo ti postopki potekali v Transkribusu. V naslednjem postopku, v orodju Digitisation Manager, je bil izdelan celoten paket, ki ga običajno potrebujemo za objavo na dLib.si, saj smo za uvoz v Transkribus potrebovali datoteko PDF.



Slika 1: *Vmesnik orodja Narodne in univerzitetne knjižnice Digitisation Manager po opravljeni obdelavi skenogramov*

3.2 Uvoz v Transkribus

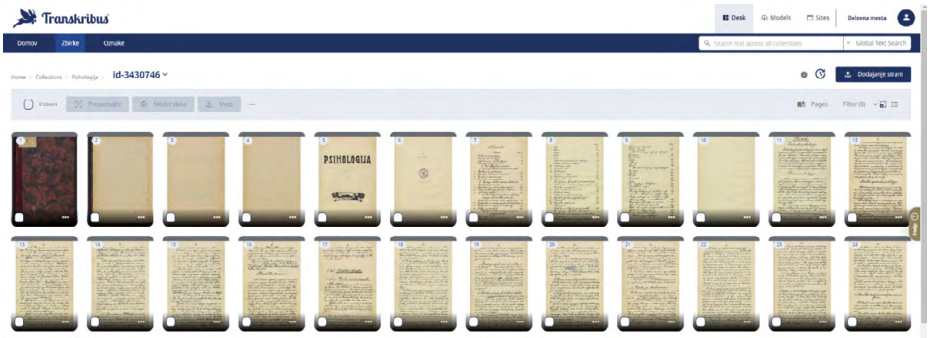
Uvoz v Transkribus omogoča naslednje formate datotek: JPEG/JPG (do 10 MB), PNG (do 10 MB) in PDF (do 200 MB) z največ 3000 stranmi. Zaradi teh omejitev smo že na začetku za uvoz pripravili PDF, izdelan z digitalizacijo v orodju Digitisation Manager. Če bi želeli uvoziti druge formate, bi morali opraviti pretvorbo formatov, saj med digitalizacijo v Narodni in univerzitetni knjižnici nastaneta datotečna formata TIFF in JP2, ki ne le, da nista podprta, tudi velikosti presegajo omejitve 10 MB.

Ustvarjanje zbirke

Za dotični projekt smo v Transkribusu ustvarili novo zbirko, z naslovom *Psihologija*. Predlagamo, da se za vsak večji projekt ustvari ločena zbirka, da imamo vse projekte primerno urejene. Še bolje bi bilo, če bi v poimenovanje zbirke dodali tudi leto izdaje in COBISS-ID, da se v primeru več projektov lažje najde točno določeno gradivo (npr. Psihologija, 1905, 3430746), ter bi hkrati ohranjali isti sistem poimenovanja, kot ga uporabljamo v orodju Digitisation Manager. V našem primeru v začetku testiranja na to nismo pomislili, saj nismo pričakovali večje količine gradiv v Transkribusu, a smo pri vseh nadaljnjih gradivih upoštevali omenjeni dodatni zapis leta izdaje in COBISS-ID.

Naložitev datotek v zbirko

Ko je zbirka ustvarjena, je treba dodati datoteke, v našem primeru datoteko PDF. Ko datoteko izberemo, se nam odpre Transkribusov seznam delovnih postopkov (angl. *Transkribus Server Jobs*), kjer lahko spremljamo, kdaj bo uvoz dokumenta zaključen.



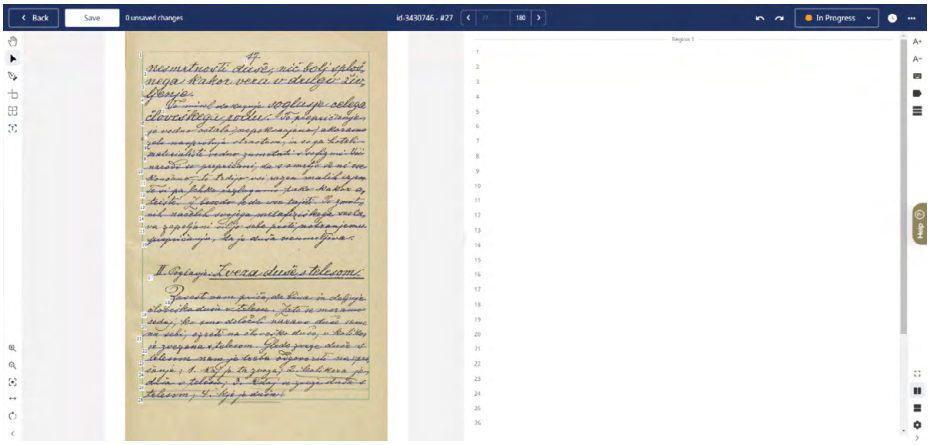
Slika 2: *Vmesnik Transkribusa po uvozu datoteke PDF*

Ob kliku na ustvarjeni dokument se nam v naslednjem oknu odprejo vse strani knjige, in sicer vsaka posebej. Od tu naprej sledi uporaba postavitvenega modela in dalje učenje besedilnega modela za prepoznavo besedila rokopisa.

