

NARODNA IN UNIVERZITETNA KNJIŽNICA

Janez Žumer

**OPTIČNA PREPOZNAVA ZNAKOV V PROCESIH DIGITALIZACIJE
KNJIŽNIČNEGA GRADIVA: UPORABNOST PLATFORME
TRANSKRIBUS ZA NADALJNI RAZVOJ DIGITALNE KNJIŽNICE
SLOVENIJE**

Pisna naloga za bibliotekarski izpit

Ljubljana, 2025

Ključna dokumentacijska informacija

Ime in PRIIMEK: Janez ŽUMER

Naslov pisne naloge: Optična prepoznavna znakov v procesih digitalizacije knjižničnega gradiva: uporabnost platforme Transkribus za nadaljnji razvoj Digitalne knjižnice Slovenije

Kraj: Ljubljana

Leto: 2025

Št. strani: 34

Št. slik: 10

Št. preglednic: 1

Št. prilog: /

Št. strani prilog: /

Št. referenc: 17

Strokovno usposabljanje za bibliotekarski izpit je potekalo v: Narodni in univerzitetni knjižnici
Mentor v času strokovnega usposabljanja: Janko Klasinc, Narodna in univerzitetna knjižnica

UDK 02:005.522.52:004.352.242

Ključne besede: knjižnične storitve, digitalno okolje, digitalizacija, digitalna tehnologija, digitalne knjižnice, digitalna obdelava

Izveček: Optična prepoznavna znakov kot ena izmed storitev, ki jih nudi Digitalna knjižnica Slovenije, je ključnega pomena za zagotavljanje dostopa do informacij, saj digitalno preslikano besedilno gradivo iz berljivih slik preobraža v iskalno besedilo in s tem uporabniku omogoča hitrejša in zanesljivejša iskanje informacij. A medtem ko so rezultati optične prepoznavne znakov v tiskanih besedilih, ki jo zaposleni v Oddelku za digitalne vsebine Narodne in univerzitetne knjižnice znotraj programskega okolja Digitization Manager izvajajo s programskim orodjem ABBYY FineReader Engine 11, dobri in zanesljivi, temu v primeru tipkopisov zaradi slabše razločnosti besedila dostikrat ni tako. Iz tega razloga smo v raziskavi preučili drugo rešitev, s pomočjo katere bi bilo mogoče izboljšati uporabniško izkušnjo portala dLib.si. Ko gre za tipkopise, se je kot ena od možnih poti do dviga kakovosti optične prepoznavne znakov v tipkopisih kazala uporaba platforme Transkribus, ki deluje na temelju strojnega učenja branja besedila. Da bi preverili, ali je temu res tako, smo z obema orodjema izvedli optično prepoznavo znakov v petih izbranih tipkopisih, ki so objavljeni na portalu dLib.si, pri čemer smo na platformi Transkribus pred zagonom postopka optične prepoznavne za vsako od petih besedil pripravili lasten jezikovni model. Po pridobitvi rezultatov, za katero smo s pomočjo odprtokodnega orodja Dinglehopper izračunali odstotke napak na ravni posameznih znakov v vseh desetih izvedbah optične prepoznavne, smo preučili še možnosti za vpeljavo platforme Transkribus v potek dela na področju digitalizacije knjižničnega gradiva v Narodni in univerzitetni knjižnici. Kot se je izkazalo, bi vpeljavo platforme Transkribus tako z vidika izboljšanja optične prepoznavne znakov v tipkopisih kot z vidika možnosti za njeno sistemsko integracijo v programsko okolje Digitization Manager lahko pomenila korak v razvoju Digitalne knjižnice Slovenije.

KAZALO VSEBINE

1 UVOD.....	1
1.1 OPREDELITEV PROBLEMA.....	1
1.2 NAMEN IN CILJ NALOGE	2
1.3 METODOLOGIJA DELA.....	2
2 OPTIČNA PREPOZNAVA ZNAKOV IN PREPOZNAVA ROKOPISNIH BESEDIL	4
2.1 OPTIČNA PREPOZNAVA ZNAKOV	4
2.2 PREPOZNAVA ROKOPISNIH BESEDIL	5
2.3 PROGRAMSKO ORODJE ABBYY FINEREADER ENGINE 11.....	7
2.4 PLATFORMA TRANSKRIBUS	9
3 Raziskava.....	11
3.1 PRIPRAVA BESEDILNIH DATOTEK Z UPORABO PROGRAMSKEGA ORODJA ABBYY FINEREADER ENGINE 11	12
3.2 PRIPRAVA Besedilnih DATOTEK Z UPORABO PLATFORME TRANSKRIBUS.....	12
3.3 IZRAČUN NAPAK NA RAVNI POSAMEZNIH ZNAKOV V Besedilnih DATOTEKAH, PRIDOBLENJIH Z UPORABO PROGRAMA ABBYY FINEREADER IN PLATFORME TRANSKRIBUS.....	15
3.4 REZULTATI IZRAČUNOV NAPAK NA RAVNI POSAMEZNIH ZNAKOV V PRIDOBLENJIH BESEDILNIH DATOTEKAH	15
3.4.1 Dejavnika razločnosti besedil in prisotnosti rokopisnih popravkov	15
3.4.2 Predstavitev rezultatov	18
3.5 MOŽNOSTI INTEGRACIJE PLATFORME TRANSKRIBUS V PROGRAMSKO OKOLJE DIGITIZATION MANAGER	20
4 ZAKLJUČEK	24
5 NAVEDENI VIRI IN LITERATURA	26

KAZALO SLIK

Slika 1: Prikaz rabe platforme Transkribus in sodelovanja uporabnikov pri njenem razvoju (Muehlberger idr., 2019)	10
Slika 2: Prepoznavna regije in vrstic besedila, ki jo je omogočila analiza postavitve besedila, izvedena z uporabo modela Universal Lines	13
Slika 3: Rezultat analize postavitve besedila in njegov temeljni prepis	14
Slika 4: Izsek doktorske disertacije z naslovom Angleška poetično-politična drama v letih 1930-1940 (COBISS.SI-ID - 9680384)	16
Slika 5: Izsek doktorske disertacije z naslovom Literarna kritika in estetska ideologija Paula Bourgeta (COBISS.SI-ID - 19745378)	16
Slika 6: Izsek doktorske disertacije z naslovom Kiparstvo na slovenskem Štajerskem v 18. stoletju (COBISS.SI-ID - 196615680)	16
Slika 7: Izsek doktorske disertacije z naslovom Stensko slikarstvo od 15. do srede 17. stoletja na Slovenskem (COBISS.SI-ID - 196893440)	16
Slika 8: Izsek doktorske disertacije z naslovom Gospodarska geografija pokrajine med Snežnikom z Javorniki, Vremščico in Slavnikom (COBISS.SI-ID - 249056000)	17
Slika 9: Rezultat optične prepoznavne znakov na platformi Transkribus po predhodni prilagoditvi vrstic in izpeljavi postopka treniranja lastnega jezikovnega modela	18
Slika 10: Rezultat optične prepoznavne znakov v istem delu besedila s programskim orodjem ABBYY FineReader Engine 11	18

KAZALO PREGLEDNIC

Preglednica 1: Rezultati primerjave optično prepoznanih besedil z njihovimi temeljnimi prepisi, izvedene z odprtokodnim orodjem Dinglehopper	18
--	----

ZAHVALA

Za pomoč pri pripravi naloge ter koristne napotke in informacije se zahvaljujem mentorju Janku Klasincu in zaposlenim v Službi za informacijsko podporo Narodne in univerzitetne knjižnice.

1 UVOD

Optična prepoznavna znakov je ključnega pomena za zagotavljanje dostopa do informacij, saj digitalno preslikano besedilno gradivo iz berljivih slik preobraža v iskalno besedilo in s tem uporabniku omogoča hitrejša in zanesljivejša iskanja informacij. Digitalno preslikana besedila, objavljena na portalu Digitalne knjižnice Slovenije (dLib.si), bi v grobem lahko razdelili v dve skupini. Gre bodisi za tiskana besedila, pri katerih je izvedba optične prepoznavne znakov razmeroma enostavna in natančna, bodisi za rokopisna besedila, ki vključujejo tudi tipkopise in pri katerih pretvorba digitalnega posnetka v besedilno obliko predstavlja večji izziv. Izmed številnih orodij, ki so dostopna na trgu, je v Oddelku za digitalne vsebine Narodne in univerzitetne knjižnice (dalje NUK) za optično prepoznavo znakov v tiskanih besedilih in tipkopisih uporabljeno programsko orodje ABBYY FineReader Engine 11. A medtem ko omenjeni program, kot že rečeno, v primeru tiskanih besedil nudi razmeroma dobre rezultate, je kakovost optične prepoznavne znakov v tipkopisih zaradi slabše razločnosti besedil pogosto okrnjena.

1.1 OPREDELITEV PROBLEMA

Kot izhaja iz Manifesta za digitalne knjižnice, ki ga je izdala IFLA in sprejela Generalna skupščina UNESCO (IFLA/UNESCO Manifesto for digital libraries, 2011), je poslanstvo digitalne knjižnice v zagotavljanju neposrednega dostopa do digitalnih in ne-digitalnih informacijskih virov v strukturirani obliki. Kakovost tega dostopa v primeru ne-digitalnih oziroma digitaliziranih gradiv ni odvisna le od kakovosti digitalne preslikave ter ustrezne strukturiranosti in bogastva metapodatkov, ki digitalni objekt opisujejo, temveč – v primeru besedilnih gradiv – tudi od kakovosti optične prepoznavne znakov. Rezultati optične prepoznavne znakov v tipkopisih, ki jo v Oddelku za digitalne vsebine NUK opravljamo s programskim orodjem ABBYY FineReader Engine 11, so relativno nizke kakovosti. Iz tega razloga bi bilo smiselno preučiti druge, ustrežnejše rešitve, s pomočjo katerih bi bilo mogoče izboljšati uporabniško izkušnjo portala dLib.si. Ko gre za tipkopise, se kot ena od možnih poti do dviga kakovosti optične prepoznavne znakov kaže uporaba platforme Transkribus, ki je bila v okviru projekta TranScriptorium razvita leta 2015 in ki za strojno učenje branja rokopisnih besedil uporablja sistem umetnih nevronske mreže (Muehlberger idr., 2019).

