

PRIMER UPORABE PROGRAMA TRANSKRIBUS IN IZDELAVA MODELA ZA AVTOMATSKO OPTIČNO PREPOZNAVANJE ZNAKOV ZA POENOSTAVLJENO TRANSKRIBIRANJE ROČNO PISANE GOTICE

Jože Glavič

Zgodovinski arhiv Ljubljana, Enota za Dolenjsko, Novo mesto, Slovenija
joze.glavic1@guest.arnes.si

Izvleček:

V prispevku je predstavljeno orodje, ki s pravim pristopom olajša proces transkribiranja. Na podlagi praktičnega primera transkribiranja korespondence družine Terpinč, ki jo hrani Zgodovinski arhiv Ljubljana, sta prikazana proces transkribiranja korespondence in izgradnja modela HTR. Izgradnja modela HTR predstavlja osnovo za nadaljnji proces avtomatiziranega transkribiranja večje količine gradiva. Namen članka je prikazati prednosti in izzive uporabe programskih orodij na zgodovinskih virih, hkrati pa umestiti skupno uporabo procesa transkribiranja in tehnoloških orodij v širši koncept razvoja digitalne humanistike na področju arhivistike in obdelave arhivskih virov.

Ključne besede:

Transkribus, gotica, transkripcija, optično prepoznavanje znakov, orodje za transkribiranje

Abstract:

Use case of the Transkribus Software and the Creation of a Model for Automatic Optical Character Recognition for Simplified Transcription of Written Gothic

The article focuses on a tool, which simplifies the process of transcribing. The process of transcribing and creating a HTR model is demonstrated on a case study based on archival records of the Terpinč family, which is kept in the Historical Archives Ljubljana. The construction of the HTR model forms the basis for a further process of automated transcription of a large amount of material. The purpose of the article is to outline the advantages and challenges of using software tools on historical resources, while at the same time establishing how the transcription process and technology tools can be contextualized in the broader concept of digital humanities, the field of archival science and archival resource processing.

Key words:

Transkribus, Gothic script, transcription, optical character recognition, transcription tool

1. UVOD

Arhivi in arhivistika se poleg hrambe arhivskega gradiva ukvarjajo tudi z njegovo dostopnostjo in promocijo. V kontekst dostopnosti arhivskega gradiva se umešča tudi pričujoči prispevek. Da je arhivsko gradivo dostopno čim širši množici uporabnikov, poskrbijo tako sami arhivi kot tudi arhivska društva s posebnimi objavami raznovrstnih virov iz različnih zgodovinskih obdobj v različnih oblikah.

Pri uporabi in obdelavi starejšega arhivskega gradiva večkrat naletimo tudi na pisavo, ki se zaradi preteka časa in specifik zgodovinskega obdobja, v kateremu je nastala, razlikuje od današnje. Včasih manj, včasih bolj. Tipičen primer takšne pisave je gotica (nemška kurenta), ki jo zna velik del zgodovinarjev in arhivistov vsaj pogojno zadostno brati, večini smrtnikom pa predstavljata branje in razumevanje gotice poseben izziv. Za reševanje tega izziva poznamo proces transkribiranja.¹

Proces transkribiranja, tj. spreminjanja tuje pisave v današnjo razumljivo in vsesplošno uporabljeno latinico, zahteva ne samo poznavanje pisave starejših obdobj, ampak zaradi zgodovinskega razvoja tudi poznavanje tujih jezikov (nemščine, latinščine itd.), v katerih so arhivski viri zapisani. Potrebno pa je izpostaviti, da se pisava ne razlikuje samo od obdobja do obdobja, v katerem je bilo arhivsko gradivo ustvarjeno, ampak tudi od osebe, ki je dokumente zapisovala, kar si lahko predstavljamo s primerom iz šolskih klopi – branjem sošolčevih zapiskov. Proces transkribiranja pa ne zahteva le usposobljenih ljudi, ampak je lahko tudi časovno potraten. Večja ko je količina gradiva, večji časovni vložek je potreben za transkripcijo.

Z razvojem tehnologije, tako strojne kot tudi programske opreme, se je pojavila možnost poenostavitve procesa transkribiranja in njegova avtomatizacija. Začetki procesa optičnega prepoznavanja znakov (ang. Optical Character Reader (OCR)), ki omogoča prepoznavanje in pretvorbo tipkanih ali tiskanih znakov v digitalno obliko (Optical character recognition, 2020), segajo že v leto 1870(!) (Schantz, 1982), prvi aplikativni procesi pa v 50. leta 20. stoletja (The First OCR System: "GISMO", 2004–2020). V primeru ročnih zapisov pa je zadeva težavnejša zaradi specifičnosti ročne pisave, predvsem zaradi odstopanja od normativov in unikatnosti posameznikove pisave. Zaradi strojnih omejitev računalnikov je tehnologija napredovala počasi, kljub začetnemu skupnemu razvoju s tehnologijo OCR, zato njen razvoj v 80. in 90. letih 20. stoletja ni prinesel aplikativne uporabnosti (Muehlberger, 2019).

V začetku 21. stoletja je razvoj na področju umetne inteligence, naprednega prepoznavanja vzorcev in povečevanja računalniške procesorske moči neposredno vplival tudi na razvoj optičnega prepoznavanja ročno pisanih dokumentov. Eden izmed rezultatov tega procesa je izdelava programa Transkribus, razvila ga je skupina strokovnjakov, ki jih je vodil Günter Muehlberger z Univerze v Innsbrucku. Program je nastal v okviru evropskega projekta READ (Recognition and Enrichment of Archival Documents) (READ, 2016). Eden izmed pglavitnih ciljev projekta je narediti arhivsko gradivo bolj dostopno z razvojem ustreznih orodij in tehnologij.