4 Uporaba postavitvenega modela

Pri fazi uporabe postavitvenega modela smo najprej izbrali skenograme, za katere želimo uporabiti model – prazne strani smo pustili neizbrane, saj s tem koristimo manj kreditov. Na seznamu javno dostopnih postavitvenih modelov smo izbrali model *Universal lines*, ki je bil naučen na velikem vzorcu strani (24.000+), na rokopisnem gradivu, in ima 8.94 % napačno prepoznanih znakov. Sistem nam tudi prikaže potrebno število kreditov za izbrani postopek. Ob zagonu je bil postopek dodan na seznam delovnih postopkov in počakali smo na dokončano izvedbo.

Če po izvedenem postopku odpremo vsako stran posebej, lahko vidimo (slika 3), da so na njih prepoznani odsek (eden, ker gre za preprosto, enostolpično strukturo strani) in vrstice. V tej fazi besedilo še ni prepoznano.



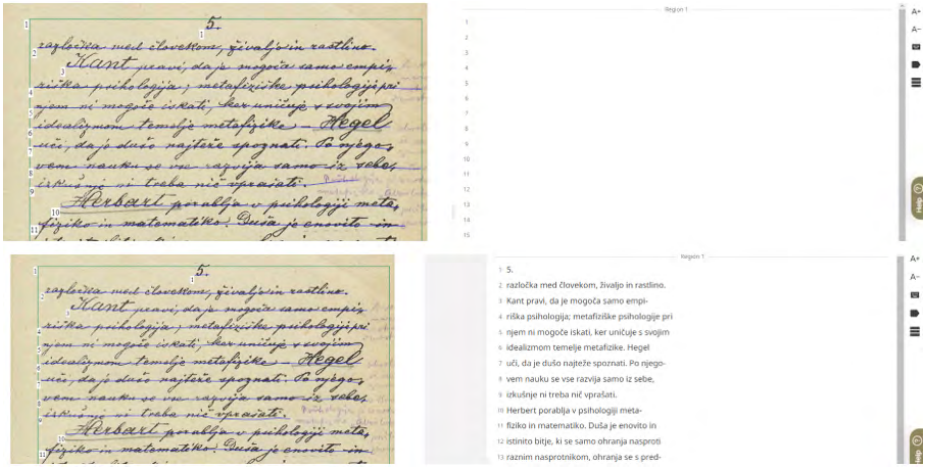
Slika 3: *Vmesnik Transkribusa po izvedbi postavitvenega modela; označena sta odsek in vsaka vrstica besedila*

5 Učenje besedilnega modela in optična prepoznava znakov

5.1 Pred učenjem besedilnega modela

Pred začetkom učenja besedilnega modela smo morali opraviti pregled prepoznanega postavitvenega modela – urejanje odsekov in prepoznanih vrstic na vsaki strani. Odseki morajo zajemati obseg besedila, vrstice pa morajo biti pravilno označene in pravilno dolge, kot je samo besedilo.

Za vse strani smo opravili naslednji postopek:
- Pregled in morebitno preoblikovanje prepoznanih odsekov na vseh straneh – večanje odsekov in izbris nepotrebnih ali napačno prepoznanih odsekov.
- Urejanje prepoznanih vrstic, kar vključuje:
  - brisanje nepotrebno prepoznanih vrstic – če so vrstice prazne ali je prepoznana vrstica, kjer je nekdo nekaj dopisal, in tega ne želimo imeti v prepoznanem besedilu;
  - dodajanje manjkajočih vrstic;
  - urejanje dolžin vrstic – podaljševanje in krajšanje vrstic ali postavitev teh tako, da so bolj ravne.
- Za hitrejše postopke smo uporabljali ukaze z bližnjicami: M – združi, B – orodje za dodajanje vrstic, O – orodje za izbiro, CTRL + S – shrani.



