



PAM Pokrajinski
arhiv
Maribor

Moderna
arhivistika

Časopis arhivske teorije in prakse
Journal of Archival Theory and Practice

ISSN 2591-0884

<https://doi.org/10.54356/MA>

Letnik 6 (2023), št. 1 / Year 6 (2023), No. 1

Maribor, 2023

Prejeto / Received: 05. 05. 2023

1.02 Pregledni znanstveni članek

1.02 Review article

<https://doi.org/10.54356/MA/2023/JXVT8129>

OPERATIVNA UPORABNOST ORODIJ ZA PREPOZNAVANJE Z ROKO NAPISANEGA BESEDILA NA PODLAGI ANALIZE RAZPOZNANIH VSEBIN

Jan COTIČ

arhivist, Pokrajinski arhiv Koper, Slovenija
jan.cotic@arhiv-koper.si

Ivančica SABADIN

informatik, Pokrajinski arhiv Koper, Slovenija
ivancica.sabadin@arhiv-koper.si

Izvleček:

Prepoznavanje z roko napisanega besedila oz. rokopisa (angl. *handwritten text recognition*) je tehnologija, ki bi olajšala problematiko prepoznavanja vsebine tovrstnih dokumentov. Cilj uporabe orodja za prepoznavanje z roko napisanega besedila je ustvaritev besedila, ki bi omogočilo hitrejše razumevanje, večjo dostopnost in lažjo uporabo specifičnega gradiva ter omogočil prosto iskanje po celotnem besedilu (angl. *full-text search*). Avtorja sta čez analizo treh dostopnih orodij za prepoznavanje z roko napisanega besedila predstavila postopek in dobljene rezultate na digitaliziranem gradivu Pokrajinskega arhiva Koper. Rezultati so pokazali, da je povprečna napaka za vsa tri orodja 57 %, kar pomeni, da so orodja zaznala narobe več kot polovico besed.

Ključne besede:

prepoznavanje z roko napisanega besedila, Transkribus, Microsoft Azure's Computer Vision, Google Cloud Vision AI, umetna inteligenca

Abstract:

Operational Usefulness of Handwritten Text Recognition Tools

Handwritten Text Recognition (HTR) is a technology that would ease the problem of recognising the content of written documents. The aim of using handwritten text recognition technology is to create a text that makes certain material more understandable, accessible and easy to use. It also allows for full-text searching. The authors present the procedure and the results obtained on the digitised material of the Regional Archives Koper through an analysis of three available tools for handwritten text recognition. The results showed that the average error for all three tools was 57%, which means that the tools recognised more than half of the words incorrectly.

Key words:

Handwritten Text Recognition, Transkribus, Microsoft Azure's Computer Vision, Google Cloud Vision AI, Artificial Intelligence

1 UVOD

Pričujoči prispevek o analizi in primerjavi dostopnih orodij za prepoznavanje z roko napisanega besedila (rokopisa) ali HTR (angl. handwritten text recognition) obravnava natančnost in uporabnost te tehnologije, ki s procesom »transkribiranja« ustvari prepis tujih pisav v latinico (Glavič, 2020, str. 87). Omenjeni proces omogoča dostopnost težje berljivih zapisov in lažje razumevanje le-teh širši populaciji.

Za raziskavo sta avtorja uporabila deset različnih dokumentov, ki jih hrani Pokrajinski arhiv Koper. Koprski arhiv hrani celoten fond piranske komune od leta 1173, statute od 1307, notarske in vicedomske knjige, dukale, testamente itd. Bonin (2006) pravi, da so za starejšo zgodovino prav tako pomembni tudi fondi koprskih patricijskih družin, sodni fondi, gradiva šol, različnih zavodov itd.

Ozemlje severozahodne Istre je bilo od konca 13. st. dalje formalnopravno del Beneške Republike, je pa ga vpliv z ozemlja zahodne obale Jadranskega morja v kulturnem, pravnem in političnem smislu zaznamoval že prej. Po prihodu pod beneško oblast so istrska mesta v glavnem sprejela beneško zakonodajo in jo prilagodila svojim potrebam (Darovec, 2009). Zgodovinske okoliščine so v Istri skozi več stoletij narekovale drugačen razvojni tok. Na tem območju je še danes stičišče različnih narodov, to je botrovalo k prepletu številnih kultur in jezikov, kar se pri arhivskemu gradivu odraža z bogato in raznovrstno pisno dediščino, napisano v različnih jezikih (npr. v latinščini, italijanščini, slovenščini, nemščini itd.). Avtorja sta se v ta namen odločila izbrati tiste primere dokumentov, ki s svojo vsebino izpričujejo pomembne in zanimive dogodke iz različnih časovnih obdobjih severozahodne Istre z okolico.

Za dodatne primerjalne analize izbranih orodij za transkribiranje sta namensko bila izbrana dva stilsko in slogovno med seboj podobna si piranska dukala iz beneškega obdobja.

2 METODOLOGIJA

V prispevku je uporabljenih več znanstvenih metod. Z izkustveno metodo je izbranih deset dokumentov iz arhivskega gradiva Pokrajinskega arhiva Koper (tabela 1).

Tabela 1: Izbrani dokumenti iz arhivskega gradiva Pokrajinskega arhiva Koper

Oznaka dokumenta	Naslov	Jezik
1	Piranski dukal z dne 8. 4. 1570	Italijanski jezik (ITA)
2	Piranski dukal z dne 30. 3. 1570	Italijanski jezik (ITA)
3	Tartinijevo pismo (prva stran)	Italijanski jezik (ITA)
4	Doktorska diploma padovske univerze Jacobusa Vittorija, sina Petra, 1806	Latinski jezik (LAT)

5	Ustanovitev koprške knjižnice	Italijanski jezik (ITA)
6	Pismo Francesca Grisonija	Italijanski jezik (ITA)
7	Izvleček iz Piranskega statuta	Latinski jezik (LAT)
8	Prospero Petronio, Brani della parte prima delle Memorie sacre e profane dell'Istria 1680-1681	Latinski jezik (LAT)
9	Listina avstrijskega cesarja Franca Jožefa I. z dne 15. januarja 1911, s katero je povzdignil Ilirsko Bistrico v trg	Slovenski jezik (SLO)
10	Šolska kronika državne osnovne šole v Gabrovici na Krasu za šolsko leto 1953/54	Slovenski jezik (SLO)