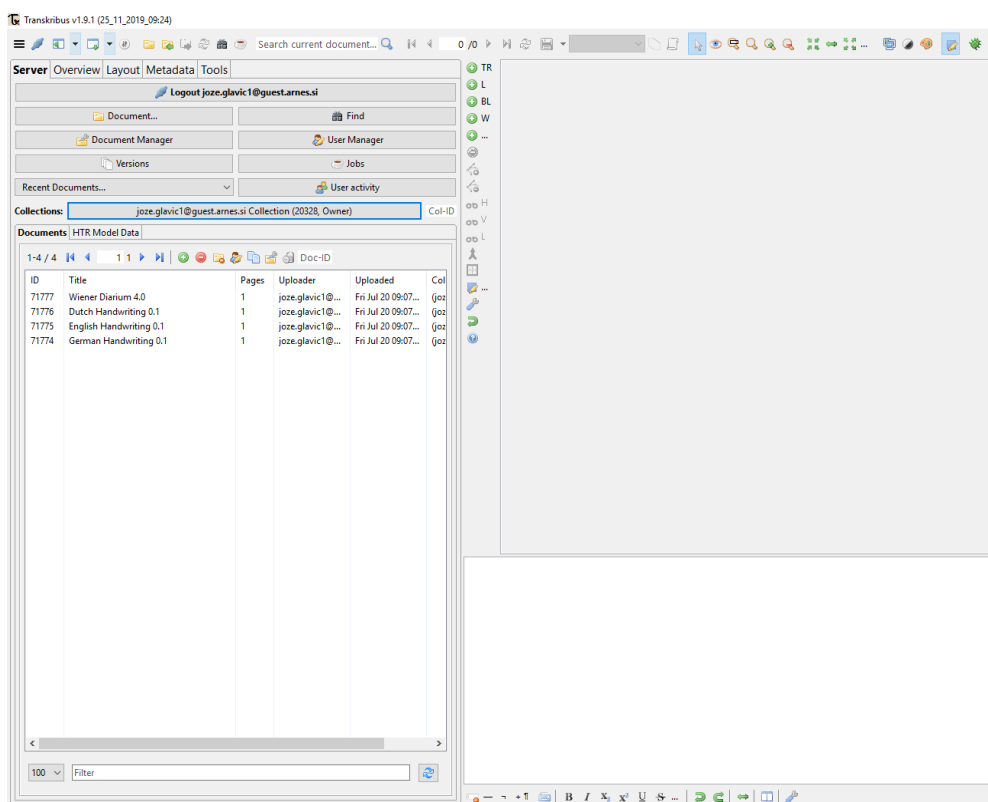
¹ S pomočjo spletnega portala Fran Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, naslednika portala Slovenske in besedilne zbirke, najdemo terminu transkripcija sinonim v besedi prečrkovanje. Prečrkovanje pomeni zapis črk določene pisave s črkami druge pisave. Zaradi vsesplošne razširitve tega termina bom tudi sam v nadaljevanju uporabljal izraza transkripcija in transkribiranje (proces uporabe transkripcije).

2. PREIZKUS ORODJA V PRAKSI

Transkribus je programsko orodje, ki se uporablja za transkribiranje dokumentov, objavljenih v znanstvenih izdajah, za ustvarjanje in usposabljanje modelov, ki omogočajo avtomatsko pretvorbo ročno pisanih tekstov in vsesplošno pomoč pri transkribiranju. Njegova uporabnost pa se izkaže predvsem v primeru, ko je potrebno transkribirati večjo zbirko ročno pisanih dokumentov.

V Zgodovinskem Arhivu Ljubljana hranimo korespondenco družine Terpinč. Pisma so del večjega fonda Graščina Fužine (SI_ZAL_LJU/0296). Sama korespondenca vsebuje pisma med sorodniki in prijatelji ter tudi nekaj poslovnih pisem. Pisma in korespondente je v 41. izdaji Virov podrobneje opisala in umestila N. B. Kodrič (2018), jih tudi uredila, transkribirala vire ter napisala spremno študijo in kazalo. Tema tega članka ni vsebina pisem in njihova umestitev v zgodovinski prostor, ampak sam proces izgradnje tako imenovanega modela HTR s programom Transkribus, ki omogoča poenostavljen, avtomatski in hitrejši proces transkribiranja pisem, ki so ostala še netranskribirana.

Na tem mestu naj omenim, da je za uporabo programa predhodno potrebno same vire pretvoriti v digitalno obliko. V primeru korespondence družine Terpinč so bila pisma že predhodno digitalizirana in hranjena na strežnikih Zgodovinskega arhiva Ljubljana. Idejna zasnova projekta, ki smo ga izvedli, je bil prenos digitaliziranih in že transkribiranih besedil v program Transkribus. Tam je bilo potrebno povezati transkripcije z digitaliziranimi posnetki virov, nato pa z funkcionalnostjo programa zgraditi model HTR+, s katerim bi nato še ne transkribirane digitalizate transkribirali avtomatsko z omenjenim programom. O tem, kako je bil postopek izveden, je pojasnjeno v nadaljevanju.