1.2 NAMEN IN CILJ NALOGE

Namen pisne naloge je opisati značilnosti procesov optične prepoznavne znakov (ang. optical character recognition (OCR)) in prepoznavne rokopisnih besedil (ang. handwritten text recognition (HTR)), pobliže predstaviti razvoj, ustroj in uporabniške možnosti platforme Transkribus in programskega orodja ABBYY FineReader Engine 11 ter oceniti uporabnost platforme Transkribus za nadaljnji razvoj Digitalne knjižnice Slovenije in za posledično izboljšanje uporabniške izkušnje portala dLib.si. To oceno bomo podali tako na podlagi empirične primerjave kakovosti optične prepoznavne znakov v tipkopisih, kakršno nudita platforma Transkribus in programsko orodje ABBYY FineReader Engine 11, kot tudi na podlagi ovrednotenja možnosti za vpeljave platforme Transkribus v potek dela na področju digitalizacije knjižničnega gradiva v NUK.

Cilj pisne naloge bo zatorej potrditi oziroma ovreči naslednji hipotezi:

- H1: Platforma Transkribus v primeru optične prepoznavne znakov v tipkopisih nudi boljše rezultate kot programsko orodje ABBYY FineReader Engine 11. S tem omogoča lažje iskanje informacij in odpira možnost za izboljšanje uporabniške izkušnje portala dLib.si.
- H2: Funkcionalnosti platforme Transkribus ni mogoče integrirati v programsko okolje Digitization Manager. Raba platforme Transkribus kot zunanega orodja za izvedbo optične prepoznavne znakov v tipkopisih bi zato zahtevala prilagoditev poteka dela na področju digitalizacije knjižničnega gradiva in trajnega ohranjanja digitalnih objektov.

1.3 METODOLOGIJA DELA

V pisni nalogi bomo raziskovali uporabnost platforme Transkribus za nadaljnji razvoj Digitalne knjižnice Slovenije in za izboljšanje uporabniške izkušnje portala dLib.si., in sicer tako z vidika kakovosti optične prepoznavne znakov v tipkopisih kot z vidika možnosti za vpeljavo platforme Transkribus v potek dela na področju digitalizacije knjižničnega gradiva v NUK.

Pisna naloga bo zajemala teoretični in empirični del. V teoretičnem delu bomo najprej opredelili pojma optične prepoznavne znakov in prepoznavne rokopisnih besedil. V drugem delu teoretičnega

dela bomo predstavili razvoj in ustroj platforme Transkribus in programskega orodja ABBYY FineReader Engine 11, ki ga za izvedbo optične prepoznavne znakov v tipkopisih v NUK že uporabljamo, ter na ta način bralca uvedli v empirični del naloge.

V empiričnem delu naloge bomo predstavili potek in rezultate raziskave, s katero smo želeli preučiti učinkovitost programskega orodja ABBYY FineReader Engine 11 in platforme Transkribus v postopkih optične prepoznavne znakov v tipkopisih. Optično prepoznavo znakov smo izvedli v petih izbranih tipkopisih, ki so objavljeni na portalu dLib.si. Kot smo zapisali zgoraj, platforma Transkribus deluje na temelju strojnega učenja branja besedil, kar pomeni, da bomo za njeno uporabo predhodno pretipkali približno 5.000 besed iz vsakega od petih besedil, ki jih bomo uporabili. Ti temeljni prepisi besedil bodo nato po eni strani služili za izvedbo optične prepoznavne znakov s pomočjo platforme Transkribus, po drugi pa tudi za primerjavo rezultatov obeh programov oziroma za oceno števila napak, ki se bodo pojavile v obeh procesih optične prepoznavne. Število napak bo ocenjeno na ravni posameznih znakov (ang. character error rate (CER)) in bo zapisano v odstotkih. Empiričnemu delu bo sledilo še krajše ovrednotenje možnosti za vpeljavo platforme Transkribus v potek dela na področju digitalizacije knjižničnega gradiva v NUK. Gre namreč za dejavnik, ki ga ne gre spregledati, ko govorimo o nadaljnjem razvoju Digitalne knjižnice Slovenije. Ovrednotenje bo izvedeno na podlagi preučitve možnosti za integracijo aplikacijskega programskega vmesnika (ang. application programming interface (API))¹ platforme Transkribus v programsko okolje Digitization Manager, ki se za digitalizacijo knjižničnega gradiva uporablja v Oddelku za digitalne vsebine NUK.

¹ Aplikacijski programski vmesnik pomeni povezavo med različnimi programskimi orodji, ki omogoča, da ta med seboj komunicirajo ter delijo podatke in funkcije (Goodwin, 2024).

2 OPTIČNA PREPOZNAVA ZNAKOV IN PREPOZNAVA ROKOPISNIH BESEDIL²

Vključitev optične prepoznavne znakov, s tem pa tudi prepoznavne rokopisnih besedil v procese digitalizacije knjižničnega gradiva je ključnega pomena, saj uporabniku s tem, ko digitalno preslikano besedilo iz berljivih slik preobrazimo v iskalno besedilo, omogočimo hitrejše in zanesljivejše iskanje informacij. Razvoj optične prepoznavne znakov in prepoznavne rokopisnih besedil sega v prvo polovico 20. stoletja, za njuno izvajanje pa je bilo do danes razvitih veliko orodij (Memon idr., 2020). A medtem ko so bila tekom razvoja teh orodij tiskana besedila zaradi lažje strojne berljivosti obravnavana bistveno pogostejše kot rokopisna besedila, je, kot še zapišejo Memon idr. (2020), v zadnjih petnajstih letih velik zagon dobilo raziskovanje novih pristopov k strojnemu učenju.³ Ti pristopi – združeni s tehnologijami za procesiranje slik – so omogočili razvoj in nadgrajevanje naprednih programskih orodij za izvajanje prepoznavne rokopisnih besedil.

2.1 OPTIČNA PREPOZNAVA ZNAKOV

Pojem optične prepoznavne znakov se danes nanaša na proces, tekom katerega so digitalno preslikana tiskana besedila konvertirana v strojno kodirano besedilo. Ta proces poteka na osnovi sprotne primerjave posameznih znakov znotraj izhodiščne slike z množico tipografij (Muehlberger idr., 2019), izmed katerih program, ki ga uporabljamo za izvedbo optične prepoznavne znakov, izbere najustreznejšo.

² Na tem mestu naj opozorimo, da je v literaturi izraz »optična prepoznavna znakov« po eni strani rabljen kot nadpomenka za prepoznavo tako tiskanih kot rokopisnih besedil, med katera, kot smo že omenili, uvrščamo tudi tipkopise, po drugi in pogostejše pa kot sopomenka za optično prepoznavo tiskanih besedil. Čeprav je raba obeh naslovnih izrazov v prvem pomenu, torej kot nadpomenke in podpomenke, točnejša, bomo v tem poglavju izraza »optična prepoznavna znakov« in »prepoznavna rokopisnih besedil« uporabljali tako, kot ju najpogostejše uporabljajo raziskovalci obravnavanega področja. Za prepoznavo tipkopisov, s katero se bomo ukvarjali v empiričnem delu, bomo v nalogi vseskozi uporabljali izraz »optična prepoznavna znakov v tipkopisih«.

³ Pojem strojnega učenja se nanaša na algoritme, ki omogočajo računalnikom, da posnemajo človeško učenje, samostojno opravljajo naloge in na podlagi novih podatkov samodejno izboljšujejo svojo učinkovitost in natančnost (IBM, 2021).

Tekom prvih desetletij razvoja je tehnologija za optično prepoznavo znakov napredovala do te mere, da je postala tržno dostopna, in sicer v obliki bralnika IBM 1287, ki je bil med drugim zmožen brati rokopisne številke in je bil prvič predstavljen na tehnološkem sejmu v New Yorku leta 1965. V sedemdesetih letih 20. stoletja so se razvijalci osredotočili na izboljšanje zmogljivosti in odzivnosti tovrstnih bralnikov, med letoma 1980 in 2000 pa so razvili in vseskozi izboljševali programska orodja za izvedbo optične prepoznavne znakov. Ta orodja – med njimi tudi program ABBYY FineReader, ki je v svoji prvi različici na trg prišel leta 1993 (Nockels idr., 2022) – so zelo kmalu prešla v rabo v izobraževalnih ustanovah in postala še posebej zanimiva po vpeljavi pred-procesnih binarizacijskih tehnik,⁴ ki so znatno prispevale h kakovosti optične prepoznavne znakov in raziskovalcem bistveno olajšale vpogled v tiskane zgodovinske dokumente (Memon idr., 2020).

2.2 PREPOZNAVA ROKOPISNIH BESEDIL

Prepoznavna rokopisnih besedil se, kot že ime pove, od zgoraj opredeljene optične prepoznavne znakov razlikuje v tem, da se nanaša na besedila, ki niso tiskana in so zato bistveno močnejše zaznamovana z zgodovinskimi, kulturnimi in drugimi pogojenostmi njihovih avtorjev. Podobno kot optična prepoznavna znakov sicer tudi programska prepoznavna rokopisnih besedil deluje na osnovi sprotne primerjave posameznih znakov znotraj izhodiščnega skenograma s predlogo. Tehnologiji pa se razlikujeta v tem, da programi za izvedbo optične prepoznavne znakov zajemajo iz baze tipografij, medtem ko programi za prepoznavo rokopisnih besedil, kot je platforma Transkribus, ki jo bomo podrobneje predstavili v nadaljevanju, črpajo iz t. i. velikih jezikovnih modelov (ang. large language model (LLM)), ki niso statični, temveč se vseskozi širijo in postajajo vse kompleksnejši.⁵

⁴ Binarizacija slike je postopek pretvorbe digitalnega barvnega ali sivinskega posnetka v binarnega. Binarna slika ima samo dve možni vrednosti za vsako slikovno piko, ki je po izvedbi postopka bodisi črna bodisi bela (Binarization, 2025).