Slika 4: *Zgornji primer prikazuje stanje pred urejanjem vrstic in prepisom besedila, spodnji pa stanje po opravljenih postopkih (skrajšane in urejene vrstice ter prepis besedila)*

Po končanem postopku smo začeli z ročnim prepisovanjem besedila, in sicer za vsako označeno vrstico posebej. Pri tem smo bili pozorni na naslednje dodatne postopke:

– Prepis besedila: Za učenje besedilnega modela je priporočljivo, da ročno vnesemo vsaj 5000 besed ali 20 strani besedila. Če kakšne besede nismo znali prebrati, smo jo izpustili in nadaljevali z besedilom, ki je preprostejše za prebrati. Pri prepisu smo morali biti konsistentni. V našem primeru smo se odločili konsistentno navajati deljenje besed (zapis ,, smo spremenili v -), posodobiti okrajšave (n. pr. v npr. in i. t. d. v itd.) ter zapisati številke strani s piko.
– Vsaki dokončani strani, ki smo jo želeli uporabiti za učenje modela, smo spremenili status iz *in progress* (slov. v pripravi) v *ground truth* (slov. referenčni prepis).

5.2 Učenje besedilnega modela

Ko smo presodili, da imamo zadostno število prepisanih besed, smo začeli učenje besedilnega modela za ta rokopis. Ta faza vsebuje štiristopenjski postopek:

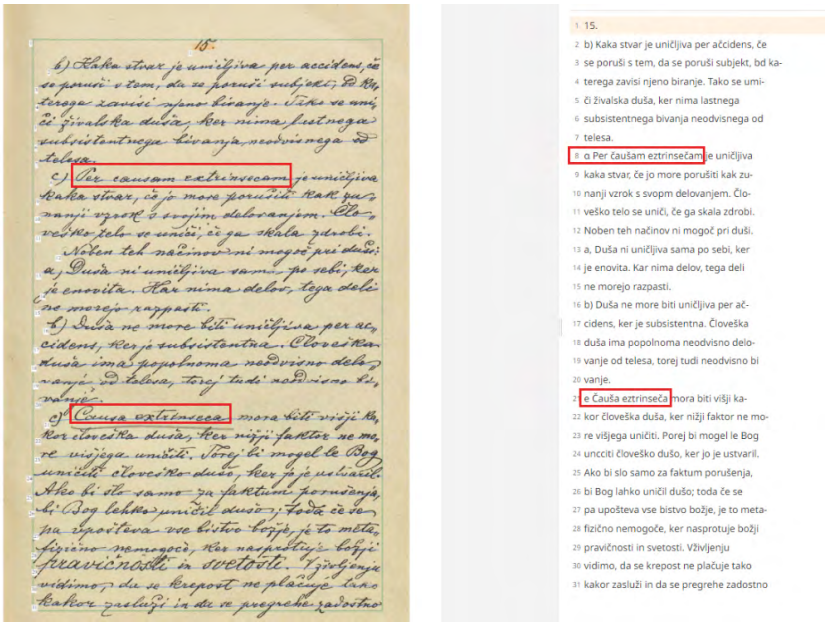
1. **Izbor podatkov za učenje modela.** Sistem nam na izbranem delu samodejno izbere strani, ki vsebujejo status *Ground Truth*. Svetuje nam, naj bo vsaj dvajset strani prepisanega gradiva.
2. **Validacija podatkov.** V tej fazi nismo ničesar spremenili in smo ohranili generične nastavitve.
3. **Nastavitve besedilnega modela.** V tej fazi smo nastavili ime in druge lastnosti modela ter izbrali že obstoječi besedilni model, ki smo ga vzeli za osnovo. V našem primeru je to bil model *The German Giant I*, saj ima sprejemljivo dober odstotek napačno prepoznanih znakov, podobno obdobje rokopisa in visok osnovni set podatkov besedila (prek 15 milijonov), na katerem je bila naučena prepoznava besedila.
4. **Začetek.** V tej fazi zaženemo učenje besedilnega modela. Postopek se avtomatično doda na seznam delovnih postopkov.

Po končanem postopku smo preverili odstotek napačno prepoznanih znakov besedilnega modela, ki je v našem primeru znašal 3 %, kar je zadovoljiva vrednost. Naš model je zaseben, vse dokler se ne odločimo, da ga nastavimo kot javnega.

5.3 Zagon optične prepoznave na podlagi ustvarjenega besedilnega modela

Za testiranje novonastalega modela smo v pogledu vseh skenogramov izbrali le nekaj strani, na katerih še ni bila narejena optična prepoznava. V postopku smo med zasebnimi modeli našli ustvarjenega in počakali, da se je postopek zaključil.