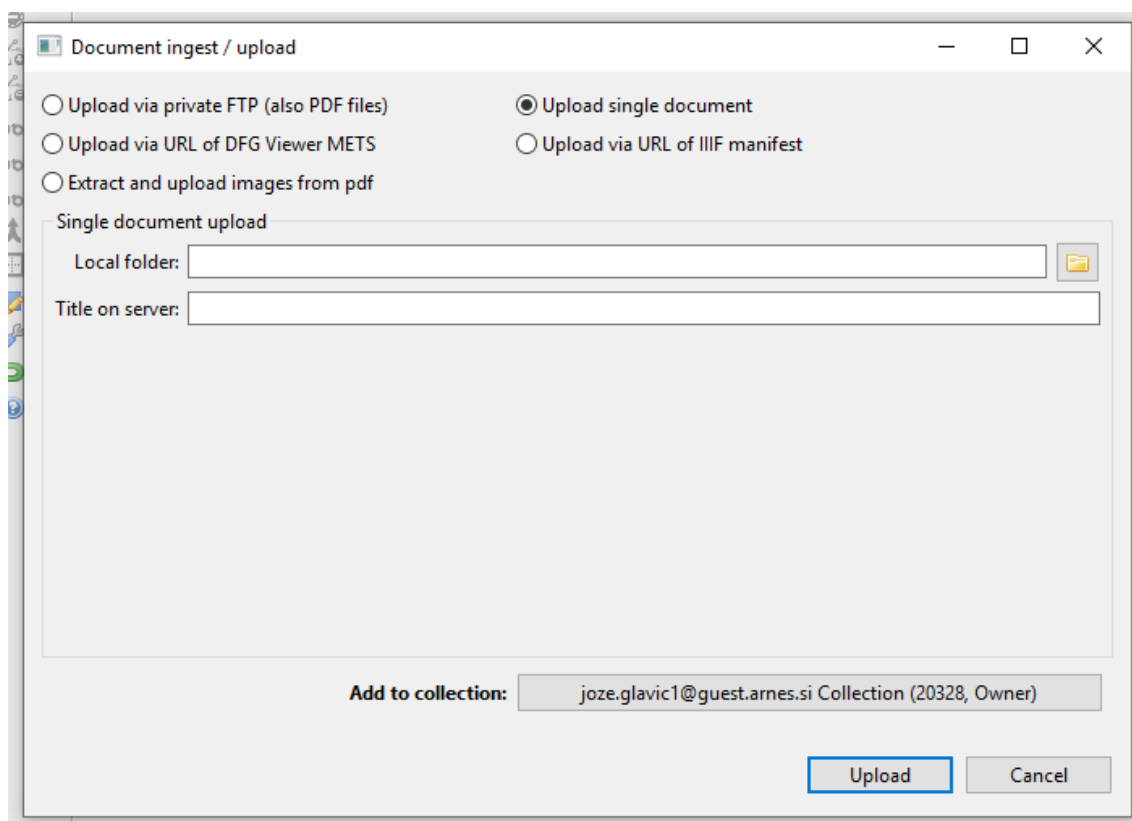


Slika 1: Zajeta slika programa Transkribus, osnovni pregled

3. POSTOPEK UPORABE PROGRAMA ZA TRANSKRIBIRANJE KORESPONDENCE DRUŽINE TERPINC IZ GOTICE V LATINICO

Program Transkribus je dostopen na spletni strani <https://transkribus.eu/Transkribus/>. Na strani se moramo predhodno registrirati z e-mail naslovom, nato pa program prenesemo s spleta in ga enostavno namestimo na sistem. Program zaženemo in se vanj z ustvarjenim registriranim e-mailom vpišemo. Registracija in profil programa omogočata shranjevanje in nadzor nad delom, ki ga opravljamo v programu. S profilom nam je omogočeno hkratno urejanje in upravljanje več zbirk hkrati, omogočeno pa je tudi hkratno sodelovanje več različnih uporabnikov programa Transkribus, tako da lahko delajo na transkripciji iste zbirke digitalizatorov.

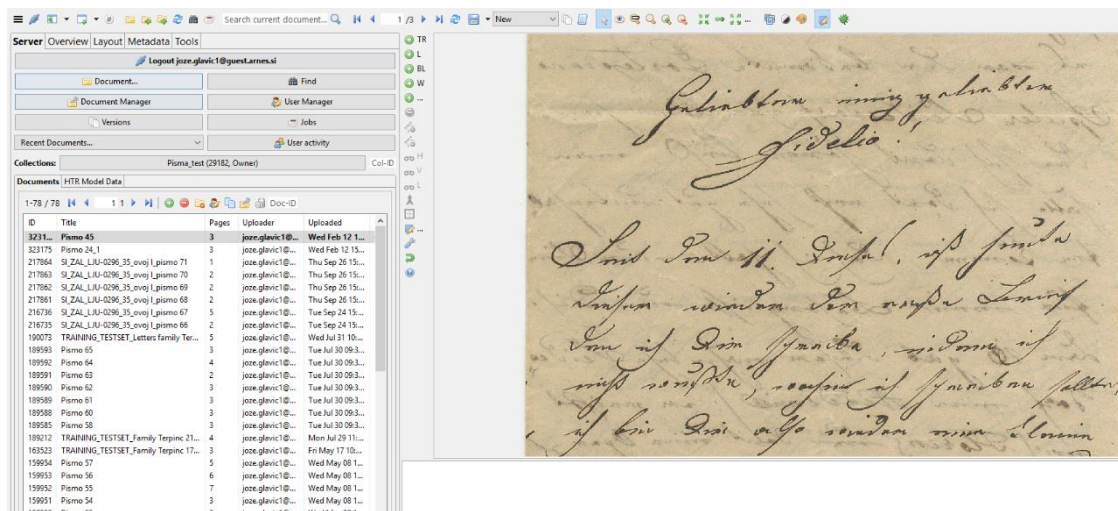
Program omogoča lokalno obdelovanje digitalizatorov, to pomeni z računalniškega diska, ali obdelovanje s strežnika (Transkribus Cloud Service, 2017), ki ga naložimo z lokalnega računalnika preko vmesnika (primer viden na sliki 2). Če želimo izgraditi HTR-model, ki omogoča avtomatsko transkribiranje, je potrebno izbrati opcijo, pri kateri naložimo datoteke na Transkribusov strežnik. Na strežniku naloženi digitalizati so zasebni in vidni samo uporabniku, ki jih je naložil. Po potrebi se lahko uporaba le-teh dovoli točno določenim uporabnikom, za katere želimo, da sodelujejo z nami.