⁵ Kot zapišejo Nockels idr. (2024), je bil platformi Transkribus v aprilu leta 2023 dodan nov jezikovni model z imenom *German Giant*, ki je bil treniran na besedilih, pisanih med 16. in 21. stoletjem v latiničnih pisavah in v nemški kurzivi (nem. Kurrentschrift). Kot dodajajo, tovrstni splošni algoritmi povečujejo podatkovni bazen, iz katerega lahko zajemajo raziskovalci, ki trenirajo lastne jezikovne modele na besedilnih korpusih,

Kljub temu, da se obravnavani tehnologiji danes v marsičem razlikujeta, je bil razvoj prepoznavne rokopisnih besedil dlje časa tesno povezan z razvojem optične prepoznavne znakov. V tem prvem obdobju razvoja so bile tehnike za prepoznavo rokopisnih besedil v uporabi predvsem v finančnem in komercialnem sektorju, in sicer za namene prepoznave poštnih naslovov, procesiranja bančnih čekov in verifikacije podpisov. Kasneje sta se z razvojem naprednejših tehnologij za prepoznavanje vzorcev in prvimi koraki na področju umetne inteligence v devetdesetih letih 20. stoletja, razvojem globokega učenja⁶ v tem tisočletju ter nenazadnje s povečevanjem procesne zmogljivosti računalnikov in vsesplošnim napredkom na področju digitalnega preslikovanja uporabnost in domet prepoznavne rokopisnih besedil izjemno povečala, ukvarjanje s strojno prepoznavo rokopisnih besedil pa se je znotraj računalništva in informatike izoblikovalo v samostojno raziskovalno področje (Muehlberger idr., 2019). Izziv, ki je bil gonilo tega razvoja, je vseskozi predstavljala prepoznavna rokopisnih besedil, ki bi bila v vseh primerih enako kakovostna, natančna in hitra ter bi omogočila iskanje po celotnem besedilu (Nockels idr., 2024). Kot še dodajajo Nockels idr. (2024), je ta cilj danes povečini dosežen, saj orodja za prepoznavo rokopisnih besedil že ustvarjajo zanesljiva strojno berljiva besedila in omogočajo celovito transkripcijo obsežnih zbirk zgodovinskih dokumentov.⁷

Čeprav je tehnologija za prepoznavo rokopisnih besedil z uporabo umetne inteligence in strojnega učenja prišla izjemno daleč, pa njen potencial ostaja skorajda neizčrpen. Če namreč k rokopisnim zbirkam prištejemo še zbirke inkunabul oziroma prvotiskov, ki so, kot pravijo Muehlberger idr. (2019), pogosto preveč kompleksne, da bi bile lahko učinkovito transkribirane s pomočjo optične prepoznavne znakov, ter novost, ki jo prinaša tehnologija prepoznavne rokopisnih besedil,

s katerimi se ukvarjajo. Vendar pa proces ni enosmeren. Z uporabo manjših, bolj specifičnih jezikovnih modelov namreč tudi splošni algoritmi postajajo natančnejši in zmožni pokrivati širša časovna razdobja.

⁶ Pojem globokega učenja se nanaša na metodo umetne inteligence, ki za učenje iz podatkov uporablja umetne nevronske mreže. Te so zasnovane po zgledu človeških možganov in jih je mogoče uporabiti za reševanje najrazličnejših problemov, vključno s prepoznavanjem slik, obdelavo naravnega jezika in prepoznavanjem govora (Holdsworth in Scapicchio, 2024).

⁷ Kot zapišejo Muehlberger idr. (2019), avtomatizirana prepoznavna rokopisnih besedil, ki ima za osnovo dovolj obširen temeljni prepis besedila (ang. ground truth), na podlagi katerega poteka globoko strojno učenje, znotraj računalništva in informatike velja za rešen znanstveni problem.

primerjamo s preteklimi metodami raziskovanja besedilnih korpusov, lahko trdimo ne le, da ta tehnologija izboljšuje in razširja dostop do zbirk ter uporabnikom omogoča hitro in učinkovito najdevanje tem, besed, oseb, krajev in dogodkov, temveč tudi, da lahko s pomnožitvijo raziskovalnih možnosti⁸ spremeni razumevanje konteksta, v katerem se pojavljajo (Muehlberger idr., 2019).

2.3 PROGRAMSKO ORODJE ABBYY FINEREADER ENGINE 11

Kot smo že omenili, je podjetje ABBYY prvo različico programskega orodja FineReader izdalo leta 1993. Tehnologija optične prepoznavne znakov je s tem prvič v zgodovini postala dostopna širši javnosti (Nockels idr., 2022). Od tedaj do danes so v podjetju ABBYY razvili številne različice programa. A medtem ko zadnja verzija, ki nosi ime ABBYY FineReader PDF 16 in je bila izdana leta 2022, velja za eno najbolj naprednih na trgu, je bil program v prvi polovici svojega obstoja v marsičem omejen. Nockels idr. (2022) v zgoraj povzetem članku zapišejo še, da je različica 9.0, ki je bila izdana leta 2008, uporabniku predstavljala omejitve pri analizi postavitve besedila (ang. document layout analysis), medtem ko so imeli uporabniki predhodnih verzij težave z omejeno prilagodljivostjo programa njihovim potrebam, z omejitvami pri številu strani, ki so jih lahko obdelali, in s tem povezanimi stroški ter s slabšo učinkovitostjo programa pri besedilih s kompleksnejšimi in spreminjajočimi se tipografijami.

V Oddelku za digitalne vsebine NUK sta danes v uporabi dve različici programskega orodja FineReader, in sicer ABBYY FineReader Engine 11 in ABBYY FineReader Server 14. Obe verziji sta integrirani v interno razvito programsko okolje Digitization Manager, ki zaposlenim po digitalni preslikavi bodisi besedilnega bodisi slikovnega gradiva poleg izvedbe optične prepoznavne znakov omogoča obdelavo skenogramov, zajem metapodatkov iz vzajemnega kataloga slovenskih knjižnic (COBIB) ter pripravo paketov datotek, namenjenih objavi na portalu dLib.si in trajnemu ohranjanju digitalnih objektov. Na tem mestu naj omenimo še, da v procesih digitalizacije besedilnih gradiv v NUK programsko orodje ABBYY FineReader Server 14 služi optični prepoznavi tiskanih besedil

⁸ Avtorji povzetega članka k novim raziskovalnim možnostim štejejo semantične, lingvistične in geo-prostorske računalniške analize primarnih zgodovinskih virov (Muehlberger idr., 2019).

v gotici, medtem ko je za tiskana latinična besedila in tipkopise v uporabi programsko orodje ABBYY FineReader Engine 11.

Glede na to, da se bomo v nadaljevanju besedila ukvarjali z empirično primerjavo kakovosti optične prepoznavne znakov v tipkopisih, kakršno nudita platforma Transkribus in programsko orodje ABBYY FineReader Engine 11, je prav, da to orodje predstavimo nekoliko natančneje.

V primeru te različice programa gre za t. i. komplet za razvoj programske opreme (ang. software development kit (SDK)), ki omogoča integracijo tehnologij večjezične prepoznavne znakov in konvertiranja dokumentov v zunanje aplikacije. Sama prepoznavna, ki poleg optične prepoznavne znakov v strojno tiskanih besedilih (OCR) vključuje še prepoznavo znakov v ročno tiskanih besedilih (ang. intelligent character recognition (ICR)), prepoznavo kljukic (ang. optical mark recognition (OMR)) ter enodimenzionalnih in dvodimenzionalnih črtnih kod, poteka v šestih korakih. Ti poleg prvega vnosnega koraka vključujejo še:

- Pred-procesiranje slik: v tem koraku programsko orodje slike očisti šumov, izravna linije v primerih, da je bilo gradivo digitalizirano s pomočjo fotoaparata, izravna slike v primeru nagiba, ki ni večji od 20 stopinj, ter slike samodejno obreže in jih razdeli po sredini v primeru, da prikazujejo dve strani dokumenta.
- Analizo dokumenta in postavitve: ta korak, ki za izvedbo postopkov uporablja umetno inteligenco, služi določitvi polj, na katerih je potrebno izvesti prepoznavo.
- Prepoznavo: programsko orodje podpira 202 jezika za izvedbo optične prepoznavne znakov v strojno tiskanih besedilih in 126 jezikov za izvedbo prepoznavne znakov v ročno tiskanih besedilih.
- Verifikacijo in uporabniški poseg: v tem koraku ima uporabnik možnost preveriti rezultate analize dokumenta in prepoznavne ter vnesti potrebne popravke.
- Izvoz rezultatov optične prepoznavne: za izvedbo tega koraka je v programskem orodju na voljo veliko formatov, in sicer TXT, CSV, HTML, RTF, ODT, DOCX, XLS(X), PPTX, EPUB, FictionBook, ABBYY XML in ALTO XML (ABBYY, b. d.).

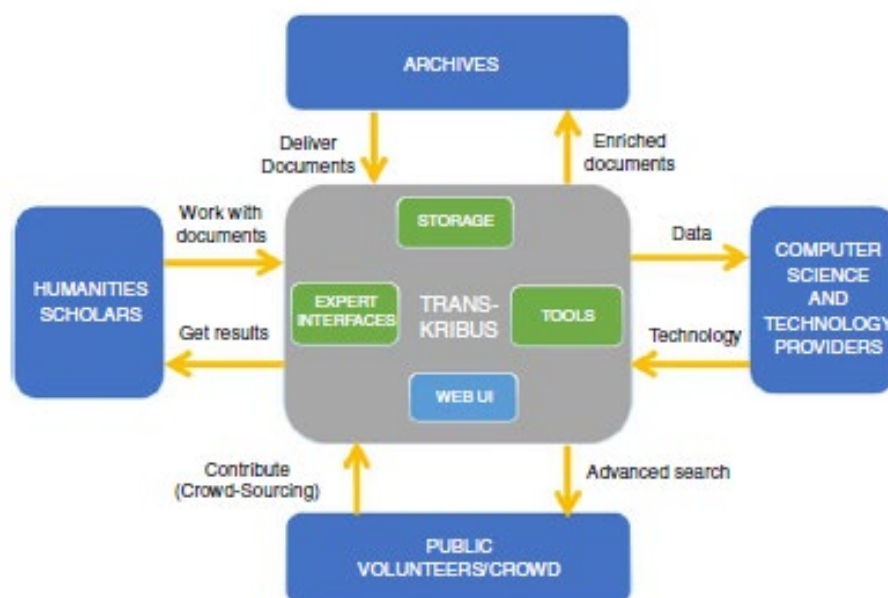
2.4 PLATFORMA TRANSKRIBUS

Posredno smo o platformi Transkribus nekaj že povedali v podpoglavju 2.2, a jo bomo v nadaljevanju predstavili še nekoliko natančneje.