Izbor navedenih dokumentov je bil analiziran in primerjan v dveh različnih ločljivostih, 300 DPI¹ in 72 DPI, z namenom ugotoviti, ali resolucija optično preslikanih izvirkov vpliva pri samem procesu transkribiranja. Za ustrezen izračun odstotka napak je potrebno točno število besed besedila v izvirkniku, ki ga dobimo s pripravo natančnega in točnega prepisa. Dva od desetih izbranih dokumentov (št. 3² in št. 7³) imata že obstoječ prepis besedila v obliki tipkopisa, za preostale dokumente sta avtorja samostojno pripravila natančne prepise. Novonastale transkripcije so rezultat procesa spoznavanja in razumevanja različnih besedil ter primerjanja med seboj vsebinsko podobnih si dokumentov, kjer je bilo treba seznaniti se s pisavami, jezikovnimi posebnostmi, krajšavami besed in drugimi značilnostmi. Na tovrstni proces priprave točnega prepisa opozarja že Glavič, J. (2020), saj v svojem strokovnem članku izpostavlja, da gre za sistematičen, dolgoročen in zamuden postopek.

Z metodo primerjave sta avtorja primerjala avtomatično transkribiran dokument z ustvarjenim prepisom. Za primerjavo je uporabljeno odprtokodno orodje Notepad++. Poudarek prispevka je na razumevanju besedil, v ta namen sta avtorja uporabila enačbo za računanje napake (slika 1).

Slika 1: Enačba za računanje napake (Patel, 2019, str. 35)

$$\text{Napaka} = \frac{\text{Število napačno prepoznanih besed}}{\text{Število besed v izvirknem besedilu}}$$

Rezultati primerjave so obdelani s statistično metodo, ki je bila uporabljena za ustvarjanje tabel in grafičnih prikazov obdelanih podatkov. Za opisovanje delovanja orodij je uporabljena deskriptivna metoda.

¹ DPI – angl. dots per inch.

² Tartinijevo pismo (prva stran).

³ Izvleček iz Piranskega statuta.

3 TRANSKRIBUS

Opis orodja

Transkribus je celovita rešitev, ki obsega prepoznavanje besedila, transkripcijo in iskanje dokumentov (READ - COOP, b. d.). Platforma Transkribus Univerze v Innsbrucku je izdelana v sklopu evropskega projekta Horizon 2020 »READ«⁴ za potrebe raziskovanja, izobraževanja, znanosti in poslovanja. Orodje je namenjeno štirim skupinam uporabnikov: arhivistom, znanstvenikom s področja humanistike, informatikom in uporabnikom, ki jih zanima preučevanje in uporaba zgodovinskih dokumentov. Interesi skupin uporabnikov se med seboj prepletajo in vsaka od njih prispeva k razvoju infrastrukture Transkribusa (Muehlberger et al., 2019). Vsak uporabnik Transkribusa lahko ustvari svoj model glede na pisavo, vsebino in jezik dokumenta. Glavič (2020) pravi, da se uporabnost Transkribusa izkaže, ko je treba transkribirati večjo zbirko ročno pisanih dokumentov. Za ustvarjanje modela v Transkribusu potrebuje raziskovalec okoli 50 strani ročno napisanega besedila. Za dostop sta na voljo dve možnosti. S pomočjo brskalnika se dostopa do vmesnika, ki omogoča le osnovne možnosti. Za uporabnike, ki želijo bolj podrobno analizo in ustvarjanje modela, je na voljo aplikacija, ki se lahko namesti na računalnik. Transkribus ni brezplačno orodje, temveč ima vsak uporabnik ob začetku uporabe na voljo 500 kreditnih točk, ki jih lahko uporabi za transkribiranje in ustvarjanje modelov.

Cilj trenutne raziskave ni bil ustvarjanje modela in merjenje njegove zanesljivosti, ampak primerjava orodij Transkribus, Google Cloud Vision AI in Microsoft Azure's Computer Vision. Zato so bili večinoma izbrani dokumenti, za katere že obstajajo modeli v Transkribusu, razen dveh dokumentov v slovenščini, ki nista bila analizirana s Transkribusom.

Transkribus sodi med bolj zastopana orodja v kontekstu prepoznavanja ročno napisanega besedila. Orodje temelji na obstoječih modelih prepoznavanja jezikov, kar uporabniku omogoči lastno izbiro algoritma za posamezen jezik oz. jezikovno skupino sorodnih jezikov. Edini ustrezen model za naš izbor dokumentov je bil model Univerze v Torontu za latinski jezik. Ostali modeli za prepoznavanje pisave, ki prav tako vključujejo latinščino, niso ustrezali prepoznavanju besedila. Model »University of Toronto« ni brezhiben, saj ima 'naučenih' 140.158 besed, medtem ko so drugi modeli (npr. za nemški, nizozemski jezik itd.) dopolnjeni z več milijoni besed. Poleg neustreznih modelov za latinščino ni modelov za slovenski jezik.