Slika 2: Zajeta slika uvoza

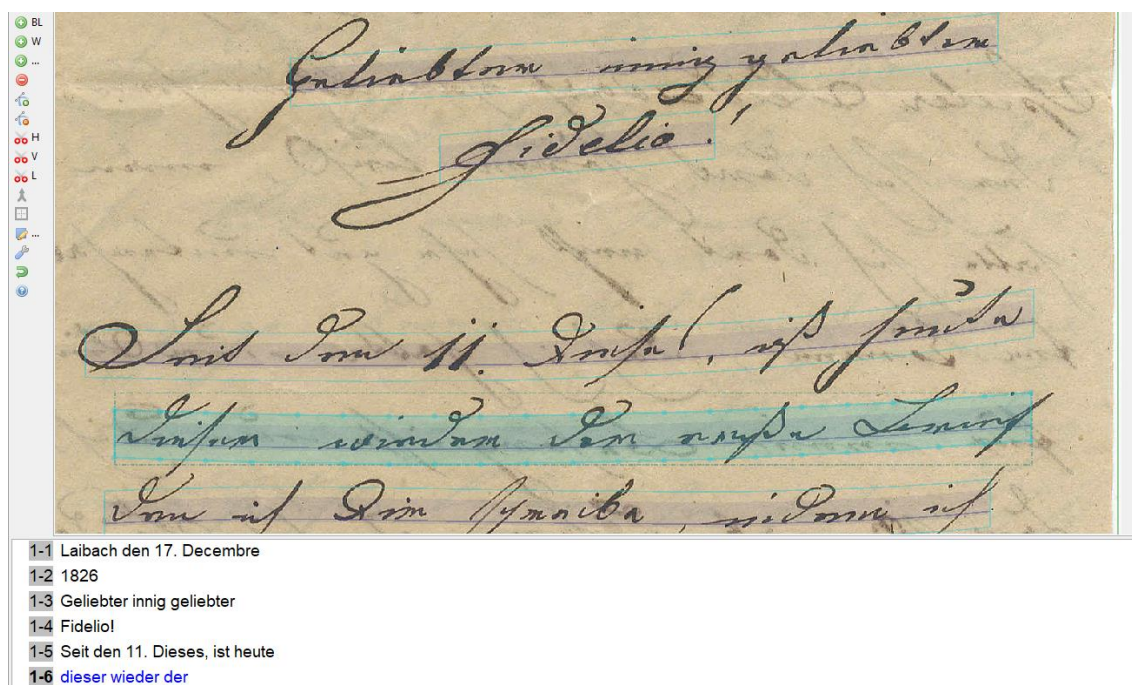
Način shranjevanja gradiva v programu je odvisna od primera do primera oziroma od potreb uporabnika. V našem primeru smo naložili vsako pismo posebej, lahko pa uporabnik tudi naloži vsa pisma naenkrat v eni mapi, kar omogoča hitrejšo avtomatsko obdelavo.

Ko so digitalizati naloženi, se lotimo transkribiranja. V levem desnem oknu s seznama izberemo digitalizat, ki se nam pokaže v desnem zgornjem oknu. Primer je viden na sliki 3.



Slika 3: Zajeta slika transkribiranja

Da program razloči, kaj je besedilo in kaj ozadje, najprej v zavihku orodja (ang. Tools) izberemo analizo postavitve teksta. Tu program v ozadju določi tekst in vrstice besedila. Te tudi ustrezno zaporedno oštevilči. S tem postopkom se digitalizat poveže z urejevalnikom besedila oziroma se poveže besedilo s sliko.



Slika 4: Zajeta slika vnosa teksta

Nato ročno vnesemo ustrezno besedilo v urejevalnik teksta, ki je lociran v desnem spodnjem polju programa. Prednost takega načina je v tem, da lahko hkrati gledamo digitalizat kot tudi tipkamo besedilo. Obenem lahko digitalizat približamo, oddaljimo, usmerimo pozornost na specifično vrstico ter si s tem olajšamo prepoznavanje znaka ali besede. Sam program omogoča tudi vnašanje posebnih znakov, kot so tisti, ki jih standardna tipkovnica ne pozna (npr. ß,—,ω,ξ).