Kot smo že omenili v uvodu, začetki platforme Transkribus segajo v leto 2015. Platforma je bila tedaj v okviru projekta TranScriptorium (2013–2015), financiranega s strani Evropske komisije, dokončno domišljena in zagnana. A medtem ko so se razvijalci v okviru omenjenega projekta posvečali predvsem računalniškim raziskavam in možnostim uresničitve zadanih ciljev, sta bila implementacija in razvoj platforme postavljena v središče naslednjega projekta, imenovanega Recognition and Enrichment of Archival Documents (READ, 2016–2019). Tako kot predhodnega, je tudi ta projekt financirala Evropska komisija, in sicer v okviru programa Evropske unije za raziskave in inovacije Obzorje 2020 (Muehlberger idr., 2019). Kot zapišejo Colutto idr. (2019), je bila v letu 2019 nato ustanovljena Evropska zadruga READ-COOP (European Cooperative Society (READ-COOP SCE)), ki platformo Transkribus upravlja še danes in ki tako (javnim) institucijam kot fizičnim osebam omogoča, da se v razvoj storitve vključijo ne le kot njeni uporabniki,⁹ temveč tudi kot soodločevalci. Kot izhaja iz poročila zadruge READ-COOP z dne 12. 5. 2023, ki ga navajajo Nockels idr. (2024), se je v zadrugo do tistega trenutka vključilo 85 institucij in 50 posameznikov iz 30 držav, medtem ko je danes, glede na podatek, ki ga najdemo na spletni strani zadruge READ-COOP, članov že več kot 200 (READ-COOP SCE, 2023).

Kot smo nakazali v zgornjem odstavku in v opombi 5, vtem ko govorimo o razvoju platforme Transkribus, ne smemo spregledati njenih uporabnikov. V tem kontekstu Muehlberger idr. (2019) zapišejo, da gre v primeru platforme Transkribus za orodje, ki daje tehnologijo prepoznave rokopisnih besedil na voljo štirim interesnim skupinam: arhivom oziroma knjižnicam, raziskovalcem, pripadnikom splošne javnosti ter računalničarjem oziroma razvijalcem tehnologije. Kot dodajajo in pokažejo z diagramom, prikazanim na Sliki 1, pa te skupine od uporabe platforme Transkribus nimajo le koristi, temveč na svoj način prispevajo tudi k njeni rasti in izboljšanju.

⁹ Glej opombo 5.



Slika 1: Prikaz rabe platforme Transkribus in sodelovanja uporabnikov pri njenem razvoju (Muehlberger idr., 2019)

Uporaba platforme Transkribus omenjenim skupinam prinaša naslednje koristi (Muehlberger idr., 2019):

- arhivi in knjižnice lahko svojim uporabnikom zagotovijo dostop do digitalnih zbirk rokopisnih besedil, ki so iskalna;
- raziskovalci lahko po izvedbi prepoznavne rokopisnih besedil gradivo, s katerim se ukvarjajo, raziskujejo učinkoviteje;
- pripadniki splošne javnosti lahko učinkoviteje raziskujejo družinsko zgodovino;
- računalničarji oziroma razvijalci tehnologije lahko zaprosijo za ponovno uporabo podatkov v obliki berljivih slik in njihovih transkripcij ter te podatke uporabijo za raziskovanje tehnologije prepoznavne rokopisnih besedil.

Iz diagrama na Sliki 1 je poleg tega mogoče razbrati, da uporabniki platforme Transkribus k njenemu razvoju tudi prispevajo, in sicer tako neposredno kot posredno. Kajti medtem ko lahko računalničarji in razvijalci tehnologije, ki delujejo v zadrugi READ-COOP, ugotovitve iz svojih raziskav aplicirajo neposredno na razvoj platforme, gre pri aktivnosti drugih skupin uporabnikov za posreden prispevek k napredku tehnologije prepoznavne rokopisnih besedil. Kot izhaja iz zgoraj navedenega poročila, je na platformi Transkribus meseca maja leta 2023 sodelovalo 127.332

uporabnikov, ki so dnevno procesirali 12.946 strani rokopisnih besedil. K temu velja dodati, da je bilo od začetkov platforme Transkribus v letu 2015 do 12. 5. 2023 na strežnike naloženih 46,9 milijonov digitaliziranih strani zgodovinskih dokumentov in da je bilo do tistega trenutka v uporabi 21.032 jezikovnih modelov, ki so bili trenirani s pomočjo konvolucijskih nevronske mreže¹⁰ (Nockels idr., 2024). Rastoča mreža uporabnikov je ključnega pomena za uspeh platforme Transkribus. Povratne informacije uporabnikov in poročanje o napakah (hroščih) so namreč neposredno vpeti v razvojno delo na platformi. Poleg tega pa je tu še dejstvo, da gre pri platformi Transkribus v osnovi za sistem strojnega učenja, ki z vsakim procesiranim skenogramom postane močnejši. Dokumenti, naloženi na strežnike prek Transkribusovega grafičnega uporabniškega vmesnika (ang. graphic user interface (GUI)), ostajajo zasebni. Ne glede na to pa se v ozadju nevronske mreže vseskozi učijo ter z vsako digitalizirano stranjo in spremljevalnim temeljnim prepisom besedila,¹¹ ki ju posredujejo uporabniki, postajajo učinkovitejše pri prepoznavi najrazličnejših pisav in jezikov (Muehlberger idr., 2019).

3 RAZISKAVA

Kot smo omenili v uvodu, se bo empirični del pisne naloge nanašal na preizkus učinkovitosti programskega orodja ABBYY FineReader Engine 11, ki je kot integralni del programskega okolja Digitization Manager v rabi v Oddelku za digitalne vsebine NUK, ter platforme Transkribus, ki se, ko gre za tipkopise, kaže kot ena od možnih poti do dviga kakovosti optične prepoznavne znakov.

Za namen preizkusa obeh orodij oziroma potrditve prve hipoteze (H1) smo na portalu dLib.si poiskali pet doktorskih disertacij iz petdesetih in šestdesetih let preteklega stoletja z nekoliko

¹⁰ Konvolucijske nevronske mreže so specializirane vrste algoritmov za globoko učenje, zasnovane predvsem za naloge, ki zahtevajo prepoznavanje objektov in klasifikacijo slik (IBM, b. d.).

¹¹ Kot smo že omenili v opombi 5, imajo uporabniki platforme Transkribus poleg uporabe jezikovnih modelov, ki so v danem trenutku na voljo, možnost trenirati lastne jezikovne modele. Za ta namen morajo poleg skenogramov v sistem posredovati tudi temeljni delni prepis besedila, ki ga želijo transkribirati in na podlagi katerega se nevronske mreže naučijo prepoznave rokopisa. Ob tem velja omeniti še, da so uporabniki, ki svoje jezikovne modele delijo s skupnostjo, za to nagrajeni s krediti, ki jih lahko uporabijo za nadaljnje delo na platformi Transkribus (Colutto idr., 2019). Na ta način je še dodatno poskrbljeno za krepitev platforme in strojnega učenja, ki poteka v ozadju.

različnimi tipografijami, manjšim številom ročnih popravkov in predvsem z različnimi stopnjami razločnosti besedila. Kot bomo videli, je bila prav razločnost besedila ključni dejavnik, od katerega je bila znatno odvisna kakovost optične prepoznavne znakov s programskim orodjem ABBYY FineReader Engine 11, medtem ko na kakovost optične prepoznavne, izvedene na platformi Transkribus, razločnost besedila ni pomembneje vplivala.

3.1 PRIPRAVA BESEDILNIH DATOTEK Z UPORABO PROGRAMSKEGA ORODJA ABBYY FINEREADER ENGINE 11

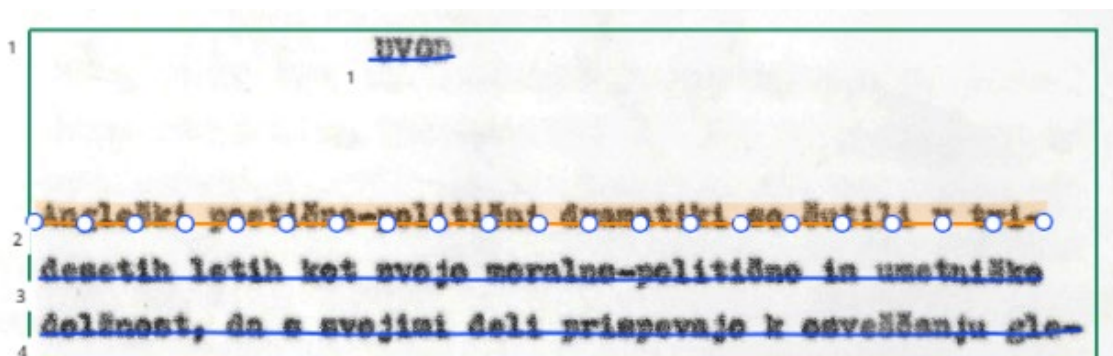
Programsko orodje ABBYY FineReader Engine 11 je, kot smo zapisali, integrirano v programsko okolje Digitization Manager, ki ga zaposleni v Oddelku za digitalne vsebine NUK uporabljajo ne le za izvedbo optične prepoznavne znakov, temveč tudi za zajem metapodatkov iz vzajemnega kataloga slovenskih knjižnic (COBIB), obdelavo skenogramov digitalno preslikanega besedilnega ali slikovnega gradiva in pripravo paketov datotek, ki so po eni strani namenjeni objavi na portalu dLib.si, po drugi pa trajnemu ohranjanju digitalnih objektov. Iz tega razloga smo pred pričetkom dela s programom ABBYY FineReader iz digitalnega arhiva NUK pridobili osemdeset izvornih skenogramov od vsakega od petih izbranih tipkopisov in v programskem okolju Digitization Manager ustvarili pet projektov z enoznačnimi poimenovanji. Na ta način smo ustvarili pet delovnih okolij, v katera je bilo mogoče pridobljene skenograme uvoziti in na njih z izbiro enega od enajstih postopkov, ki so na voljo v modulu za obdelavo gradiva, izvesti optično prepoznavo znakov. Rezultat teh postopkov je bila pridobitev petih datotek PDF in petih besedilnih datotek (dalje datotek TXT), ki smo jih uporabili tako za nadaljnje delo na platformi Transkribus kot za končno oceno števila napak na ravni posameznih znakov (ang. character error rate (CER)).