Rezultati analize

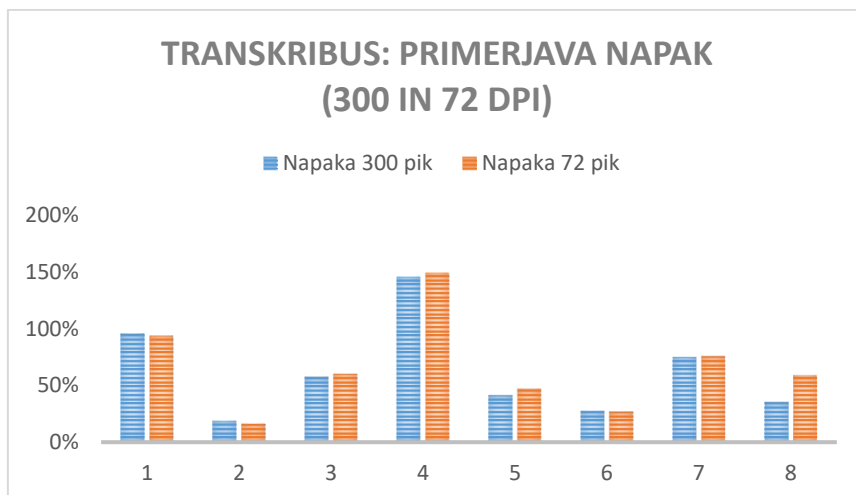
Prvo analizirano orodje je bilo Transkribus. Rezultati so prikazani v tabeli 2. Vrednosti so pokazale, da je povprečna napaka 64 %, minimalno 16 %, maksimalno pa 149 % za 16 skeniranih primerov od 20. Štirih primerov ni bilo mogoče vključiti v analizo, ker ni ustreznega modela za slovenski jezik. Pri dokumentu št. 4 je zaznana napaka od 145 % za 300 DPI in 149 % za 72 DPI, ker je Transkribus zaznal 53 % več besed, kot jih je v izvorniku.

Na podlagi opravljene analize niso opažena večja odstopanja pri spremembi ločljivosti, izjemoma pri dokumentu št.8. Povprečna napaka za slike v 300 DPI je 62 % in 66 % za slike v 72 DPI. Podrobna analiza primerjave napak med ločljivostjo dokumentov je prikazana na sliki 2.

⁴ Več informacij dostopno na: <https://readcoop.eu/about/> (24. 4. 2023).

Tabela 2: Rezultati analize transkribiranja z orodjem Transkribus

Datoteka	Jezik	Uporabljeni model	300 DPI					72 DPI				
			Število besed v izvorniku	Zaznano besed	Točno besed	Narobe besed	Napaka	Število besed v izvorniku	Zaznano besed	Točno besed	Narobe besed	Napaka
1	ITA	Italian administrative hands 1500-1700	130	145	21	124	95 %	130	144	22	122	94 %
2	ITA	Italian administrative hands 1500-1700	149	149	121	28	19 %	149	146	122	24	16 %
3	ITA	Transkribus Italian Handwriting	416	425	185	240	58 %	416	427	177	250	60 %
4	LAT	University of Toronto	216	331	17	314	145 %	216	328	6	322	149 %
5	ITA	Transkribus Italian Handwriting	51	57	36	21	41 %	51	56	32	24	47 %
6	ITA	Transkribus Italian Handwriting	157	151	108	43	27 %	157	151	109	42	27 %
7	LAT	University of Toronto	95	93	22	71	75 %	95	92	20	72	76 %
8	LAT	Transkribus Print M1	34	23	11	12	35 %	34	20	0	20	59 %



Slika 2: Primerjava napak pri 300 DPI in 72 DPI orodja Transkribus

4 Google Cloud Vision AI

Opis orodja

Google Cloud Vision AI je del platforme Google Cloud, ki uporabnikom ponuja veliko možnosti, kot so označevanje slik, zaznavanje obrazov, besedil in logotipov. Za uporabo vseh možnosti mora uporabnik imeti dostop do platforme Google Cloud. V primerjavi s Transkribusom je delo z orodjem Google Cloud Vision AI primerno za zahtevnejše uporabnike. Na spletni strani Google Cloud je na voljo demo različica, ki ima omejitve pri nalaganju slik. Velikost slik ne sme presegati 4 MB⁵. Za potrebe raziskave je uporabljen »REST API«⁶ oziroma ukazi, ki so omogočili komunikacijo z vmesnikom Google Cloud Vision AI. Rezultat je bila datoteka v obliki JSON⁷, ki je z ustreznim programskim orodjem pretvorjena v TXT⁸ obliko, iz katere je razvidno besedilo. Pri prepoznavanju ročno napisanega besedila z orodjem Google Cloud Vision AI lahko uporabnik s pomočjo ustreznega ukaza ponudi nasvet ali »languageHints« glede jezika. Po uporabniških navodilih v večini primerov »languageHints« ni treba uporabljati, saj je orodje opremljeno s funkcijo avtomatičnega zaznavanja jezika (Google Cloud, b. d.).

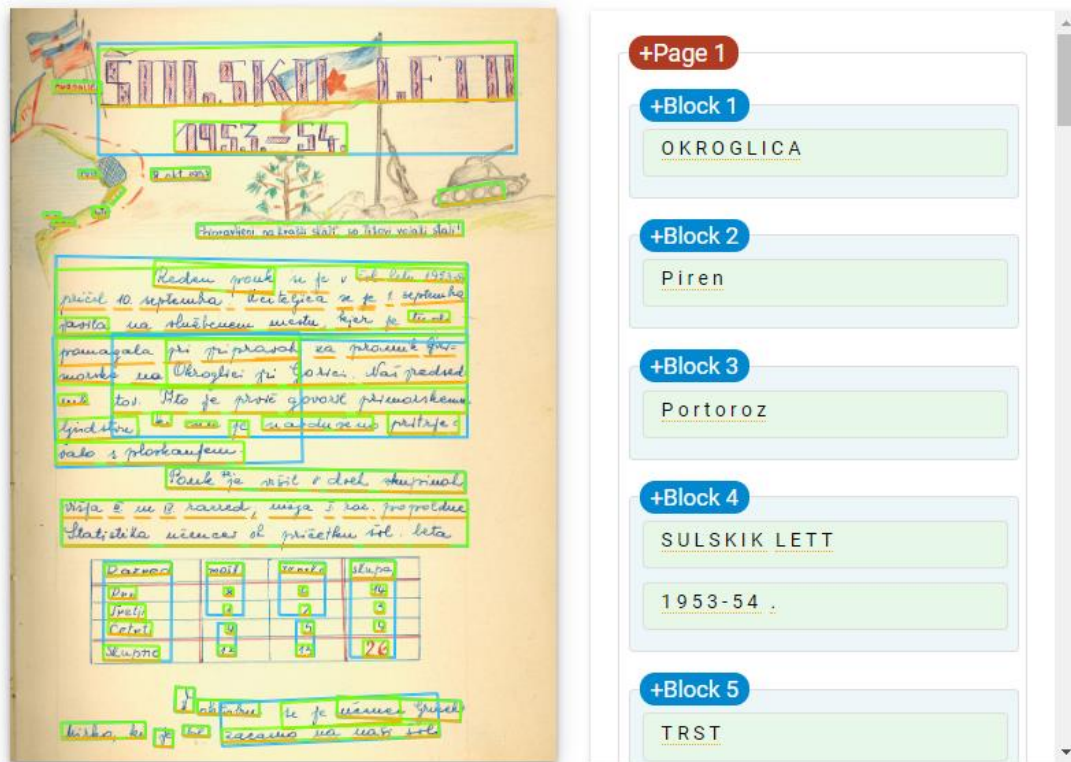
Google Cloud Vision AI deluje na podlagi hierarhije zaznavanja besedila: stran (angl. Page), okvirji (angl. Block), ki zaznajo odstavke, besede in presledke (slika 3). Pri samodejnem zaznavanju jezika in elementov na dokumentu lahko pride do napačne interpretacije vsebine.