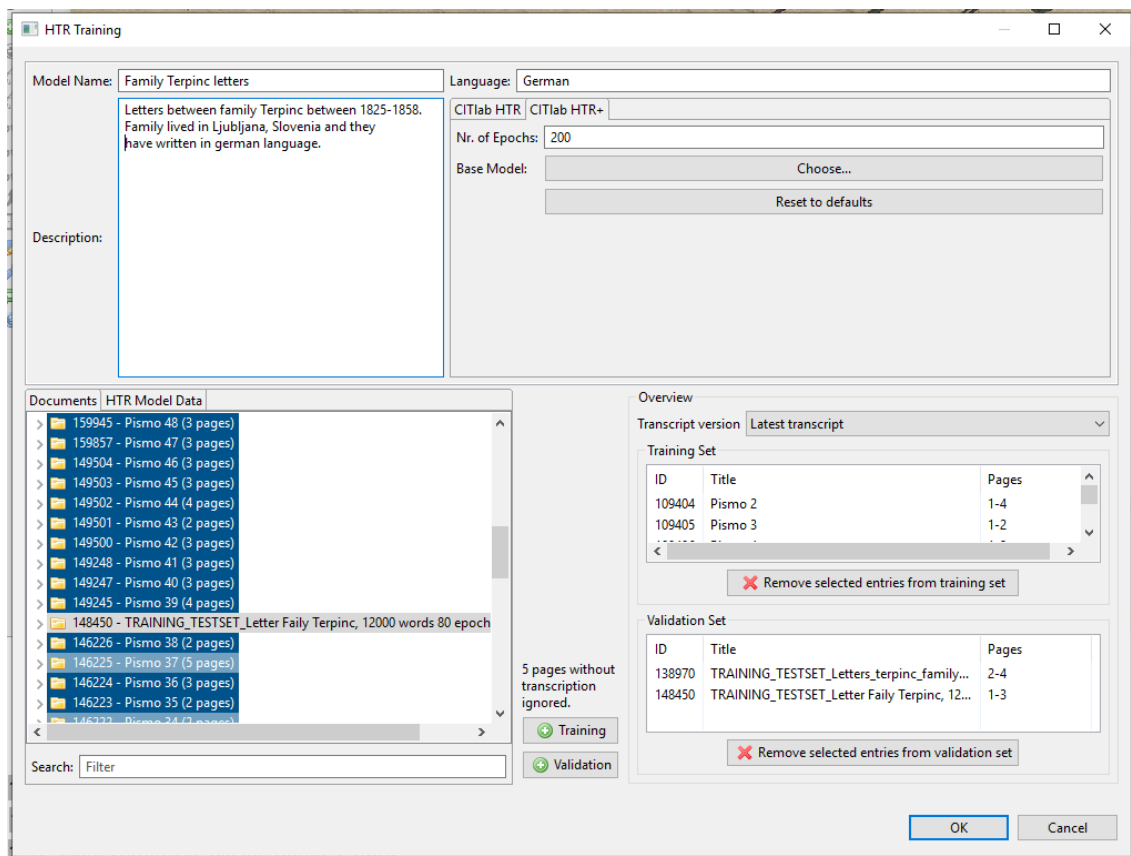
4. IZGRADNJA MODELA IN AVTOMATIZACIJA TRANSKRIBIRANJA

V primeru večjega števila dokumentov, ki jih je napisal isti pisar, dokumentov, kjer je pisava slogovno podobna ali napisana v istem obdobju, lahko samo transkribiranje pospešimo s tako imenovanim procesom optične prepoznave ročne pisave (ang. Handwritten Text Recognition (HTR), ki je vgrajen v program Transkribus. Program ima namreč pripravljenih več modelov splošno razširjenih pisav, med drugimi na primer nemško tiskano gotico iz 19. stoletja, nemško kurzivo iz 18. stoletja in podobno. Vendar zaradi dekorativnih posebnosti in unikatnega sloga posameznega pisarja lahko prihaja do večjih odstopanj v obliki črk kot tudi v načinu zapisa, tako da pri uporabi privzetih modelov velikokrat prihaja do napak in neustreznega transkribiranja.

V primeru korespondence družine Terpinc je bilo tako potrebno izgraditi nov model. Za izgraditev modela je bilo potrebno najprej vključiti možnost izgradnje modela v samem programu. S tem v zvezi smo se obrnili na ustvarjalce programa preko e-maila (email@transkribus.eu). Z vključeno opcijo smo potem nadaljevali delo. V primeru korespondence družine Terpinc smo uvozili digitalizate pisem, nato pa s preprosto metodo »kopiraj in prilepi« (ang. copy-paste) prenesli (s strani Nataše Budne Kodrič) že transkribirano besedilo v ustrezna polja tekstovnega urejevalca programa.

Predpogoj izgradnje modela je zadostna količina transkribiranih besedil. Priporočen je vnos med 5000 in 15.000 transkribiranih besed. V primeru korespondence družine Terpinc smo za okoli 20.000 besed potrebovali 50 pisem. Sama količina besed je odvisna od kompleksnosti in istovrstnosti pisave ter tudi števila avtorjev. Pri obdelovanju korespondence moramo biti vedno pozorni na to, da je ponavadi vključeno večje število pisarjev, saj so poleg glavnega korespondenta prisotni tudi dopisovalci. Večje ko je število besed, večjo natančnost doseže izgrajeni model, posledično je manj napak pri samem transkribiranju.

V programu smo nato izvedli izgradnjo modela. V zavihku »Orodja« (ang. Tools) smo izbrali gumb »Treniraj« (ang. Train). V oknu smo določili pisma, namenjena uporabi pri izgradnji modela, ter tista pisma, ki bodo uporabljena za validacijo in izračun pravilnosti transkribiranja. Določili smo način izgradnje modela CITlab HTR+ in število prehodov 200. S številom prehodov ali epoh (ang. Epochs) označujemo, kolikokrat program prebere niz pisem, ki smo ga izbrali za izgradnjo modela. Več kot je epoh, dalj časa traja proces učenja. Izgradnja modela HTR deluje na tehnologiji umetnih nevronske mreže. Za razlago delovanja nevronske mreže pa bi potrebovali kar nov članek.