Po zaključku smo preverili kakovost prepoznanega besedila. V našem primeru je bilo besedilo dobro prepoznano. Težave so bile vidne večinoma pri:
- oznakah za alineje – a), b), α), β), γ) ...;
- menjavi črk c, z, s, č, ž, š;
- menjavi črk: c in e, ni in m, n in u itd., kar je razumljivo, saj so si oblikovno podobne;
- nekaj težav je bilo zaznanih pri velikih začetnicah, saj so se posamezne črke v osnovi bazi podatkov redko pojavile;
- nekaj težav je bilo pri izrazih, ki niso v slovenščini, oziroma znaku x, saj so se tudi ti v osnovni bazi redko pojavili.



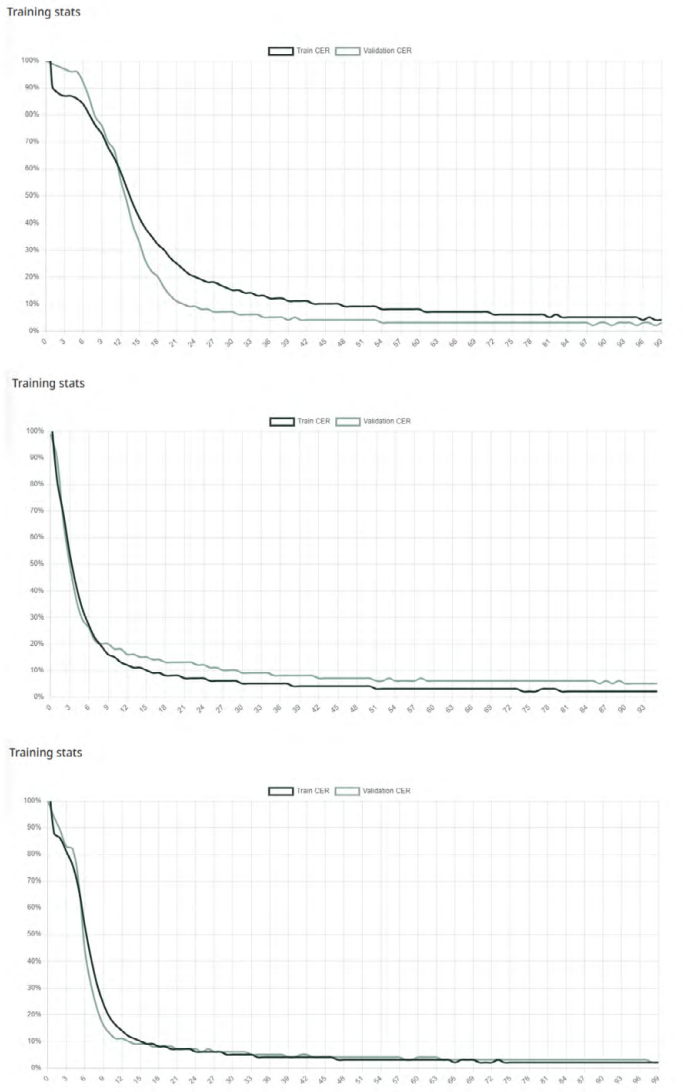
Slika 5: *Primer optično prepoznanega rokopisa na podlagi novega modela, z označenimi očitnimi napakami pri latinskih izrazih*

V tej fazi smo se za namen testiranja odločili, da bi poskušali odstotek napak še bolj zmanjšati. Povečali smo osnovni set podatkov besedila za učenje, tako da smo popravili dodatnih nekaj strani na podlagi že opravljene optične prepoznave (okvirno 10–25 popravkov je bilo na posamezno stran), kar je vsekakor potekalo hitreje kot ročno prepisovanje celotnih strani. Nato smo postopek učenja besedilnega modela ponavljali, vse dokler odstotek ni bil zadovoljiv.

Preglednica 1 in slika 6 prikazujeta statistiko učenja treh dodatno pripravljenih besedilnih modelov.