⁵ MB – megabajt.

⁶ REST API ali »RESTful Application programming interface« je vmesnik, ki omogoča komunikacijo s spletnimi storitvami.

⁷ JSON – angl. JavaScript Object Notation.

⁸ TXT – angl. text.



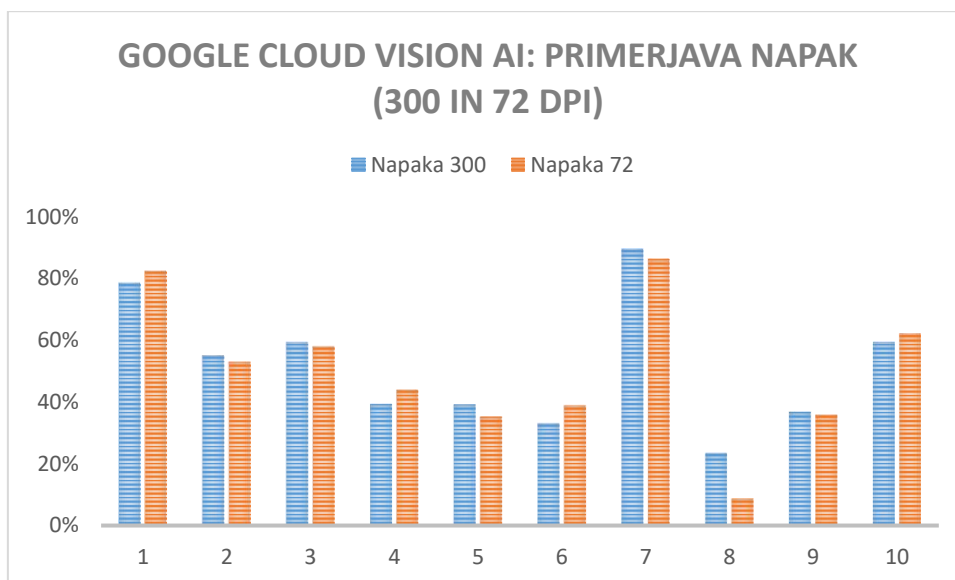
Slika 3: Prepoznavanje besedila s pomočjo orodja Google Cloud Vision AI

Rezultati analize transkribiranja z orodjem Google Cloud Vision AI so prikazani v tabeli 3. Pokazali so, da je povprečna napaka 51 %, minimalno 9 %, maksimalno 89 % za 20 skeniranih primerov.

Tabela 3: Rezultati analize transkribiranja z orodjem Google Cloud Vision AI

Datoteka	Jezik	300 DPI					72 DPI				
		Število besed v izvirniku	Zaznane besede	Točne besede	Napačne besede	Napaka	Število besed v izvirniku	Zaznane besede	Točne besede	Napačne besede	Napaka
1	ITA	130	161	28	102	78 %	130	149	23	107	82 %
2	ITA	149	157	67	82	55 %	149	154	70	79	53 %
3	ITA	416	466	169	247	59 %	416	457	175	241	58 %
4	LAT	216	208	131	85	39 %	216	204	121	95	44 %
5	ITA	51	65	31	20	39 %	51	59	33	18	35 %
6	ITA	157	153	105	52	33 %	157	153	96	61	39 %
7	LAT	95	95	10	85	89 %	95	95	13	82	86 %
8	LAT	34	34	26	8	24 %	34	34	31	3	9 %
9	SLO	106	107	67	39	37 %	106	108	68	38	36 %
10	SLO	74	79	30	44	59 %	74	76	28	46	62 %

Povprečna napaka za slike v 300 DPI je 51 % in 50 % za slike v 72 DPI. Podrobna analiza primerjave napak med ločljivosti dokumentov je prikazana na sliki 4.



Slika 4: Primerjava napak pri 300 DPI in 72 DPI orodja Google Cloud Vision AI

5 Microsoft Azure's Computer Vision

Opis orodja

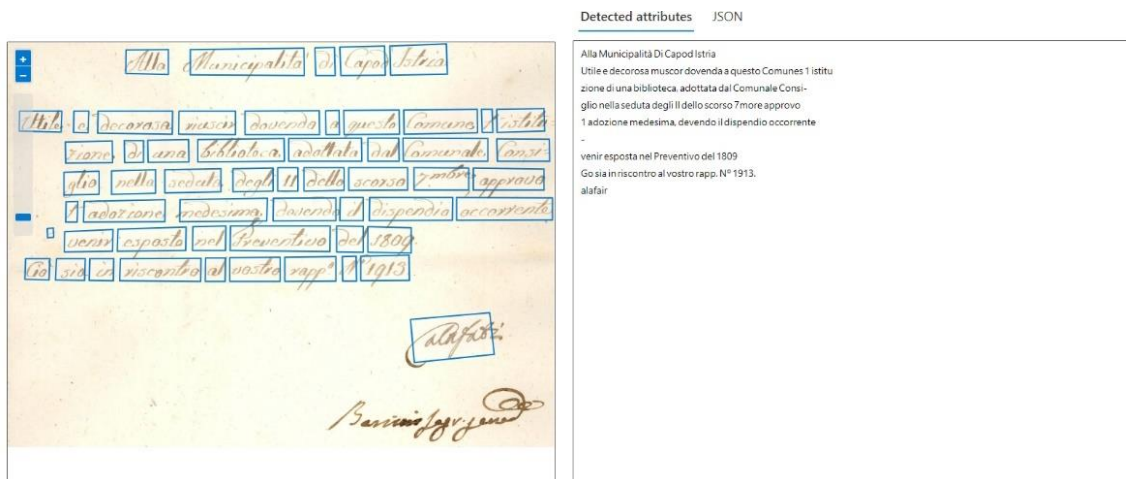
Orodje *Microsoft Azure's Computer Vision* združuje več storitev v eno platformo:

- optično zaznavanje znakov,
- analiza slik,
- zaznavanje, prepoznavanje in analiza človeškega obraza,
- prostorska analiza.