Slika 5: Zajeta slika, ki prikazuje določitev nastavitev in vzorca za izgradnjo modela (vir: lastni)

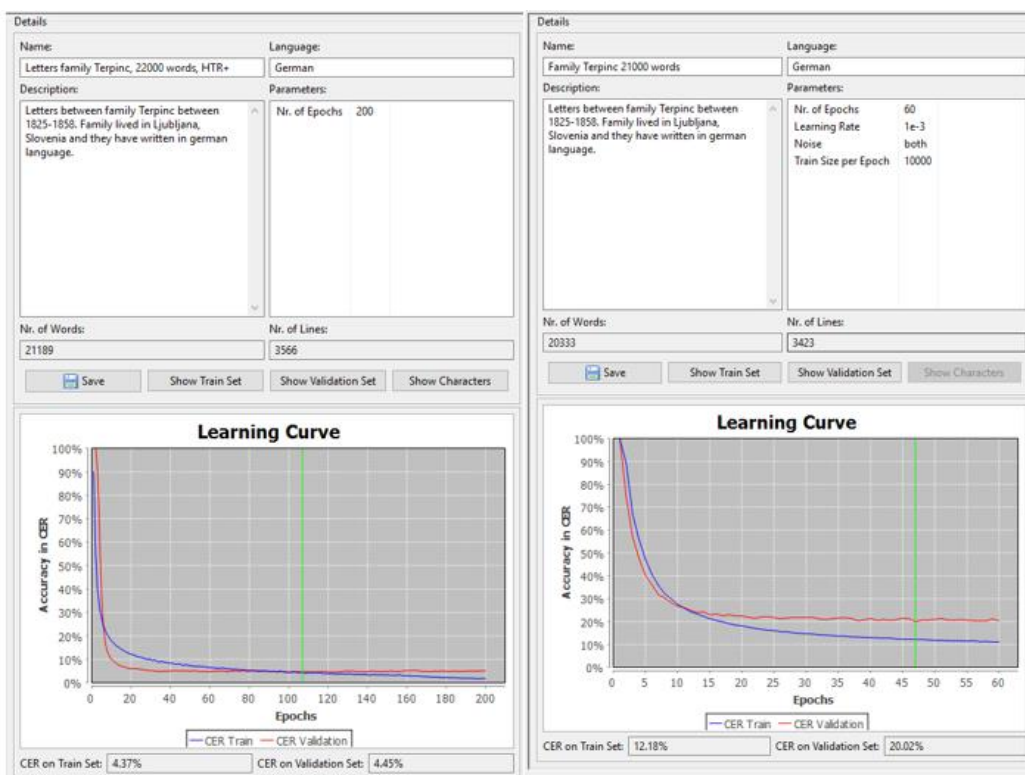
Pri načinu izgradnje modela smo sicer preizkušali več načinov. Poleg samega HTR+ tudi osnovnega HTR z različnim številom epoh. Najboljše rezultate smo dosegli z modelom CITlab HTR+. V primerjavi s HTR smo pri istem nizu dokumentov dosegli boljše rezultate ob znatnem prihranku časa.

Ko določimo vse parametre izgradnje modela, damo ukaz izgradnje. Sama izgradnja modela poteka na strežnikih ustvarjalcev programa. Samo treniranje/preskušanje modela poteka tudi do 24 ur. Potek lahko spremljamo v zavihku »strežnik« (ang. Server) pod gumbom »opravila« (ang. Jobs). Na sliki 6 je razvidno, da večje kot je število epoh, daljši je čas izgradnje. Prav tako izbira načina modela HTR+ prihrani čas kljub večjemu številu epoh. Primerjava: za isto število besed (približno 21.000) je izgradnja modela načina HTR trajala 1 dan, 4 ure in 58 minut, medtem ko je z modelom HTR+ vse skupaj trajalo 5 ur in 3 minute.

Type	State	Pa...	User...	Description	Errors	Created	Started
CITlab HTR Training: Letters_family_early_19th	FINISHED		joze.gla...	Done	0	21.03.2019 15:27:24	21.03.2019
CITlab HTR Training: Family Terpic 21000 words	FINISHED		joze.gla...	Done, duration: 1d 4h 58m 33s 320ms	0	29.07.2019 10:57:07	29.07.2019
CITlab HTR Training: Family Terpic 17000 words	FINISHED		joze.gla...	Done, duration: 1d 2h 33m 34s 449ms	0	17.05.2019 10:30:07	17.05.2019
CITlab HTR Training: Letter Family Terpic, 12000 words 80 epochs	FINISHED		joze.gla...	Done, duration: 1d 11h 18m 56s 327ms	0	15.04.2019 10:10:19	15.04.2019
CITlab HTR Training: Letter Family terpic 9000, 50 epochs	FINISHED		joze.gla...	Done, duration: 1d 1h 30m 25s 769ms	0	03.04.2019 10:11:09	03.04.2019
CITlab HTR Training: Letters family Terpic	CANCE...		joze.gla...		0	21.03.2019 15:18:55	
CITlab HTR+ Training	FINISHED		joze.gla...	Done, duration: 5h 3m 40s 918ms	0	31.07.2019 10:30:58	31.07.2019
CITlab Handwritten Text Recognition	FINISHED	2	joze.gla...	Done	0	18.12.2018 13:20:55	18.12.2018

Slika 6: Zajeta slika opravljenih opravil izgradnje modela (vir: lastni)

Rezultat izgradnje je model, kateremu je izmerjena tudi natančnost, kar pogojuje samo uporabnost modela. Natančnost modela se meri v stopnji napak v znakih ali CER (ang. Character Error Rate). Izražena je v številu odstotkov, kjer je model napačno interpretiral znak glede na testni niz, ki smo ga predhodno določili. Na sliki 7 vidimo graf, kjer je prikazana natančnost dveh modelov. Na grafu je najpomembnejša modra črta oziroma rezultat »CER on Validation Set«, kjer vidimo, kako se je program odrezal na testnih dokumentih.



Slika 7: Zajeta slika primerjave natančnosti dveh modelov, zgrajenih na isti osnovi pisem, z različnimi nastavitvami

Razlika v načinu izbire modela pokaže še eno podrobnost. Kljub daljšemu časovnemu obsegu je model HTR izgradil model, ki je napravil v celotnem tekstu kar 20 % napak. Model HTR+ pa je s krajšim časovnim obsegom napravil le 4,5 % napake. Pravilna izbira modela tako omogoča ne samo prihranek časa, ampak tudi kvalitetnejšo transkripcijo z manjšim odstotkom napak v celotnem tekstu. Za oris: 4,5-odstotna nenatančnost pomeni, da se v pismu na 4 straneh, z okoli 550 besedami, napačno transkribira le okoli 25 znakov ali 5 besed.