Preglednica 1: *Primerjava treh verzij modelov*

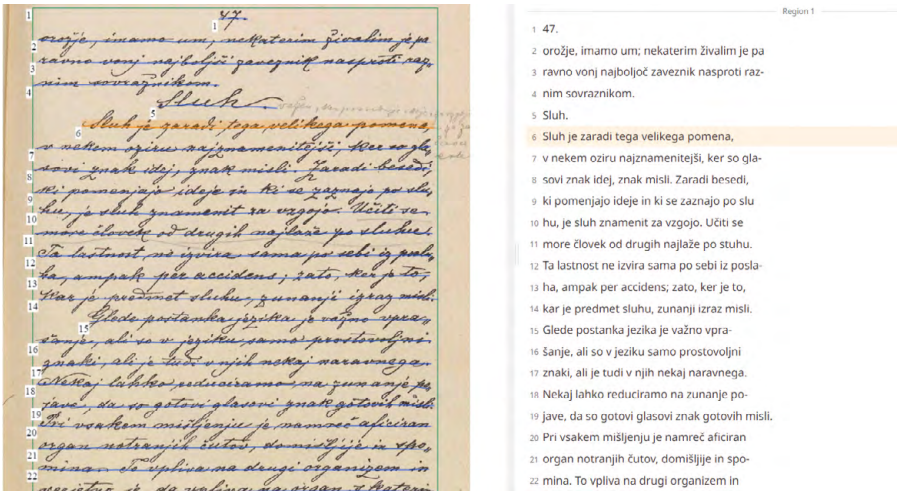
Model	Velikost osnovne baze (št. besed)	Število strani za osnovo	Napačno prepoznani znaki (%)	Osnovni model za učenje
Verzija 1	2623	24	3,00	The German Giant I
Verzija 2	4880	36	5,00	Verzija 1
Verzija 3	680	48	2,00	The German Giant I



Slika 6: *Primerjava statistike učenja, in sicer Train CER in Validation CER, za vse tri modele; vrednosti obeh črt naj bi bili čim bolj enaki*

Če primerjamo CER v odstotkih in na sliki statistiko učenja vseh treh modelov, lahko vidimo, da je zadnja, tretja verzija najustreznejša za nadaljnjo optično prepoznavo na celotnem delu. Hkrati smo ugotovili, da ni priporočljivo uporabiti svojega besedilnega modela za učenje, kar se je pokazalo pri verziji 2, saj je odstotek napak višji kot pri rabi obstoječega javnega modela (z večjo osnovno bazo podatkov besed). Vsekakor povečana osnovna baza podatkov besed pripomore tudi k manjšim odstotkom napak.

Za optično prepoznavo rokopisa smo na koncu izbrali skenograme, na katerih optična prepoznava še ni bila opravljena, in za postopek izbrali ustvarjeni besedilni model, ki ima najboljše vrednosti (model verzije 3). Kot razvidno s slike 7, je prepoznava rokopisa po opravljenem postopku zelo dobra; v samem besedilu bi bili potrebni le minimalni popravki.



Slika 7: *Končni izdelek optične prepoznave rokopisa na podlagi verzije 3 ustvarjenega besedilnega modela*

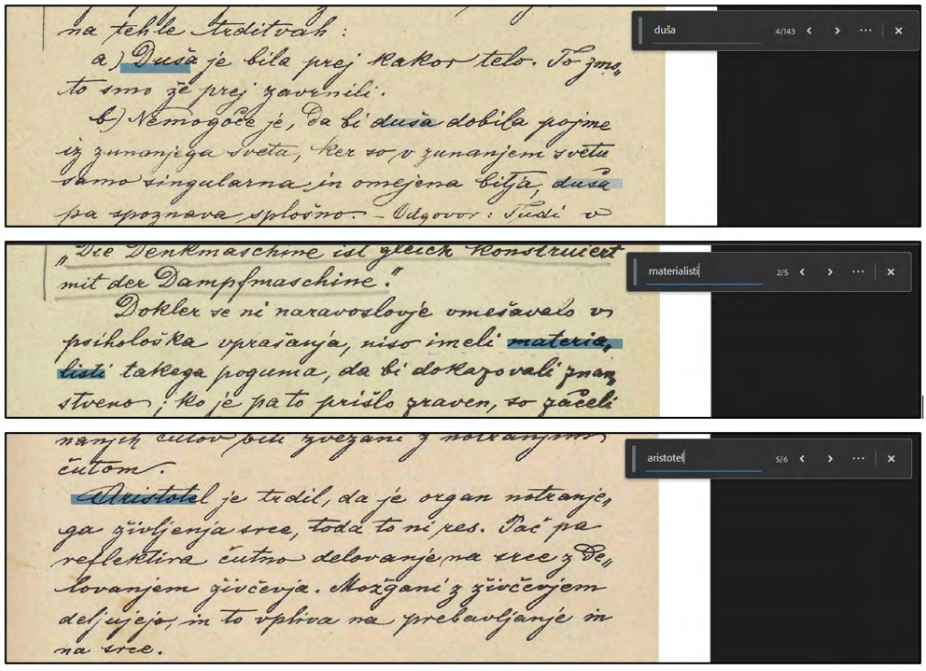
6 Izvoz iz Transkribusa

Za izvoz rezultatov iz Transkribusa nam vmesnik omogoča izbiro različnih formatov. Na voljo so nam:

- Images – orodje izvozi slike skenogramov.
- Docx files – orodje izvozi datoteko Microsoft Word, kjer so ločene vrstice združene. Na voljo je izvoz oznak (*tags*).
- Transkribus PDF – orodje izvozi iskalno datoteko PDF. Na voljo je izvoz oznak (*tags*).