3.2 PRIPRAVA BESEDILNIH DATOTEK Z UPORABO PLATFORME TRANSKRIBUS

Za namen izvedbe optične prepoznavne znakov v petih izbranih tipkopisih na platformi Transkribus, ki deluje na temelju strojnega učenja branja besedil, je bilo v prvem koraku potrebno pripraviti dovolj obširen temeljni prepis besedila, na podlagi katerega bi strojno učenje potekalo. Pri tem smo upoštevali priporočilo razvijalcev platforme Transkribus, ki uporabnikom svetujejo, da za izvedbo strojnega učenja branja besedil oziroma treniranje lastnih jezikovnih modelov uporabijo prepise,

ki obsegajo najmanj 5000 besed (READ-COOP, 2023). Da bi to dosegli, smo za izhodišče vzeli datoteke TXT, ki smo jih predhodno pridobili z uporabo programskega orodja ABBYY FineReader Engine 11 in ob zgledovanju po digitalnih posnetkih izvornih strani, prenešenih v format PDF, pripravili temeljne prepise, ki niso vsebovali napak, do katerih je prišlo ob predhodni optični prepoznavi znakov. Kot se je izkazalo, je bilo pripravo temeljnih prepisov besedil na takšen način mogoče izvesti v štirih od petih primerov, saj je bila razločnost enega od digitalno preslikanih tipkopisov okrnjena do te mere, da je bil rezultat optične prepoznavne znakov s programskim orodjem ABBYY FineReader Engine 11 povsem neuporaben. V tem primeru smo v orodje Notepad 4.685 besed ročno pretipkali in pri tem ohranili izvorno postavitev in delitev besedila.

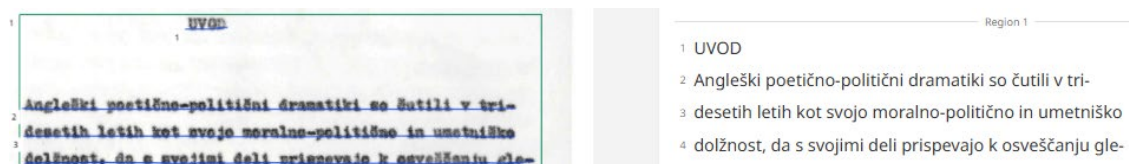
V naslednjem koraku smo na platformi Transkribus ustvarili zbirko, v katero smo v obliki ločenih projektov uvozili pet datotek PDF, ki smo jih predhodno pridobili iz programskega okolja Digitization Manager. To nam je omogočilo, da smo na vseh osemdesetih straneh petih izbranih tipkopisov s pomočjo vnaprej danega modela, imenovanega Universal Lines, izvedli analizo postavitve besedila (ang. layout analysis), katere rezultat je viden na Sliki 2.



Slika 2: Prepoznavna regije in vrstic besedila, ki jo je omogočila analiza postavitve besedila, izvedena z uporabo modela Universal Lines

Platforma Transkribus uporabnikom po izvedbi analize postavitve besedila, ki je ključnega pomena za vse nadaljnje korake, omogoča, da na straneh, ki so si jih s predhodno pripravo temeljnega prepisa besedila izbrali za treniranje lastnega jezikovnega modela, vrstice s pomočjo premičnih pik še ročno prilagodijo. Cilj, ki smo ga zasledovali na tej točki, je bil, da bi končni rezultat analize in prilagoditve postavitve besedila čim boljše odlikaval temeljni prepis, ki smo ga iz orodja Notepad

prenesli v sosednje okno in na podlagi katerega je v nadaljevanju potekalo strojno učenje branja besedila.



Slika 3: Rezultat analize postavitve besedila in njegov temeljni prepis

Po tem, ko smo označene vrstice v levem oknu v zadostni meri prilagodili temeljnemu prepisu v desnem oknu, smo v naslednjem koraku stranem, katerim smo dodali temeljni prepis besedila, spremenili status iz *In Progress* v *Ground Truth* in na ta način določili strani, ki so bile v nadaljevanju uporabljene za treniranje lastnega jezikovnega modela. To treniranje je potekalo na način, kot ga predstavljamo v nadaljevanju.

Na podstrani platforme Transkribus, ki je namenjena treniranju lastnih jezikovnih modelov, smo model najprej poimenovali, nato pa izbrali jezik, v katerem je besedilo napisano, določili stoletje njegovega nastanka ter izbrali vnaprej dan javni podporni model (ang. public base model), ki je znatno povečal bazo podatkov za strojno učenje branja besedila. V našem primeru smo pri vseh petih izbranih tipkopisih uporabili podporni jezikovni model Transkribus Typewriter, ki je bil ustvarjen 13. 8. 2021 in treniran na 654.961 besedah oziroma 2.049 straneh tipkopisov v angleščini, nemščini, nizozemščini in finščini. Povprečno število napak na ravni posameznih znakov je bilo za ta model ocenjeno na 1,2 %, kar je, glede na priporočilo razvijalcev platforme Transkribus, ki uporabnikom priporočajo doseganje največ 10 % napak (READ-COOP, 2024), pomenilo dobro izhodišče.

Po zagonu in zaključku treniranja lastnih jezikovnih modelov, ki je potekalo na temelju javnega podpornega modela Transkribus Typewriter in temeljnih prepisov besedil, smo te modele uporabili še na vseh osemdesetih straneh vsakega od izbranih tipkopisov. Rezultat tega postopka, na začetku katerega smo za izvedbo optične prepoznavne znakov izbrali le lasten zasebni model, ki je ustrezal posameznemu tipkopisu, ne pa tudi javnega podpornega modela, sta bili iskalni datoteki PDF in

TXT, ki bi bili, v primeru, da ne bi šlo za osemdešet strani dolge izseke daljših besedil, primerni za objavo na portalu dLib.si.

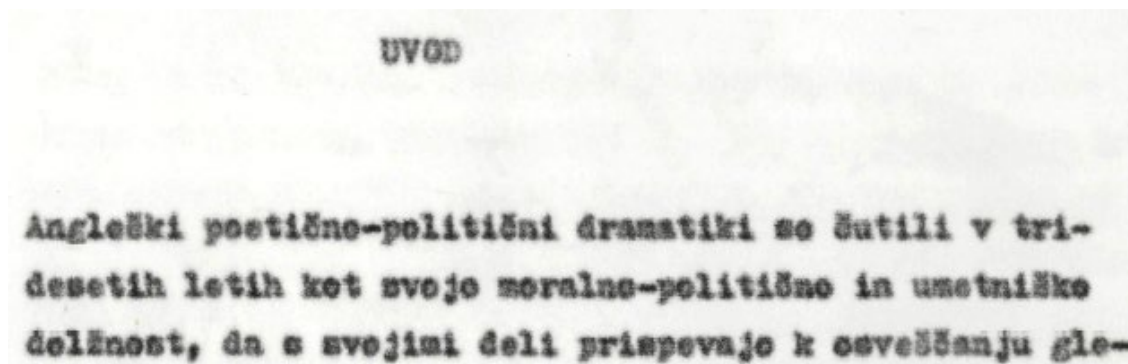
3.3 IZRAČUN NAPAK NA RAVNI POSAMEZNIH ZNAKOV V BESEDILNIH DATOTEKAH, PRIDOBILJENIH Z UPORABO PROGRAMA ABBYY FINEREADER IN PLATFORME TRANSKRIBUS

Po izpeljavi zgoraj opisanih postopkov, ki jih programsko okolje Digitization Manager in platforma Transkribus zahtevata za izvedbo optične prepoznavne znakov, je bilo potrebno izračunati odstotke napak na ravni posameznih znakov (CER), do katerih je prišlo pri uporabi enega in drugega orodja. Za ta namen smo uporabili odprtokodno programsko orodje Dinglehopper, ki nam je ob vnosu ustreznega ukaza v ukazno vrstico omogočilo strojno primerjavo temeljnih prepisov besedil v formatu TXT z enako obsežnimi besedilnimi datotekami v istem formatu, ki smo jih predhodno pridobili s programoma za optično prepoznavo znakov. Orodje Dinglehopper nam je rezultate posredovalo v formatu JSON. Ti so vključevali tako podatek o številu besed v primerjanih besedilih kot podatek o odstotku napak na ravni posameznih znakov (CER).

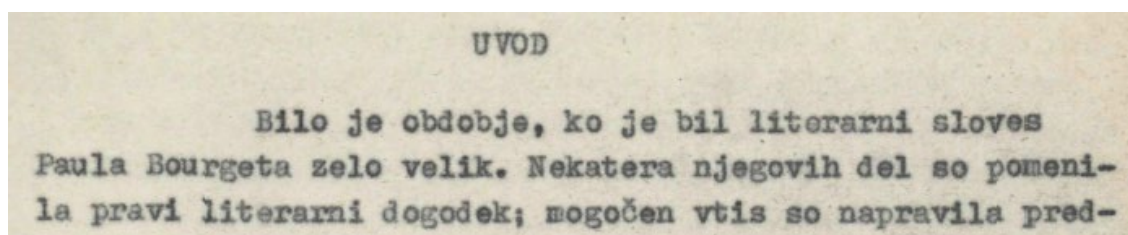
3.4 REZULTATI IZRAČUNOV NAPAK NA RAVNI POSAMEZNIH ZNAKOV V PRIDOBILJENIH BESEDILNIH DATOTEKAH

3.4.1 Dejavnika razločnosti besedil in prisotnosti rokopisnih popravkov

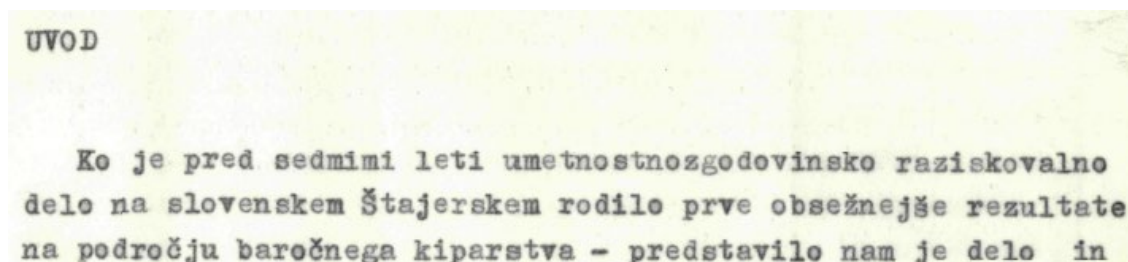
Preden preidemo k predstavitvi rezultatov, ki smo jih pridobili z uporabo orodja Dinglehopper, si nekoliko pogloblje oglejmo še sama besedila, ki smo jih uporabili za izvedbo optične prepoznavne znakov.