Ko imamo izgrajen model, tega enostavno uporabimo na naloženih in še ne transkribiranih dokumentih. Kot je že omenjeno zgoraj, sam postopek ni brez napak. V besedilih pride do odstopanj v nekaterih podobnih črkah. Praksa je pokazala, da je v primeru, ko odstotek napak preseže 10 %, smiselno transkribirati ročno, to je brez avtomatskega modela, saj bi za popravilo napak več časa porabili, kot pa bi ga prihranili (Muehlberger, 2019). 95,65 odstotkov pravilna prepoznava znakov pri posameznem dokumentu, kot v primeru korespondence družine Terpinč, je več kot odličen rezultat za nadaljevanje, saj prihranimo čas, kljub temu da vsak dokument, ki je avtomatsko transkribiran, pregleda in popravlja še arhivist. Transkribus je program, ki nam olajša samo transkribiranje in pohitri proces. A ni tu zato, da nadomesti človeški faktor.

5. POGLED NAPREJ ALI NAMESTO ZAKLJUČKA

Treba je omeniti, da pri samem programu nismo omejeni na vrsto, tip, starost ali jezik pisave, čeprav je v naslovu omenjena gotica (Leifert, 2016). Tehnologija namreč omogoča prepoznavanje pisav vse od srednjeveških listin do ljubezenskih pisem iz 19. stoletja. Le zadostno količino dokumentov je potrebno imeti.

Prikazano orodje Transkribus sicer ne nadomesti človeka – strokovnjaka, lahko pa s pravilnim pristopom in procesom znatno pohitri transkribiranje večje količine gradiva. Če se pri načrtovanju transkribiranja odločimo, da ne rabimo izdelati modela in uporabiti avtomatske transkripcije, je program še vedno koristno orodje, ki olajša transkribiranje. Prednosti uporabe progama je več: hkratno transkribiranje korpusa dokumentov več uporabnikov hkrati, pregled dokumenta, ki se transkribira, kot tudi hkratni pogled na zaslon. Prav tako nismo omejeni s samo naklado. Možnost hkratnih popravkov, delnih objav in postopne gradnje večjih digitalnih izdaj je kot češnja na vrhu torte.

Mogoče najpomembnejša prednost pa je izvoz transkripcij. Danes poznamo ne samo objave virov v fizičnih publikacijah, ampak tudi objave virov v novih formatih. Spletna tehnologija se je razvila do te mere, da omogoča objave virov s številnimi prednostmi. To so dostopnost, iskanje po tekstu, vizualizacija podatkov.

Transkribus v ta namen omogoča izvoz opravljenih transkripcij ne samo v čisti besedilni obliki, ampak tudi take z naprednimi možnostmi. Tako omogoča izvoz v formatu PDF, kjer sta povezna besedilo in tekst, v formatu DOCX, kjer lahko nadalje oblikujemo besedilo. Mogoče najpomembnejša opcija pa je možnost izvoza XSLT na podlagi standarda TEI (Text Encoding Initiative) za označevanje besedil v formatu XML² (eXtensible Markup Language). Sam izvoz pa omogoča tudi standard za opisovanje metapodatkov METS (Metadata Encoding and Transmission Standard). Prav možnosti izvoza v različnih formatih, prilagojenih spletni tehnologiji, olajšajo in poskrbijo za lažji prehod iz fizičnega dokumenta in njegovega digitalnega posnetka v obliko, ki ni samo bolj dostopna, ampak tudi bolj interaktivna, s tem pa tudi bolj uporabna in razširjena.

² XML je označevalni jezik, podoben HTML (Hyper Text Markup Language), ki omogoča zapisovanje besedil v formatu za opisovanje strukturiranih podatkov. Značilna je uporaba »<« »>« oklepajev.

Poleg vseh naštetih prednosti pa je pomemben tudi finančni in ekološki vidik, saj odpadejo vsi materialni stroški.

Naslednji korak v obdelavi korpusa pisem družine Terpinč je spletna objava, taka, ki sledi tako standardom humanistike kot tudi spletnim tehnologijam. Uporaba in integracija že razvitih spletnih standardov, ogrodij (ang. Framework), kot so na primer že omenjeni TEI, XML in IIIF (International Image Interoperability Framework)³, poskrbijo za kvalitetno objavo, ki lahko kljubuje zobu časa tudi, ko se tehnologija spreminja dnevno. O primerih dobre prakse že govori Pančur (2016), ko predstavi spletno objavo Popisi prebivalstva Slovenije 1830–1931. Lepo pa tudi predstavi uporabo tehnologije XML in TEI na primeru spletnega portala Zgodovine Slovenije – Sistory (Pančur, 2018).

Z razvojem tehnologije se je uporaba računalniških orodij razširila tudi v znanstvene vede in stroke, kjer so bila prej zapostavljena. Razvoj tehnologije prepoznavanja ročne pisave tako posredno poveča dostop do samega gradiva ter s tem omogoča uporabnikom hitro in učinkovito iskanje po temah, osebah, krajih in dogodkih. Področje tako imenovane digitalne humanistike se danes razvija tudi v Sloveniji, kar kažejo tudi specializirani študijski programi na univerzah. Eden izmed kazalcev razvoja so tudi tako imenovane digitalne edicije. Vogeler (2016) je digitalne izdaje označil kot pionirje na področju digitalne humanistike. Pionirji kot začetniki in nosilci ideje, tisti, ki orjejo ledino.