Omogočeno je tudi prepoznavanje z roko napisanega besedila, ki je del storitve *Analyze Image API*. Orodje zazna vsako besedo posebej. Rezultat je datoteka JSON, ki poleg besedila vsebuje tudi informacijo o dolžini besede in koeficient prepričanosti orodja glede natančnosti prepisa (angl. Confidence). Za zaznavanje besedila orodje ne potrebuje informacij o jeziku.

Za uporabo in testiranje Microsoftovega orodja za prepoznavanje z roko napisanega besedila je potrebno ustvariti uporabniški račun na platformi Microsoft Azure in ustvariti ustrezen resurs⁹. Za potrebe prispevka je uporabljeno okolje Vision Studio, ki omogoča testiranje storitve brez pisanja programske kode. Rezultat (slika 5) je prikazan kot besedilo ali podrobnejša datoteka JSON.

⁹ Osnovna entiteta, ki je del Microsoft Azure platforme.



Slika 5: Prepoznavanje besedila s pomočjo orodja Microsoft Azure's Computer Vision

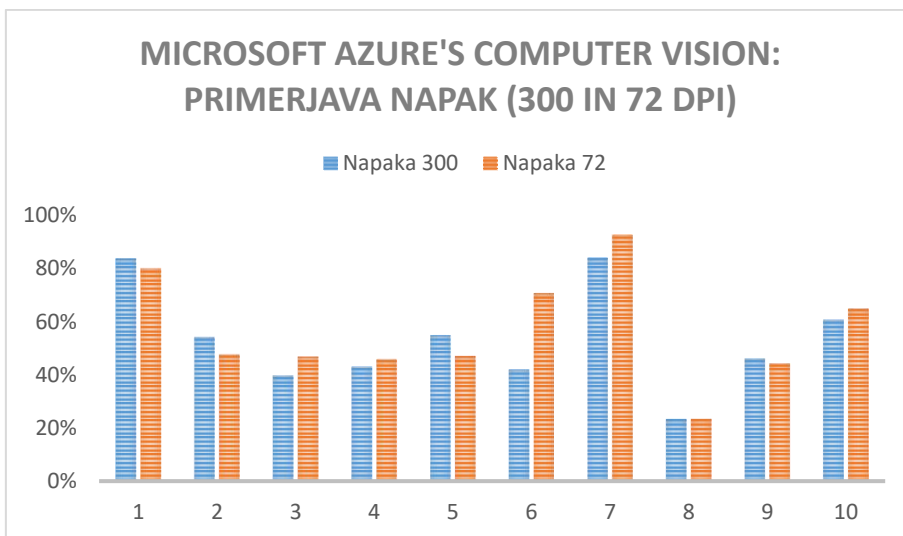
Rezultati analize

Rezultati analize transkribiranja z orodjem Microsoft Azure's Computer Vision so prikazani v tabeli 4. Povprečna napaka za 20 transkribiranih primerov je 55 %, minimalna 24 %, maksimalna 93 %.

Tabela 4: Rezultati analize transkribiranja z orodjem Microsoft Azure's Computer Vision

Datoteka	Jezik	300 DPI					72 DPI				
		Število besed v izvorniku	Zaznano besed	Točno besed	Narobe besed	Napaka	Število besed v izvorniku	Zaznano besed	Točno besed	Narobe besed	Napaka
1	ITA	130	143	34	109	84 %	130	140	36	104	80 %
2	ITA	149	149	68	81	54 %	149	143	72	71	48 %
3	ITA	416	422	256	166	40 %	416	442	247	195	47 %
4	LAT	216	215	122	93	43 %	216	216	117	99	46 %
5	ITA	51	56	28	28	55 %	51	55	31	24	47 %
6	ITA	157	151	85	66	42 %	157	203	92	111	71 %
7	LAT	95	88	8	80	84 %	95	96	8	88	93 %
8	LAT	34	34	26	8	24 %	34	34	26	8	24 %
9	SLO	106	109	60	49	46 %	106	108	61	47	44 %
10	SLO	74	73	28	45	61 %	74	73	25	48	65 %

Povprečna napaka za slike v 300 DPI je 53 %, 56 % pa za slike v 72 DPI. Podrobna analiza primerjave napak med ločljivosti dokumentov je prikazana na sliki 6.



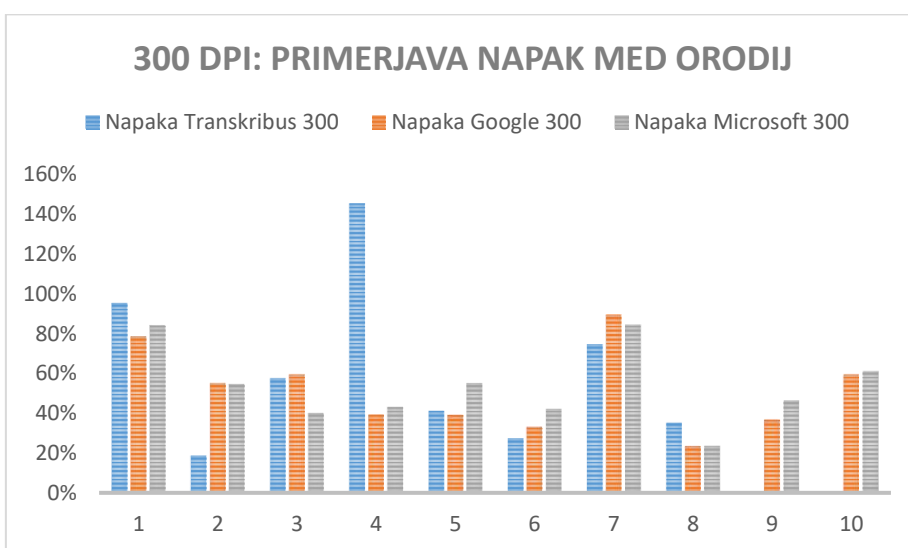
Slika 6: Primerjava napak pri 300 DPI in 72 DPI orodja Microsoft Azure's Computer Vision