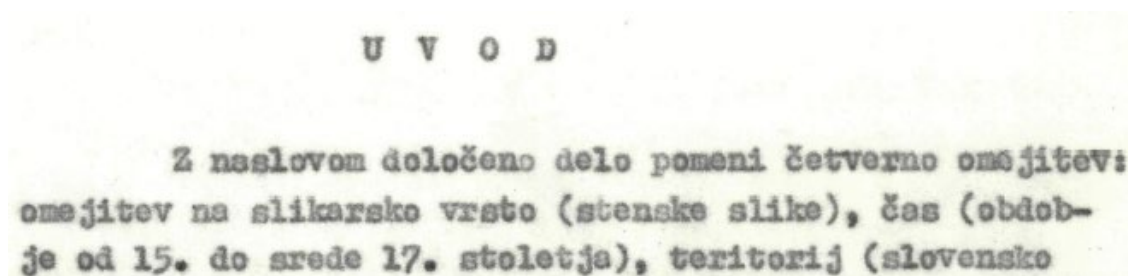
Slika 4: Izsek doktorske disertacije z naslovom Angleška poetično-politična drama v letih 1930-1940 (COBISS.SI-ID - 9680384)



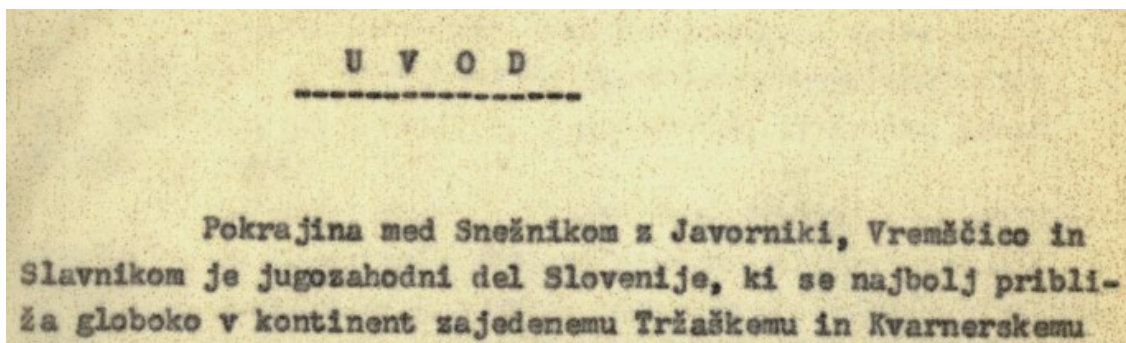
Slika 5: Izsek doktorske disertacije z naslovom Literarna kritika in estetska ideologija Paula Bourgeta (COBISS.SI-ID - 19745378)



Slika 6: Izsek doktorske disertacije z naslovom Kiparstvo na slovenskem Štajerskem v 18. stoletju (COBISS.SI-ID - 196615680)



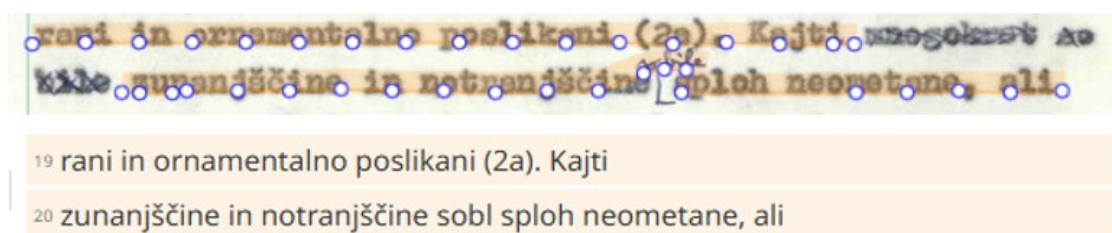
Slika 7: Izsek doktorske disertacije z naslovom Stensko slikarstvo od 15. do srede 17. stoletja na Slovenskem (COBISS.SI-ID - 196893440)



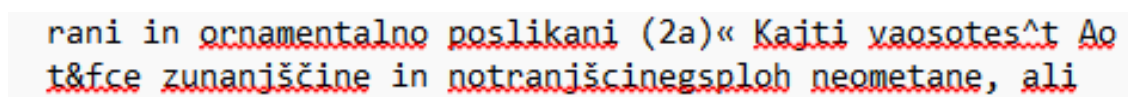
Slika 8: Izsek doktorske disertacije z naslovom Gospodarska geografija pokrajine med Snežnikom z Javorniki, Vremščico in Slavnikom (COBISS.SI-ID - 249056000)

Ob pogledu na prikazane izseke besedil kaj hitro opazimo, da je ključna razlika med tipkopisi, ki smo jih uporabili, prav njihova razločnost. Kajti medtem ko je besedilo na Sliki 6 dobro berljivo, je razločnost pri besedilih na Slikah 5, 7 in 8 že delno okrnjena. Znatno okrnjena pa je razločnost pri besedilu, prikazanem na Sliki 4, ki je bilo, kot smo mimogrede že omenili, za izvedbo optične prepoznavne znakov s programskim orodjem ABBYY FineReader Engine 11 izjemno zahtevno. Kot bomo videli v nadaljevanju, razločnost besedil na optično prepoznavo znakov, izvedeno na platformi Transkribus, ni pomembneje vplivala, saj je bila še pred zagonom postopka optične prepoznavne upoštevana v procesih strojnega učenja branja besedil.

Poleg razločnosti besedila je na kakovost optične prepoznavne znakov s programskim orodjem ABBYY FineReader Engine 11 v manjši meri vplival še en dejavnik, in sicer prisotnost rokopisnih popravkov v obravnavanih besedilih. Za razliko od platforme Transkribus, na kateri je bilo te popravke z ročno prilagoditvijo vrstic po opravljeni analizi postavitve besedila in pred zagonom treniranja lastnega jezikovnega modela mogoče v večji meri upoštevati, se jim namreč pri optični prepoznavi znakov s programom ABBYY FineReader ni bilo mogoče izogniti. Kot prikazujeta Slika 9 in Slika 10, rokopisni popravki na končni rezultat optične prepoznavne znakov na platformi Transkribus posledično skoraj niso imeli vpliva, medtem ko tega za optično prepoznavo, izvedeno s programskim orodjem ABBYY FineReader Engine 11, ne moremo trditi.



Slika 9: Rezultat optične prepoznavne znakov na platformi Transkribus po predhodni prilagoditvi vrstic in izpeljavi postopka treniranja lastnega jezikovnega modela



Slika 10: Rezultat optične prepoznavne znakov v istem delu besedila s programskim orodjem ABBYY FineReader Engine 11

3.4.2 Predstavitev rezultatov

Glede na zgoraj zapisano, rezultati primerjave datotek TXT, ki smo jih pridobili bodisi z uporabo programa ABBYY FineReader bodisi z uporabo platforme Transkribus, s temeljnimi prepisi besedil, ki napak ne vsebujejo, ne bi smeli biti presenetljivi.

Preglednica 1: Rezultati primerjave optično prepoznanih besedil z njihovimi temeljnimi prepisi, izvedene z odprtokodnim orodjem Dinglehopper

COBISS.SI-ID	Število besed	ABBYY FineReader CER	Transkribus CER
9680384	4685	44,48 %	0,47 %
19745378	5024	12,01 %	0,37 %
196615680	5023	1,29 %	0,85 %
196893440	4922	5,52 %	0,25 %
249056000	4754	3,93 %	0,21 %

Kot smo omenili v uvodu v tretje poglavje, smo za namen izvedbe raziskave oziroma potrditve prve hipoteze (H1) izbrali pet doktorskih disertacij z manjšim številom ročnih popravkov in z različnimi stopnjami razločnosti besedila. Od obeh dejavnikov, ki sta vplivala na kakovost optične prepoznavne znakov s programskim orodjem ABBYY FineReader Engine 11, je imela razločnost besedila nedvomno znatno močnejšo, če ne že kar izključne vloge. To potrjuje tako dejstvo, da v tipkopisu, delno prikazanem na Sliki 4 (COBISS.SI-ID - 9680384), in v tipkopisu, prikazanem na

Sliki 6 (COBISS.SI-ID - 196615680), rokopisni popravki skorajda niso prisotni, kot tudi dejstvo, da so v tipkopisih, prikazanih na Slikah 5, 7 in 8, rokopisni popravki razmeroma maloštevilni in med besedili dokaj enakomerno porazdeljeni. Na kar torej lahko sklepamo ob pogledu na odstopanja med odstotnimi točkami napak na ravni posameznih znakov, do katerih je prišlo ob uporabi programa ABBYY FineReader, je, da gre pri teh rezultatih za neposredno odslikavo stopnje strojne berljivosti vsakega izmed obravnavanih besedil. Ob tem naj izpostavimo še, da sta se nad mejo 10 % napak na ravni posameznih znakov, katere preseganje razvijalci platforme Transkribus uporabnikom odsvetujejo (READ-COOP, 2024), znašla le dva od petih tipkopisov, na katerih smo optično prepoznavo znakov izvedli s programom ABBYY FineReader, medtem ko smo pri ostalih dosegli vrednosti, ki nakazujejo, da je optična prepoznavna znakov s tem orodjem zadovoljiva in uporabna.

Za razliko od programskega orodja ABBYY FineReader Engine 11, razločnost besedil na kakovost optične prepoznavne znakov na platformi Transkribus ni vplivala. Kot lahko razberemo iz Preglednice 1, med odstotkoma napak na ravni posameznih znakov v najmanj (COBISS.SI-ID - 9680384) in najbolj (COBISS.SI-ID - 196615680) razločnem besedilu prihaja do nekoliko presenetljivega, a povsem zanemarljivega odstopanja v vrednosti 0,38 odstotne točke v korist manj razločnega besedila. To nakazuje na dejstvo, da je bilo strojno učenje branja besedil, ki je imelo za posledico pet lastnih jezikovnih modelov, uporabljenih za izvedbo optične prepoznavne znakov na platformi Transkribus, izjemno učinkovito in da je bil dejavnik razločnosti besedil v postopku priprave teh modelov v celoti izničen. Ob tem velja izpostaviti tudi, da smo s pripravo lastnih jezikovnih modelov, pri kateri smo uporabili vnaprej dan javni podporni jezikovni model Transkribus Typewriter, dosegli povprečen odstotek napak na ravni posameznih znakov v višini 0,43 odstotne točke. To pomeni, da smo izhodiščni odstotek napak, ki je bil za podporni model Transkribus Typewriter ocenjen na 1,2 %, v povprečju izboljšali za 0,77 odstotne točke. Ta razlika, ki sicer ni velika, a bi bila po vsej verjetnosti v primeru, da bi za optično prepoznavo znakov uporabili zgolj javni podporni model, ki je bil treniran na tipkopisih v angleščini, nemščini, nizozemščini in finščini, ne pa tudi v slovenščini, še nekoliko višja, priča o smiselnosti treniranja lastnih jezikovnih modelov. To je smiselno po eni strani zato, ker vsak odstotek napak na ravni posameznih znakov neposredno vpliva na hitrost in zanesljivost uporabnikovega iskanja informacij, po drugi pa tudi zato, ker bi se v primeru, da bi svoje modele delili z drugimi uporabniki

platforme Transkribus, povečala izhodiščna baza podatkov, na temelju katere deluje javni podporni jezikovni model Transkribus Typewriter. Po tej poti bi, skupaj z drugimi uporabniki platforme, neposredno prispevali k njeni rasti in izboljšanju ter na ta način skrbeli za dolgoročno korist vseh.