- Text files (TXT) – orodje izvozi datoteko TXT z golim besedilom. Na voljo je izvoz celotne datoteke ali razdelitev na posamezne datoteke glede na strani besedila.
- Page XML – orodje izvozi datoteke mets, metadata in datoteke HTML za vse strani, ki so bile prepisane.
- Export structural elements to Mets[12] – orodje izvozi digitalni paket, ki vsebuje bistvene osnovne informacije o datoteki, zajete v shemi Mets. The Library of Congress (2025) navaja, da je shema Mets standardna oblika za kodiranje opisnih, upravnih in strukturnih metapodatkov, ki se nanašajo na objekte znotraj digitalne knjižnice in so izraženi v jeziku sheme XML Konzorcija svetovnega spleta (W3C[13]). Opcija je na voljo le z naročnino na Transkribus Scholar ali višjo.

Za trenutne potrebe objave na portalu Digitalne knjižnice Slovenije smo potrebovali Transkribus datoteki PDF in TXT. Po zagonu procesa smo na elektronsko pošto prejeli povezavo za prenos datotek.



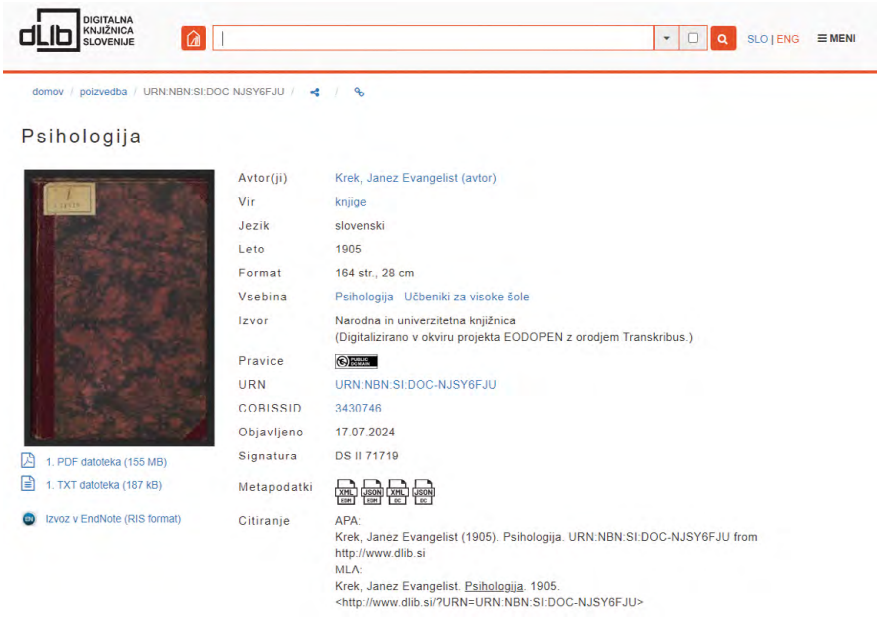
Slika 8: *Primeri iskanja v izvoženem formatu PDF; iskali smo besede: duša, materialisti in Aristotel*

[12] Mets ali Metadata Encoding and Transmission Standard.
[13] W3C ali World Wide Web Consortium.

Dodatna prednost datoteke PDF, ki smo jo izvozili iz Transkribusa, je ta, da sedaj uporabnik lahko tudi išče po besedilu, med drugim je iskanje omogočeno tudi po besedah, ki so deljene. PDF, ki smo ga pridobili z digitalizacijo v orodju Digitisation Manager, tega ni omogočal, saj optično branje znakov ni bilo izvedeno. Na sliki 8 je prikazanih nekaj primerov iskanja.

7 Objava v Digitalni knjižnici Slovenije in dodatno gradivo

Knjigo *Psihologija* smo po ustaljenih postopkih objavili na portalu Digitalne knjižnice Slovenije.[14] Ob objavi smo preverili prenesene datoteke in zajete metapodatke ter dopisali, da je bilo za digitalizacijo uporabljeno orodje Transkribus.