Tak je tudi program Transkribus in uporaba tehnologije HTR. Uporaba Transkribusa raste iz leta v leto. Tako je od leta 2015, ko je bilo uporabnikov okoli 2000, število uporabnikov v letu 2017 zraslo na okoli 17.000 (Muehlberger, 2019). Ti so dnevno naložili okoli 3000 strani, ki se nato obdelajo in ustrezno transkribirajo. Namen je, da se z uporabo programa in deljenjem izgrajenih modelov oblikuje ustrezna baza, ki bi jo lahko v prihodnosti uporabil program, brez potrebe, da se gradi nov model, ker bo ta že na voljo. S tem se prihrani čas, količina arhivskega gradiva, ki je dostopnejša uporabnikom, pa se veča.

VIRI IN LITERATURA

Leifert, G., Strauß, T., Grüning, T. and Labahn, R. (2016). *CITlab ARGUS for historical handwritten documents*. Pridobljeno 21. 2. 2020 s spletne strani <http://arxiv.org/abs/1605.08412>.

Pančur, A. (2016). *Popisi prebivalstva Slovenije 1830-1931. Orodje za transkribiranje historičnih demografskih podatkov*. Jezikovne tehnologije in digitalna humanistika. Ljubljana. Pridobljeno 21. 2. 2020 s spletne strani http://www.sdit.si/wp/wp-content/uploads/2018/09/JTDH-2018_Pancur_Trajnost-digitalnih-izdaj-Uporaba-staticnih-spletnih-strani-na-portalu-Zgodovina-Slovenije-Sistory.pdf.

Pančur, A. (2018). *Trajnost digitalnih izdaj, Uporaba staticnih spletnih strani na portalu Zgodovina Slovenije – Sistory*. Jezikovne tehnologije in digitalna humanistika. Ljubljana. Pridobljeno 21. 2. 2020 s spletne strani http://www.sdit.si/wp/wp-content/uploads/2018/09/JTDH-2018_Pancur_Trajnost-digitalnih-izdaj-Uporaba-staticnih-spletnih-strani-na-portalu-Zgodovina-Slovenije-Sistory.pdf.

³ IIIF je mednarodni okvir za interoperabilnost slik, v okviru katerega je zastavljenih več aplikativnih programskih vmesnikov, ki s standardizacijo procesov zagotavljajo enotno opisovanje in dostavo slik preko spleta.

- READ – Recognition and Enrichment of Archival Documents.** Pridobljeno 14. 2. 2020 s spletne strani <https://cordis.europa.eu/project/id/674943>.
- Kodrič Budna, N. (2018).** *Korespondenca Jožefine in Fidelija Terpinč (1825–1858)*, Viri št.41, Arhivsko društvo Slovenije, Ljubljana.
- Muehlberger, G., Seaward et al. (2019).** *Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study*, Journal of Documentation, Vol. 75 No. 5, str. 955 in str. 962. Pridobljeno 14. 2. 2020 s spletne strani <https://doi.org/10.1108/JD-07-2018-0114>.
- Muehlberger, G. (2019).** *Bringing the past back to life – Transkribus*. Pridobljeno 14. 2. 2020 s spletne strani <https://ec.europa.eu/digital-single-market/en/blogposts/bringing-past-back-life-transkribus>.
- Optical character recognition. (2020).** Cambridge University Press 2020. Pridobljeno 12. 2. 2020 s spletne strani <https://dictionary.cambridge.org/dictionary/english/optical-character-recognition>.
- Schantz, H. F. (1982).** *The history of OCR: optical character recognition*. Recognition Technologies Users Association.
- The First OCR System: "GISMO" (2004-2020).** Pridobljeno 12. 2. 2020 s spletne strani <http://www.historyofinformation.com/detail.php?entryid=885b>.
- Transkribus Cloud Services.** Pridobljeno 12. 2. 2020 s spletne strani https://transkribus.eu/wiki/index.php/Transkribus_Cloud_Services.
- Vogeler G., Burghartz S. (2016).** *Mehrwert für die Forschung: Digitales Edieren am Beispiel der Basler Jahrrechnungen im 16. Jahrhundert*, Pridobljeno 12. 2. 2020 s spletne strani <https://www.historikertag.de/Hamburg2016/events/was-sind-digital-humanities-praxisbeispiele-aus-den-geschichtswissenschaften-what-is-digital-humanities-examples-of-practice-in-historical-sciences-copy/index.html>.

SUMMARY

USE CASE OF THE TRANSKRIBUS SOFTWARE AND THE CREATION OF A MODEL FOR AUTOMATIC OPTICAL CHARACTER RECOGNITION FOR A SIMPLIFIED TRANSCRIPTION OF WRITTEN GOTHIC SCRIPT

Jože Glavič

Historical Archives Ljubljana, Novo mesto, Slovenia
joze.glavic1@guest.arnes.si

Transkribus is a software tool, developed at the University of Innsbruck, as part of the European READ project. The project aims to make archival materials more accessible through the development of necessary tools and technology.

The purpose of the tool is the automated recognition of handwritten documents. Its usefulness is shown especially when there is a need to transcribe a large collection of handwritten documents, which are usually stored in archives, but can be also found in libraries, museums and other related institutions. The tool itself offers automated processing of documents, especially the recognition of handwritten documents, layout analysis and optical character recognition (OCR).

The tool was used on letters which are a part of the Correspondence of Josephine and Fidelia Terpinč (1825-1858), published in the publication Viri 41, 2018, Ljubljana. Software was used to combine digitized documents and their handwritten transcripts. Transkribus was then used for training a so-called HTR model, which can be used to transcribe the rest of the Terpinč Family's correspondence automatically and fast.

The author will present the way to use the software, the process of creating an HTR model on the page of Josephine and Fidelia Terpinč Correspondence, all for purpose of simplifying and shortening the process when transcribing documents handwritten in gothic font.