6 PRIMERJAVA

Za potrebe prispevka je bilo opravljenih šestdeset transkripcij v ločljivosti 300 DPI in 72 DPI, s katerimi je bila narejena primerjava med orodji. Za izračun povprečnih vrednosti so uporabljeni rezultati za prvih osem dokumentov, saj Transkribus nima ustreznega modela za slovenski jezik.

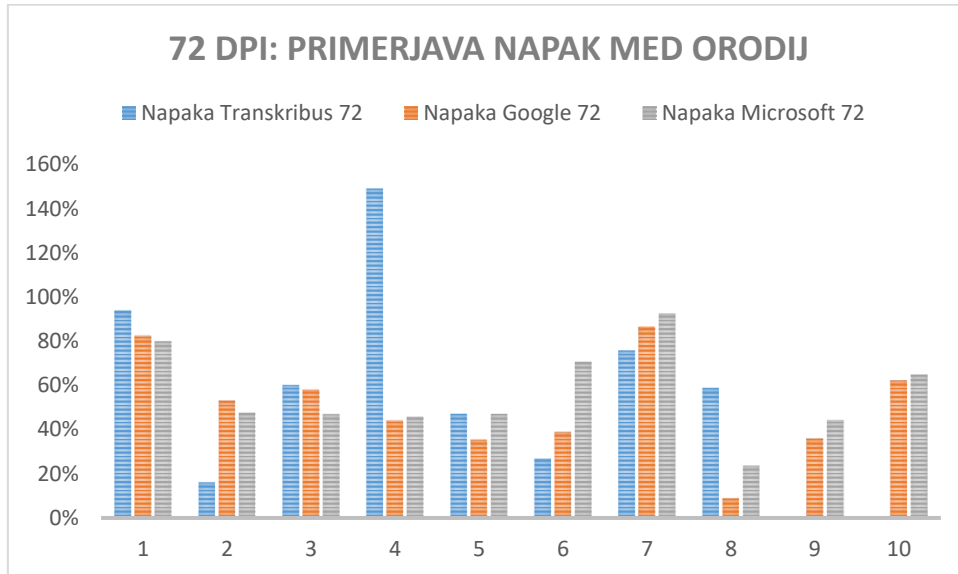
Slika 7 prikazuje primerjavo napak med orodji za dokumente v 300 DPI. Iz grafikona lahko razberemo, da Google Cloud Vision AI in Microsoft Azure's Computer Vision delujeta na podoben način.

Povprečna napaka Transkribusa za slike v 300 DPI je bila 62 %, orodja Google Cloud Vision AI 52 %, platforme Microsoft Azure's Computer Vision pa 53 %, kar pomeni, da so orodja zaznala narobe več kot polovico besed.



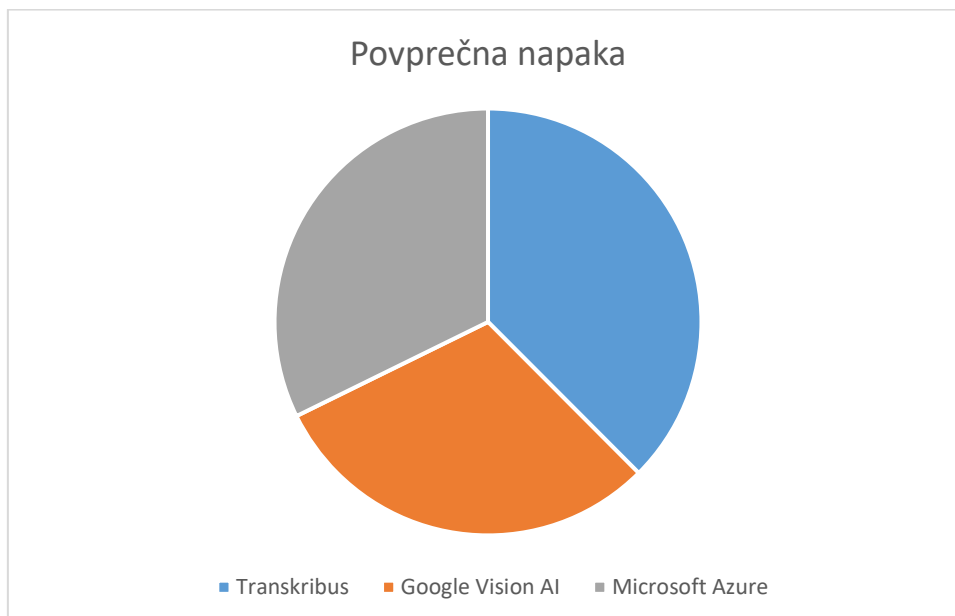
Slika 7: Primerjava zaznanih napak vseh orodij pri 300 DPI

Pri ločljivosti 72 DPI (slika 8) je bil odziv orodij Transkribus in Microsoft Azure's Computer Vision v primerjavi z orodjem Google Cloud Vision AI slabši: Transkribus je imel za 66 % napak, Microsoft Azure's Computer Vision 57 %, Google Cloud Vision AI pa 51 %.