Slika 9: *Vpogled v objavljeno delo Psihologija na portalu dLib.si*

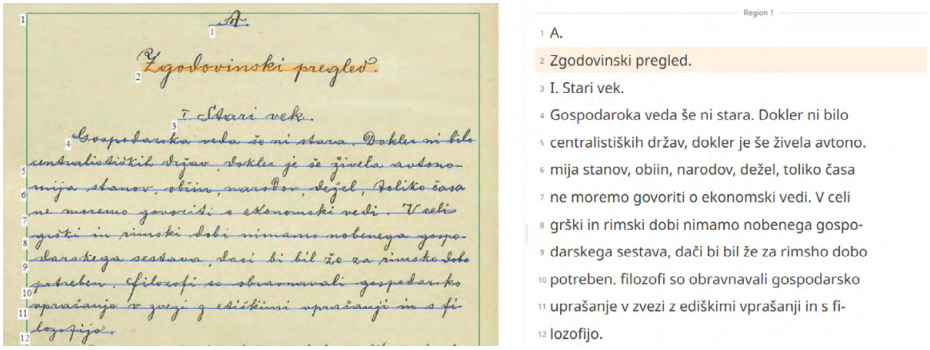
Transkribusov model smo v prvi fazi učili na podlagi pisave enega pisca semeniških predavanj, Janeza Evangelista Kreka. Zdelo se nam je smiselno, da v katalogu Narodne in univerzitetne knjižnice preverimo, ali je še kakšno delo po

[14] Delo *Psihologija* je na voljo na povezavi: https://www.dlib.si/details/URN:NBN:SI:DOC-NJSY6FJU

predavanjih Janeza Evangelista Kreka s podobno pisavo, ki bi ga lahko vključili in za optično prepoznavo rokopisa uporabili sedaj ustvarjeni besedilni model.

Po pregledu smo vključili v digitalizacijo in obdelavo v Transkribusu še tri dela, ki so sedaj že na voljo na portalu Digitalne knjižnice Slovenije in vsebujejo tudi avtomatski prepis, ustvarjen s pomočjo naučenega besedilnega modela. Ta dela so:
– *Kozmologija* (1906, COBISS-ID: 512178229)[15]
– *Naravna teologija* (1905, COBISS-ID: 512178485)[16]
– *Narodna ekonomija* (1906, COBISS-ID: 82431488)[17]

Pri slednjih treh delih je bilo opaženo, da se je kakovost prepoznanega besedila nekoliko razlikovala, saj besedilni model ni bil osnovan na pisavi teh del – pisava ni bila povsod enotna in se je nekoliko razlikovala od tiste, na kateri se je model učil. Vsekakor so bili rezultati mnogo boljši, kot če optične prepoznave rokopisa ne bi opravili (slika 10).



Slika 10: *Primer optične prepoznave dela Narodna ekonomija, kjer se pisava pisca nekoliko razlikuje od tiste, na kateri se je besedilni model učil; besedilo je še vedno dobro prepoznano, z minimalno količino napak*

8 Zaključki

Med testiranjem smo ugotovili, da lahko s Transkribusom uporabnikom omogočimo prebiranje rokopisov, iskanje po celotnem besedilu in nenazadnje tudi

[15] Delo *Kozmologija* je na voljo na povezavi: https://www.dlib.si/details/URN:NBN:SI:DOC-TPDI8QKU
[16] Delo *Naravna teologija* je na voljo na povezavi: https://www.dlib.si/details/URN:NBN:SI:DOC-L8T44RG9
[17] Delo *Narodna ekonomija* je na voljo na povezavi: https://www.dlib.si/details/URN:NBN:SI:DOC-XOSOQ9FI

podatkovno rudarjenje, ki je bilo do sedaj onemogočeno zaradi neprepoznanega besedila na skenogramih rokopisnih gradiv. Uporabnikom lahko na ta način omogočimo hitrejše in učinkovitejše iskanje določenih tem, besed, oseb, krajev in dogodkov v dokumentih ter povečamo možnosti za raziskovanje.

Opisani postopki niso zahtevni, potrebno je nekoliko več ročnega dela in vloženega časa, še posebej, ko gre za večjo količino gradiva, ki je bilo do sedaj nedostopno. Testiranje je bilo v Narodni in univerzitetni knjižnici izvedeno prvič in ponuja zametke za nadaljnje delo na tem področju, ne le za rokopisno gradivo, ampak tudi drugo, pri čemer bi se lahko testirali tako postavitveni kot besedilni modeli.