3.5 MOŽNOSTI INTEGRACIJE PLATFORME TRANSKRIBUS V PROGRAMSKO OKOLJE DIGITIZATION MANAGER

Kot smo zapisali v uvodu, bomo v zadnjem delu naloge ovrednotili možnosti za vpeljavo platforme Transkribus v potek dela na področju digitalizacije knjižničnega gradiva v NUK, saj gre pri tem za dejavnik, ki ga ne smemo spregledati, ko govorimo o nadaljnjem razvoju Digitalne knjižnice Slovenije. Tu sta se nam kazali dve možnosti, in sicer (1) možnost integracije aplikacijskega programskega vmesnika (API) platforme Transkribus v programsko okolje Digitization Manager in pa – po vsej verjetnosti v primeru, da to ne bi bilo izvedljivo in bi potrdili drugo hipotezo (H2) – (2) možnost nadgradnje programskega okolja Digitization Manager v smeri, ki bi dovoljevala prilagoditev poteka dela na način, da bi datoteke PDF, ustvarjene v programskem okolju Digitization Manager, izvozili, nato v spletni aplikaciji platforme Transkribus izvedli optično prepoznavo znakov v tipkopisih oziroma prepoznavo rokopisnih besedil, in nazadnje datoteke PDF in TXT pred pripravo paketov, namenjenih objavi datotek na portalu dLib.si in trajnemu ohranjanju digitalnih objektov, uvozili nazaj v programsko okolje Digitization Manager. Da bi preverili drugo hipotezo (H2) in izvedljivost rešitve, ki bi utegnila nadomestiti integracijo API-ja platforme Transkribus v programsko okolje Digitization Manager, smo za pogovor zaprosili zaposlene v Službi za informacijsko podporo NUK.

Vprašanja, ki smo jim jih zastavili, so bila sledeča:

1. Ali je API platforme Transkribus mogoče integrirati v programsko okolje Digitization Manager?
2. Bi bilo programsko okolje Digitization Manager mogoče nadgraditi v smeri, ki bi dovoljevala uporabo datotek PDF in TXT, ustvarjenih v spletni aplikaciji platforme Transkribus, za pripravo paketov, namenjenih objavi datotek na portalu dLib.si in trajnemu ohranjanju digitalnih objektov?
3. Katera od obeh rešitev bi bila boljša oziroma enostavnejša za izvedbo?

Glede na to, da sta bila odgovora na prvo in drugo vprašanje pritrdilna, saj tehničnih ovir po besedah zaposlenih v Službi za informacijsko podporo NUK glede na razpoložljivo dokumentacijo platforme Transkribus (Transkribus, 2025) in programskega okolja Digitization Manager ni, in da je bila s tem druga hipoteza (H2) že v začetku pogovora ovržena, je bilo kmalu potrebno odgovoriti le še na tretje vprašanje. A preden se posvetimo temu odgovoru, si nekoliko поблиže oglejmo zahteve, ki bi jih bilo za integracijo API-ja platforme Transkribus v programsko okolje Digitization Manager treba izpolniti, ter prednosti in slabosti prvega in drugega načina vpeljave platforme Transkribus v potek dela na Oddelku za digitalne vsebine NUK.

Kar se tiče integracije API-ja platforme Transkribus v programsko okolje Digitization Manager, je najprej treba reči, da je to mogoče izvesti le ob pogoju plačila polne naročnine, ki je na voljo za raziskovalne in kulturne ustanove, saj le ta omogoča dostop do API-ja, imenovanega Legacy. A za razliko od naročniškega paketa, ki ga nudi podjetje ABBYY in na podlagi katerega sta v programsko okolje Digitization Manager integrirani dve različici programskega orodja FineReader,¹² naročniški paket platforme Transkribus ne omogoča namestitve API-ja na lokalni strežnik ustanove, ki je dostop zakupila. To bi v primeru integracije API-ja Legacy v programsko okolje Digitization Manager pomenilo, da bi vsa dejavnost zaposlenih, ki izvajajo digitalizacijo besedilnih gradiv, potekala na strežnikih izven NUK in da posledično zaposleni v Službi za informacijsko podporo ne bi mogli uvesti lastnih varnostnih standardov oziroma nadgraditi obstoječih in tudi ne nadzorovati delovanja varnostnih protokolov. A glede na to, da platformo Transkribus danes uporabljajo na številnih univerzah ter v knjižnicah, arhivih in drugih dediščinskih ustanovah po svetu, lahko domnevamo, da so strežniki, ki jih za svoje delovanje uporablja platforma Transkribus, dobro zavarovani. Druga težava, ki jo s seboj prinaša dejstvo, da API-ja Legacy ni mogoče namestiti na lokalni strežnik NUK, in s katero bi se srečali tudi v primeru drugega načina vpeljave platforme Transkribus v potek dela na Oddelku za digitalne vsebine, pri katerem bi zaposleni uporabljali spletno aplikacijo platforme Transkribus, pa se tiče besedil, ki so

¹² Kot smo omenili v podpoglavju 2.3, v procesih digitalizacije besedilnih gradiv v NUK programsko orodje ABBYY FineReader Server 14 služi optični prepoznavi tiskanih besedil v gotici, medtem ko je za tiskana latinična besedila in tipkopise v uporabi programsko orodje ABBYY FineReader Engine 11.

avtorskoppravno zaščiten. Kot je zapisano v členu 6.3.1. Splošnih pogojev, objavljenih na spletni strani platforme Transkribus (Transkribus, 2024), mora namreč uporabnik, ki želi na strežnikih platforme Transkribus izvajati postopke optične prepoznavne na avtorskoppravno zaščitenem gradivu, za to prehodno pridobiti dovoljenje avtorja oziroma imetnika avtorske pravice.

Kljub tej oviri, ki bi se ji izognili, če bi bilo API Legacy mogoče namestiti na lokalni strežnik NUK, in pomislekom glede varnosti, pa ne gre spregledati dejstva, da so prednosti rabe platforme Transkribus številne. Strojno učenje, na temelju katerega platforma deluje, namreč ne omogoča le izjemno kakovostne optične prepoznavne znakov v tipkopisih, temveč tudi ustvaritev iskalnih datotek PDF digitalno preslikanih rokopisnih besedil. Iz tega razloga lahko trdimo, da je to, da namestitev API-ja Legacy na lokalni strežnik matične ustanove ni mogoča, bolj okoliščina in ne toliko dejavnik, zaradi katerega se v NUK za integracijo API-ja platforme Transkribus v programsko okolje Digitization Manager ne bi odločili.

Po besedah zaposlenih v Službi za informacijsko podporo NUK je torej z integracijo API-ja Legacy mogoče zgraditi avtomatizirano rešitev, ki bi omogočala nalaganje dokumentov na strežnike platforme Transkribus, izvajanje postopkov optične prepoznavne znakov v tipkopisih oziroma prepoznavne rokopisnih besedil ter prenos rezultatov teh postopkov neposredno v programsko okolje Digitization Manager. K temu dodajajo, da bo API platforme Transkribus, sodeč po dokumentaciji, objavljeni na spletu (Transkribus, 2025), v letu 2025 korenito spremenjen in nadgrajen ter da bi bilo iz tega razloga z integracijo API-ja v programsko okolje Digitization Manager potrebno počakati, saj bodo postopki optične prepoznavne najverjetneje spremenjeni, trenutna rešitev pa čez čas zelo verjetno ne bo več podprta. Raba API-ja Legacy, ki je trenutno še aktiven, sicer vključuje več zaporednih korakov. V prvem koraku je potrebno izvesti avtentikacijo uporabnika, ki jo ta opravi bodisi s klicem na /auth/login bodisi z uporabo trajnega žetona API, ki ga je pridobil prek uporabniškega vmesnika platforme Transkribus. Po uspešni prijavi ima uporabnik v naslednjem koraku možnost ustvariti novo zbirko, v katero naloži dokumente. Vsaka nova zbirka prejme svojo identifikacijsko kodo (ID), na podlagi katere poteka nadaljnje upravljanje z dokumenti in s postopki optične prepoznavne besedil. V tretjem koraku uporabnik vhodne datoteke v formatih TIFF, JPEG ali PDF pošlje na strežnik platforme Transkribus z uporabo klica POST na /uploads. Strežnik uporabniku na tej točki vrne identifikator dokumenta (docId), ki je potreben za

nadaljnje delo. Po uspešni naložitvi dokumenta uporabnik v četrtem koraku sproži postopek optične prepoznavne znakov v tipkopisih oziroma prepoznave rokopisnih besedil s klicem na `/recognition/{docId}/{pageNr}`, za izvedbo katerega so potrebni identifikator dokumenta, podatek o številki strani (oziroma o tem, da so bile izbrane vse strani) ter izbira jezikovnega modela. Po uspešni izvedbi postopka optične prepoznavne znakov uporabnik v petem koraku prek klica na `/documents/{docId}/pages/{pageNr}/text` od strežnika prejme optično prepoznano besedilo, in sicer tako v obliki iskalne datoteke PDF kot v obliki datoteke TXT. Z drugimi besedami, programsko okolje Digitization Manager bi po izpeljavi integracije API-ja na standardni dvosmerni način (REST API) komuniciralo z oblačno rešitvijo API Legacy in bi prek zgoraj opisanih korakov (avtentikacija, ustvaritev zbirke, naložitev dokumentov, optična prepoznavna besedila, prenos besedila) zaposlenemu, ki izvaja digitalizacijo besedilnih gradiv, znotraj programskega okolja Digitization Manager omogočalo izvedbo optične prepoznavne znakov v tipkopisih oziroma prepoznavne rokopisnih besedil.