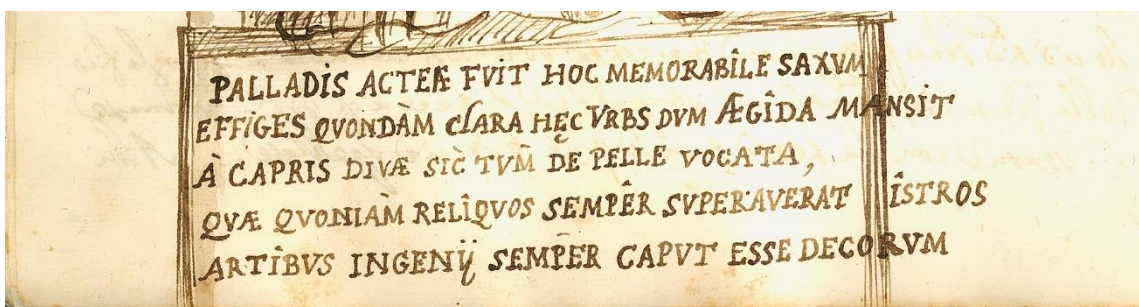
Slika 8: Primerjava zaznanih napak vseh orodij pri 72 DPI

Povprečna napaka Transkribusa za 16 transkribiranih dokumentov je bila 64 %, orodja Google Cloud Vision AI 51 % in orodja Microsoft Azure's Computer Vision 55 % (slika 9). Najboljši rezultat je pokazalo orodje Google Cloud Vision AI.



Slika 9: Primerjava povprečne napake med orodji

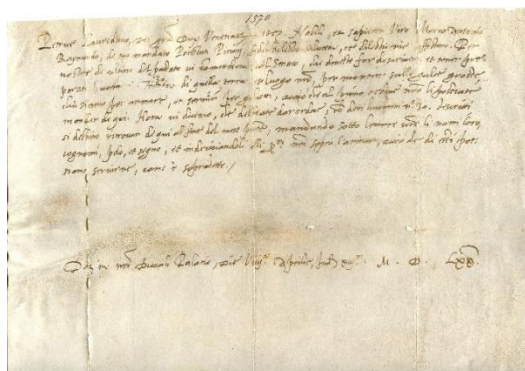
Izpostavili bi dokument št. 8 (slika 10), ker gre za primer besedila, napisanega v latinici, ki je v osnovi razumljiv za branje širši množici uporabnikov. Pojavi se vprašanje, ali je za tovrstne dokumente smiselno uporabljati umetno inteligenco. Transkribus je pri transkripciji tega dokumenta imel precej težav. Zaradi neustreznosti modela ali lastnosti skeniranega dokumenta (barva, svetloba in nasičenost) orodje ni prepoznalo besedila v ločljivosti 72 DPI. Isti dokument smo preizkusili in analizirali z različnimi barvnimi lastnostmi, vendar pri rezultatih ni bilo razlike. Dokument smo predstavili v model »Transkribus Print M1«, saj je le-ta prikazal ustrezne rezultate.



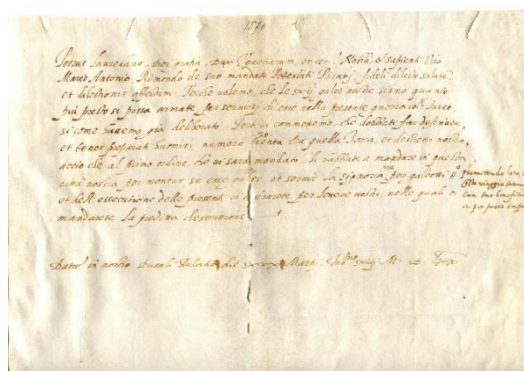
Slika 10: Prospero Petronio, Brani della parte prima delle Memorie sacre e profane dell'Istria 1680-1681

7 Primerjava dukalov

Dukali ali t. i. doževa pisma so dokumenti, ki se delijo na kolektivne in individualne. Njihova vsebina se nanaša na specifičen ukaz, ki ga je dož poslal naslovnikom v zvezi s točno določenim problemom, ki ga je bilo treba takoj reševati. »Druga vrsta dukalov predstavljajo trajne zakone, ki so dopolnjevali mestne statute, s katerimi so uravnavali delo in odnose med državo in lokalno samoupravo« (Furlan, 1978, str. 3). V obeh primerih (slika 11 in slika 12) je razvidno, da gre za doževu diktiranje istemu naslovniku, v tem primeru piranskemu podestatu Antoniu Rimondiju, saj ga pošiljatelj naslavlja na enak način: »Pietro Loredan, po milosti božji dož Benetk, plemenitemu in učenemu možu Marku Antoniju Rimondiju, po njegovem ukazu podestatu v Piranu, zvestemu, ljubljenemu pozdrav in ljubeznivo naklonjenost« (Piranski dukalet, 8. 4. 1570); Piranski dukalet (30. 3. 1570). Podobnost vsebine izbranih dukaletov omogoča poleg načrtovane analize orodji za transkribiranje besedila tudi primerjavo delovanja le-teh pri dokumentih s podobno ali enako vsebinsko strukturo.



Slika 11: Piranski dukalet z dne 8. 4. 1570



Slika 12: Piranski dukalet z dne 30. 3. 1570

Primerjava podobnih vsebinskih dokumentov je omogočila dodatno primerjalno analizo vseh treh orodij. Rezultati so pokazali, da podobnost vsebine, slog in način pisanja ne predstavlja pogoja za ustrezen prepis (tabela 5).

Tabela 5: Primerjava rezultatov delovanja orodij pri dveh dukalih s podobno vsebinsko strukturo

Datoteka	300DPI			72DPI		
	Transkribus	Google Cloud Vision AI	Microsoft Azure's Computer Vision	Transkribus	Google Cloud Vision AI	Microsoft Azure's Computer Vision
Piranski dukal z dne 8. 4. 1570	95 %	78 %	84 %	94 %	82 %	80 %
Piranski dukal z dne 30. 3. 1570	19 %	55 %	54 %	16 %	53 %	48 %

8 ZAKLJUČEK

Arhivi stremijo k uporabi arhivskega gradiva v elektronski različici. Digitalizacija arhivskega gradiva pripomore k lažji, hitrejši in preglednejši uporabi gradiva tako zaposlenih kot uporabnikov arhivov in bo nadomestila uporabo izvornikov ter posledično zaščitila originale pred zunanjimi vplivi. Z uporabo tehnologije HTR, ki temelji na prepoznavanju z roko napisanega besedila, je namen približati arhivsko gradivo širši množici uporabnikov.