Bistveno je tudi zavedanje, da nam današnja tehnologija optične prepoznave rokopisov in optične prepoznave znakov odpira nove možnosti za razvoj in stik s preteklostjo. Pri samem razvoju so bistveni tudi usposabljanje in poznavanje omejitev, kot tudi prednosti ter spremljanje napredka, ki ga prinaša UI tudi za to področje bibliotekarstva.

S konkretnim primerom smo želeli pokazati, da so na voljo nove in dodatne možnosti, ki rokopisno gradivo naredijo bolj uporabno in dostopno vsakomur. Na ta način se širi tudi razvoj optične prepoznave besedila v slovenskem jeziku, ki je trenutno v Transkribusu še vedno pomanjkljiva. Menimo, da je to področje vredno dodatnih raziskav in testiranj, ki bi rokopisno gradivo ali katerokoli gradivo, ki ga je trenutno nemogoče procesirati z običajnimi postopki optične prepoznave znakov, približalo širši publiki ter tako omogočilo prenos znanja in zgodovine iz preteklega časa.

Viri in literatura

Dolenec, I. (2024). Krek, Janez Evangelist. *Slovenska biografija*. Slovenska akademija znanosti in umetnosti, Znanstvenoraziskovalni center SAZU. https://www.slovenska-biografija.si/oseba/sbi302887/

Hodel, T. (2022). Chapter 6: Supervised and unsupervised: approaches to machine learning for textual entities. V Jaillant, L. (ur.), *Archives, access and artificial intelligence: working with born-digital and digitized archival collections*, (157–177). transcript Verlag, Bielefeld University Press. https://www.jstor.org/stable/jj.11425482

Klasinc, J., Kragelj, M., Grčar, U., Zorko, T., Šavnik, M., Malešič, J., Vovk, D., Kozjek, A., in Krstulović, Z. (2023). *Enotne zahteve in postopkovni model izvajanja interne digitalizacije knjižničnega gradiva v Narodni in univerzitetni knjižnici, različica 1.1*. Narodna in univerzitetna knjižnica.

The Library of Congress (15. 9. 2025). *METS: Metadata Encoding & Transmission Standard*. https://www.loc.gov/standards/mets/

Milioni, N. (2020). *Automatic transcription of historical documents: Transkribus as a tool for libraries, archives and scholars* [Magistrsko delo]. Uppsala universitet, department of ALM. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-412565

Muehlberger, G., Seaward, L., Terras, M., Oliveira, S. A., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., ... Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, *75*(5), 954–976. https://www.emerald.com/insight/content/doi/10.1108/JD-07-2018-0114/full/html

Nockels, J., Gooding, P., in Terras, M. (2024). The implications of handwritten text recognition for accessing the past at scale. *Journal of Documentation*, *80*(7), 148–167. https://www.emerald.com/insight/content/doi/10.1108/JD-09-2023-0183/full/html

Peter Handke Notizbücher. (24. 6. 2024). *Projektinformation*. https://edition.onb.ac.at/fedora/objects/o:hnb.red-projectinformation/methods/sdef:TEI/get?mode=info#toc-2-2

Prebor, G. (2024). From digitization and images to text and content: Transkribus as a case study. *Manuscript Studies*, *9*(1), 72–89. https://doi.org/10.1353/mns.2024.a930877

READ-COOP SCE, (b. d. a). *A cooperative to unlock our written past*. https://readcoop.org/

READ-COOP SCE. (b. d. b). *Unlock the past with Transkribus*. https://www.transkribus.org/

READ-COOP SCE. (b. d. c). *Our story*. https://readcoop.eu/our-story/

READ-COOP SCE. (b. d. č). *+Searching handwritten manuscripts at Greifswald University Library*. https://www.transkribus.org/blog/searching-handwritten-manuscripts-at-greifswald-university-library

Slovenska biografija. (2013). Slovenska akademija znanosti in umetnosti, Znanstvenoraziskovalni center SAZU. https://www.slovenska-biografija.si

Smernice za digitalizacijo knjižničnega gradiva. (2010). Narodna in univerzitetna knjižnica. https://www.dlib.si/details/URN:NBN:SI:DOC-ZU0LQ5EO

Terras, M. (2022). Chapter 7: Inviting AI into the archives: the reception of handwritten recognition technology into historical manuscript transcription. V Jaillant, L. (ur.), *Archives, access and artificial intelligence: working with born-digital and digitized archival collections* (179–204). transcript Verlag, Bielefeld University Press. https://www.jstor.org/stable/jj.11425482

Andreja Hari

Narodna in univerzitetna knjižnica, Turjaška ulica 1, 1000 Ljubljana
e-pošta: andreja.hari@nuk.uni-lj.si