Problem prepoznavanja znakova iz starijih ruskih knjiga tijekom procesa digitalizacije na primjeru Gramatike M. V. Lomonosova

Cencelj, Ivana

Sveučilište u Zagrebu

Filozofski fakultet u Zagrebu

Odsjek za informacijske i komunikacijske znanosti

Katedra za arhivistiku i dokumentalistiku

Odsjek za istočnoslavenske jezike i književnosti

Katedra za ruski jezik

Ak. god. 2018./ 2019.

Ivana Cencelj

Problem prepoznavanja znakova iz starijih ruskih knjiga tijekom procesa digitalizacije na primjeru Gramatike M. V. Lomonosova

Diplomski rad

Mentor: red. prof. dr. sc. Hrvoje Stančić

Neposredni voditelj: dr. sc. Jozo Ivanović, v. arhivist

Mentorica: izv. prof. dr. sc. Željka Čelić

Zagreb, travanj 2019.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

_____

(potpis)

Sadržaj

1. Uvod

U današnjem modernom dobu informacijsko-komunikacijska tehnologija pojavljuje se u svim aspektima ljudske djelatnosti. Proizvodi suvremene informacijske tehnologije sve se češće koriste u društvenim znanostima. Zato su se i arhivi, knjižnice i muzeji morali prilagoditi zahtjevima svojih korisnika, te modernizirati sadržaj svojih ustanova. Od osobite je važnosti da se sve tiskane knjige digitaliziraju, ne samo kako bi bile dostupne široj publici, već i kako bi se sačuvale od propadanja. To je posebno važno za stare dokumente i knjige koje su se tijekom godina istrošile i oštetile, te kojima prijeti opasnost od uništenja.

Budući da sve institucije imaju ograničen budžet i vremenski rok za digitaliziranje materijala, logično je da će se umjesto prepisivanja tekstova iz knjiga, koristiti programima za optičko prepoznavanje teksta. No iako se nude mnogi napredni programi za OCR (optičko prepoznavanje znakova), još uvijek dolazi do određenih grešaka, osobito kod digitalizacije starih knjiga, pisanih zastarjelim pismima.

U prvom dijelu ovog rada opisat će se postupak digitalizacije i detaljnije proučiti svaka njegova faza zasebno. Nakon toga, objasnit će se postupak optičkog prepoznavanja znakova, zajedno s opisom povijesti njegova razvoja, načinima upotrebe te fazama kroz koje OCR program prolazi. Nabrojat će se i opisati pogreške koje se najčešće pojavljuju prilikom OCR-a, a posebna će se pozornost dati starim povijesnim knjigama.

U zadnjem dijelu rada provest će se istraživanje, u kojem će se knjiga *Ruska gramatika* (1755) Mihaila Vasiljeviča Lomonosova provući kroz dva različita programa za optičko prepoznavanje znakova: Abbyy Finereader i Transkribus. Dobiveni rezultati proučit će se i međusobno usporediti. Uz to će se spomenuti i dostupnost ruskih knjiga na Internetu. Na samome kraju dat će se zaključak provedenog istraživanja s preporukama za digitalizaciju i obradu starijih knjiga.

2. Digitalizacija

Iako su dokumenti i knjige u papirnatom obliku i dalje u velikoj mjeri zastupljeni u svim ljudskim djelatnostima, umjesto papira i knjiga sve se više koristi računalo i dokumenti u digitalnom obliku. Kako bi se već postojeće gradivo sačuvalo od propadanja, ali i omogućilo korištenje njegovih digitalnih verzija velikom broju ljudi na različitim mjestima, potrebno ga je digitalizirati. U hrvatskoj se enciklopediji digitalizacija definira kao „pretvorba teksta, slike, zvuka, pokretnih slika (filmova i videa) ili trodimenzijskog oblika nekog objekta u digitalni oblik, u pravilu binaran kod zapisan kao računalna datoteka sa sažimanjem podataka ili bez sažimanja podataka, koji se može obrađivati, pohranjivati ili prenositi računalima i računalnim sustavima“ (Hrvatska enciklopedija).

Puno je lakše pretraživati digitalno gradivo, pa ćemo tako neki pojam puno brže i jednostavnije pronaći u digitalnom rječniku, nego listajući analogni rječnik. Digitalni rječnik zauzima minimalno memorije, za razliku od analognog, koji često zauzima puno mjesta na polici, jer ima mnogo stranica. Digitalno gradivo je i fleksibilno. Ono se može kopirati i printati neograničen broj puta i time ga nećemo trošiti ili uništavati, a dobivene kopije bit će identične originalu. Za razliku od digitalnog gradiva, papir, film ili magnetska vrpca se svakim korištenjem ili kopiranjem troši. Osim zaštite originala, povećanja njegove dostupnosti i pojednostavljivanja njegova korištenja, razlozi digitalizacije mogu biti i stvaranje novih ponuda različitih ustanova i popunjavanje fonda, u kojem su dijelovi uništeni ili nestali. Istraživači koji se nalaze u različitim dijelovima svijeta tako mogu skupiti materijale koje imaju i objediniti ih u digitalnoj zbirci koja onda može služiti za njihova daljnja istraživanja ili istraživanja korisnika kojima je omogućen jednostavan pristup toj zbirci. U budućnosti možemo očekivati da će se papir sve manje koristiti, što zbog financijskih, a što zbog ekoloških razloga. Većina gradiva će tako odmah nastati u digitalnom obliku, te se neće ni ispisivati, pa će tada biti potrebno samo održavati to gradivo.