Če se sedaj na kratko pomudimo še pri vpeljavi platforme Transkribus v potek dela na področju digitalizacije knjižničnega gradiva v NUK na način, ki bi dovoljeval izvoz datotek PDF iz programskega okolja Digitization Manager, njihovo obdelavo v spletni aplikaciji platforme Transkribus ter naknadni uvoz datotek PDF in TXT v programsko okolje Digitization Manager, je v prvi vrsti treba reči, da bi za tovrstno rešitev prišel v poštev le timski naročniški paket, kakršnega nudi platforma Transkribus. V primeru zakupa dostopa do platforme Transkribus po tej poti bi v NUK spletno aplikacijo lahko uporabljalo pet zaposlenih, ki bi bili omejeni na sto zagonov postopka treniranja lastnih jezikovnih modelov na mesec in na 1 TB prostora za hrambo dokumentov. Poleg tega bi morali zaposleni, ki bi želeli izvajati postopke optične prepoznavne na avtorskopravno zaščitenem gradivu – enako kot pri zgoraj opisani rešitvi, ki bi temeljila na integraciji API-ja Legacy v programsko okolje Digitization Manager – za to pridobiti dovoljenje avtorja oziroma imetnika avtorske pravice. Za razliko od zgoraj opisane rešitve, katere izvedba zahteva plačilo polne naročnine, pa bi bili zaposleni v okviru timskega naročniškega paketa deležni okrnjene podpore strankam. In če primerjamo obe rešitvi, je nenazadnje potrebno vzeti v obzir tudi dejstvo, da bi se v primeru rabe spletne aplikacije platforme Transkribus delovni postopki dodatno zapletli, saj bi morali zaposleni za izvedbo bodisi optične prepoznavne znakov v tipkopisih bodisi prepoznavne rokopisnih besedil prehajati med dvema programskima okoljema.

Sodeč po zgoraj zapisanem lahko trdimo, da je druga od obeh možnosti – čeprav cenejša in lažje izvedljiva – slabša. Res je sicer, da omejitve pri uporabi spletne aplikacije platforme Transkribus zaposlenih v NUK najverjetneje ne bi ovirale. Za izvedbo postopkov optične prepoznavne znakov v tipkopisih oziroma prepoznavne rokopisnih besedil bi namreč uporabljali prostorsko manj potratne datoteke PDF in ne izvornih skenogramov v formatu TIFF, s čimer bi se izognili nevarnosti izrabe prostora za hrambo dokumentov. Prav tako bi zaposleni v primeru rabe spletne aplikacije platforme Transkribus le stežka presegli število zagonov postopka treniranja jezikovnih modelov, saj pri optični prepoznavi znakov v tipkopisih ne bi nujno trenirali lastnih jezikovnih modelov, temveč bi lahko uporabili javne podporne modele, kot je na primer Transkribus Typewriter. Kljub temu pa bi po besedah zaposlenih v Službi za informacijsko podporo NUK druga od obeh možnosti prišla v poštev le v primeru, da bi se prva zaradi tehničnih ovir in posledične nezdružljivosti programskega okolja Digitization Manager z API-jem Legacy izkazala za neizvedljivo. Temu pa, kot smo pokazali, ni tako.

4 ZAKLJUČEK

V prvem, teoretičnem delu pisne naloge smo opisali značilnosti procesov optične prepoznavne znakov in prepoznavne rokopisnih besedil ter na podlagi opisa razvoja, ustroja in uporabniških možnosti programskega orodja ABBYY FineReader Engine 11, ki se v NUK za izvajanje optične prepoznavne znakov že uporablja, in platforme Transkribus predstavili novost, ki jo je razvoj umetnih nevronske mreže in strojnega učenja prinesel v polje optične prepoznavne znakov in prepoznavne rokopisnih besedil. Kot se je izkazalo pri natančnejšem vpogledu v ustroj platforme Transkribus, se ta novost ne tiče le izjemne kapacitete tega orodja, ki raziskovalcem besedilnih korpusov odpira nove, desetletja nazaj neslutene raziskovalne možnosti, temveč zadeva tudi samo zasnovo platforme, ki štiri interesne skupine (arhive oziroma knjižnice, raziskovalce, pripadnike splošne javnosti ter računalničarje oziroma razvijalce tehnologije) povezuje na način, da platformo Transkribus ne le uporabljajo, temveč z uporabo tudi prispevajo k njenemu razvoju.

S teoretičnim delom smo bralca uvedli v osrednji, empirični del naloge, v katerem smo najprej preverjali naslednjo raziskovalno hipotezo:

- H1: Platforma Transkribus v primeru optične prepoznavne znakov v tipkopisih nudi boljše rezultate kot programsko orodje ABBYY FineReader Engine 11. S tem omogoča lažje iskanje informacij in odpira možnost za izboljšanje uporabniške izkušnje portala dLib.si.

Kot se je izkazalo na primeru petih izbranih tipkopisov, je bila platforma Transkribus pri optični prepoznavi znakov bistveno učinkovitejša od programskega orodja ABBYY FineReader Engine 11, saj so bile posebnosti obravnavanih besedil, kot sta na primer razločnost posameznih znakov in prisotnost rokopisnih popravkov, na platformi Transkribus bolje upoštevane, in sicer že samem v procesu treniranja lastnih jezikovnih modelov. Že samo ta primer, ki niti ne jemlje v obzir vseh možnosti, ki jih platforma ponuja,¹³ kaže, da je potencial platforme za izboljšanje rezultatov optične prepoznavne digitalno preslikanih besedil in s tem za povečanje dostopnosti informacij na portalu dLib.si velik.

Potrditve prve hipoteze, s katero smo dokazali smiselnost rabe platforme Transkribus za izvajanje optične prepoznavne znakov v tipkopisih, nas je vodila k preverjanju druge hipoteze:

- H2: Funkcionalnosti platforme Transkribus ni mogoče integrirati v programsko okolje Digitization Manager. Raba platforme Transkribus kot zunanega orodja za izvedbo optične prepoznavne znakov v tipkopisih bi zato zahtevala prilagoditev poteka dela na področju digitalizacije knjižničnega gradiva in trajnega ohranjanja digitalnih objektov.

Ovrednotenje možnosti za vpeljavo platforme Transkribus v potek dela na področju digitalizacije knjižničnega gradiva v NUK smo izvedli s preučitvijo možnosti za integracijo aplikacijskega programskega vmesnika (API) platforme Transkribus v programsko okolje Digitization Manager in s primerjavo te rešitve z morebitno nadomestno nadgradnjo programskega okolja Digitization Manager v smeri, ki bi dovoljevala uvoz datotek PDF in TXT, ustvarjenih v spletni aplikaciji platforme Transkribus. V pogovoru z zaposlenimi v Službi za informacijsko podporo NUK smo ugotovili, da je integracija API-ja platforme Transkribus izvedljiva in da kljub dejstvu, da API-ja Legacy ni mogoče namestiti na strežnik NUK, pomeni boljšo rešitev od zgoraj omenjene nadgradnje programskega okolja Digitization Manager.

¹³ Tu imamo v mislih predvsem prepoznavo rokopisnih besedil, s katero se v nalogi zaradi metodoloških omejitev nismo ukvarjali.

Rezultati preverjanja obeh hipotez nakazujejo, da bi bilo razvoj Digitalne knjižnice Slovenije smiselno voditi v smeri integracije aplikacijskega programskega vmesnika (API) platforme Transkribus v programsko okolje Digitization Manager, ki ga zaposleni v Oddelku za digitalne vsebine NUK uporabljajo za digitalizacijo besedilnega in slikovnega gradiva, saj bi na ta način vsebino digitalizirane pisne kulturne dediščine lahko naredili še dostopnejšo in s tem izboljšali uporabniško izkušnjo portala dLib.si.

5 NAVEDENI VIRI IN LITERATURA

- ABBYY. (b. d.). *ABBYY FineReader Engine 11: the most comprehensive SDK for recognition and document conversion*. <https://static1.abbyy.com/abbyycommedia/7292/fr-engine-11-e-a3-online-abbyy-hq.pdf>
- Binarization. (21. 4. 2025). V *Cloudinary*. <https://cloudinary.com/glossary/binarization>
- Colutto, S., Kahle, P., Guenter, H. in Muehlberger, G. (2019). Transkribus. A platform for automated text recognition and searching of historical documents. V *15th International Conference on eScience (eScience)* (str. 463–466). IEEE.
<https://doi.org/10.1109/eScience.2019.00060>
- Goodwin, M. (9. 4. 2024). *What is an API (application programming interface)?* IBM.
<https://www.ibm.com/think/topics/api>
- Holdsworth, J. in Scapicchio, M. (17. 6. 2024). *What is deep learning?* IBM.
<https://www.ibm.com/think/topics/deep-learning>
- IBM. (b. d.). *What are convolutional neural networks?*
<https://www.ibm.com/think/topics/convolutional-neural-networks>
- IBM. (22. 9. 2021). *What is machine learning?* https://www.ibm.com/think/topics/machine-learning?mhsrc=ibmsearch_a&mhq=machine%20learning
- IFLA/UNESCO Manifesto for digital libraries. (2011). International Federation of Library Associations and Institutions. <https://repository.ifla.org/handle/20.500.14598/3583>

- Memon, J., Sami, M., Khan, R. A. in Uddin, M. (2020). Handwritten optical character recognition (OCR): a comprehensive systematic literature review (SLR). *IEEE Access*, 8, 142642–142668. <https://doi.org/10.1109/ACCESS.2020.3012542>
- Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinöcker, 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://doi.org/10.1108/JD-07-2018-0114>
- Nockels, J., Gooding, P., Ames, S. in Terras, M. (2022). Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. *Archival Science*, 22(3), 367–392. <https://doi.org/10.1007/s10502-022-09397-0>
- 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://doi.org/10.1108/JD-09-2023-0183>
- READ-COOP. (9. 5. 2023). What is ground truth? *Transkribus blog*. <https://blog.transkribus.org/en/what-is-ground-truth>
- READ-COOP. (16. 4. 2024). How to improve the CER of your model. *Transkribus blog*. <https://blog.transkribus.org/en/how-to-improve-the-cer-of-your-model>
- READ-COOP SCE. (2023). *Our members*. <https://readcoop.org/members>
- Transkribus. (8. 11. 2024). *General terms and conditions*. <https://legal.transkribus.org/terms>
- Transkribus. (2025). *Transkribus legacy API*. <https://help.transkribus.org/transkribus-legacy-api>