Tehnologija za prepoznavanje pisav se je v zadnjih letih razvila v obliki številnih računalniških orodij (Transkribus, Google Cloud Vision AI, Microsoft Azure's Computer Vision, Amazon Textract ...). Za orodje Transkribus omenja Glavič (2020), da ne nadomesti človeka – strokovnjaka, lahko pa s pravilnim pristopom in procesom znatno pohitri transkribiranje večje količine gradiva. Enaka trditev velja tudi za Google Cloud Vision AI in Microsoft Azure's Computer Vision. Povprečna napaka za vsa tri orodja je 57 %, kar ne pomeni, da transkribirana besedila niso berljiva, kajti če je npr. v neki besedi napačna zgolj ena črka, se to šteje kot napaka celotne besede.

Ena od prednosti tovrstnih orodjih predstavlja spreminjanje pisav iz različnih zgodovinskih obdobj v nam razumljivo različico, ki uporabniku poenostavi branje različnih pisav.

Orodje Transkribus je od svojega začetka zanimivo raziskovalcem in zgodovinarjem, kar je razvidno v številnih prispevkih. Tega pa ne moremo trditi za Google Vision AI in Microsoft Azure. Zaradi pomanjkanja strokovnih člankov za preostala orodja so potrebne nadaljnje raziskave, ki bi pripomogle k ustrežnejši primerjavi.

9 VIRI IN LITERATURA

- Bonin, Z. (2006).** *Vodnik po fondih in zbirkah Pokrajinskega arhiva Koper*. Koper: Pokrajinski arhiv.
- Bonin, Z. in Rogoznica, D. (2010).** *Koprška pisna dediščina: od hrambe javnih dokumentov do ureditve zgodovinskega arhiva*. Koper: Pokrajinski arhiv.
- Darovec, D. (2009).** *Kratka zgodovina Istre*. Koper: Univerza na Primorskem, Znanstveno-raziskovalno središče, Založba Annales.
- Dokument, s katerim je istrski perfekt Angelo Calafati leta 1808 odobril ustanovitev kopske knjižnice z dne 2. 9. 1808, Občina Koper, t. e. 4, Pokrajinski arhiv Koper.
- Furlan, M. (1978).** *Piranski dukali: katalog k razstavi Pokrajinskega arhiva Koper (Vol. 5, p. 19)*. Koper: Pokrajinski arhiv.
- Glavič, J. (2020).** Primer uporabe programa Transkribus in izdelava modela za avtomatsko optično prepoznavanje znakov za poenostavljeno transkribiranje ročno pisane gotice. *Moderna Arhivistika: časopis arhivske teorije in prakse*. Letnik 3, številka 1, str. 86–97. Maribor, Pokrajinski arhiv Maribor. DOI: <https://doi.org/10.54356/MA/2020/1/ECAD5603>.
- Google Cloud. (b. d.)** Pridobljeno 24. 4. 2023 s spletne strani: <https://cloud.google.com/docs>.
- Muehlberger, G. 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. 954-976. Emerald Publishing Limited. DOI: <https://doi.org/10.1108/JD-07-2018-0114>.
- Patel, H. P. (2019).** Archival Document Processing Using Cognitive Computing. *ProQuest Dissertations & Theses Global*. Pridobljeno 16. 1. 2023 s spletne strani: <https://www.proquest.com/dissertations-theses/archival-document-processing-using-cognitive/docview/2353604169/se-2>.
- Piranski dukal z dne 30. 3. 1570, Beneški dukali, št. 125, Pokrajinski arhiv Koper.
- Piranski dukal z dne 8. 4. 1570, Beneški dukali, št. 126, Pokrajinski arhiv Koper.
- Prospero Petronio, Brani della parte prima delle Memorie sacre e profane dell'Istria 1680-1681, Rodbina Gravisi, t. e. 32, Pokrajinski arhiv Koper.
- READ – COOP (b. d.)**. Pridobljeno 24. 4. 2023 s spletne strani: <https://readcoop.eu/about/>.
- Šolska kronika državne osnovne šole v Gabrovici na Krasu za šolsko leto 1953/54 z dne 8. 10. 1953, Osnovna šola Gabrovica, t. e. 2, Pokrajinski arhiv Koper.

SUMMARY

OPERATIONAL USEFULLNESS OF HANDWRITTEN TEXT RECOGNITION TOOLS

Jan COTIČ

Archivist, Regional Archives Koper, Slovenia
jan.cotic@arhiv-koper.si

Ivančica SABADIN

Information Scientist, Pokrajinski arhiv Koper, Slovenija
ivancica.sabadin@arhiv-koper.si

In modern archival science, more and more problems are solved using new technologies. Nowadays, archivists, scientists, and scholars are researching the usage of Blockchain, the Internet of Things, Crowdsourcing, and especially Artificial Intelligence in archival science. Artificial intelligence can be used for archival descriptions, metadata processing, classification, and converting archival documents...

One of the usages of artificial intelligence in archival science is Handwritten Text Recognition (HTR). Handwritten text recognition operates based on artificial intelligence to enable the creation of new text from digitalized documents. In the recent years, there has been more and more research in this field because this technology could facilitate the availability and comprehension of old documents.

There are many challenges to the usage of HTR technology. Historical documents are written in different writing styles and are often degraded. In Slovenian archives, papers written in different languages (Slovenian, Latin, Italian) and scripts (Curialis, Caroline, Gothic ...) are preserved. The aim using such tools is the creation of text that could be understood more quickly, more available, easier to use, and in which full-text search could be used. Also, we must not forget the importance of the digitalization of historical documents so we could prevent future degradation.

Through analysis of Transkribus, Microsoft Azure's Computer Vision, and Google Cloud's Vision AI, authors will present the process and results of applying Handwritten Text Recognition technology with digitalized materials of Regional Archives Koper. Since the tools are using different technologies to recognize text from an image, authors will compare their accuracy on the different qualities of images. Transkribus uses deep learning technology, while Microsoft Azure's Computer Vision and Google Cloud's Vision AI use cognitive computing.

Used scientific methods are the description of tools, analysis of the tool's operation, comparison of tools characteristics, and statistical presentation of tool comparisons results.