2.1. Faze digitalizacije

Hrvoje Stančić, u svojoj knjizi *Digitalizacija*, proces digitalizacije podijelio je u 7 zasebnih faza:

1. Odabir gradiva za digitalizaciju,
2. Digitalizacija gradiva,

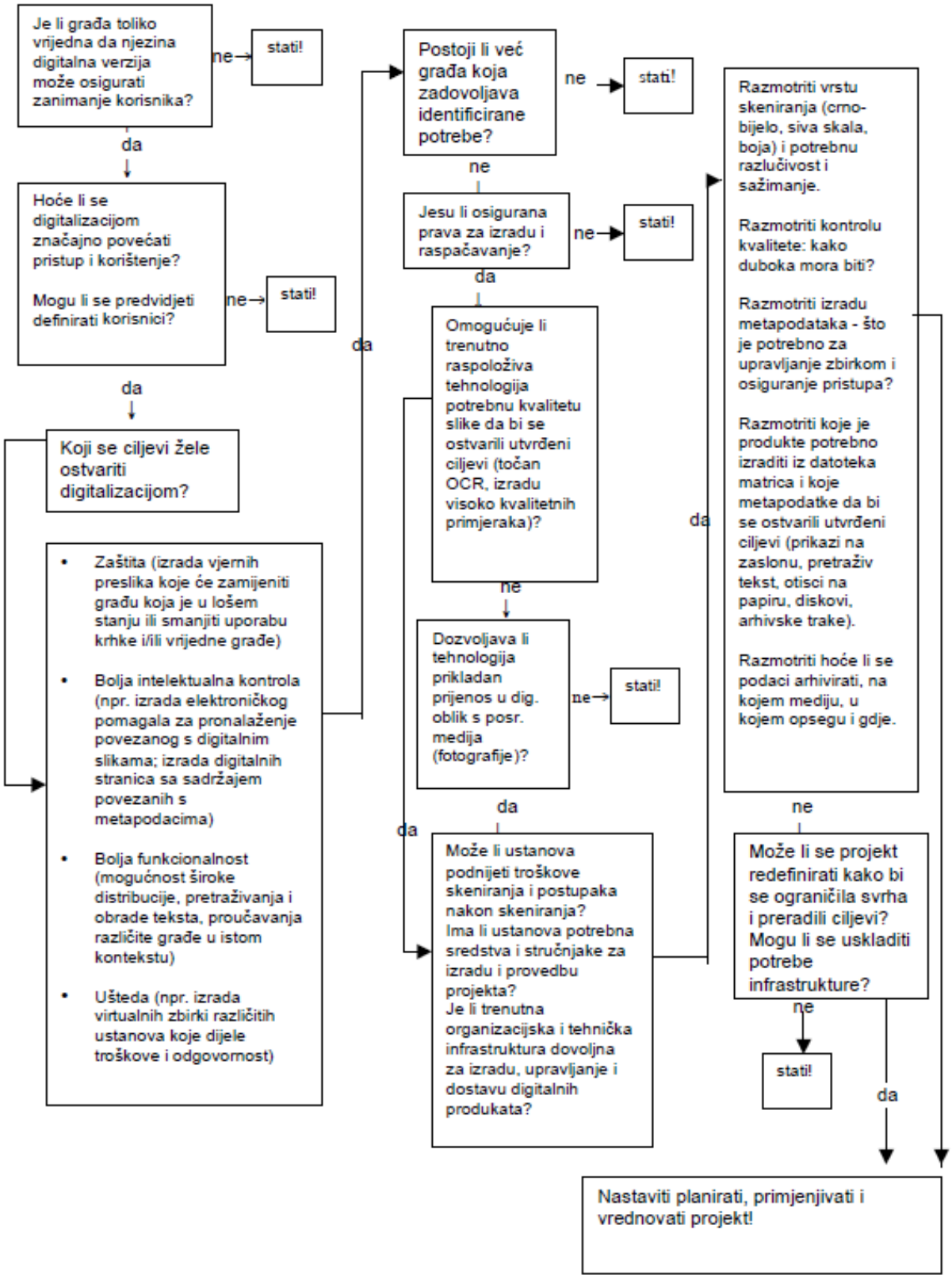
3. Obrada i kontrola kvalitete,
4. Zaštita gradiva u elektroničkoj okolini,
5. Pohrana i prijenos digitalnog gradiva,
6. Pregled i korištenje digitalnog gradiva,
7. Održavanje digitalnog gradiva (Stančić 2009: 7).

Svaka od navedenih faza detaljnije će se opisati u zasebnom poglavlju.

2.1.1. Odabir gradiva za digitalizaciju

Na samom početku procesa digitalizacije potrebno je razraditi projektni plan i jasno odrediti ciljeve digitalizacije. Prilikom pisanja projektnog plana, važno je okupiti sve sudionike projekta i postaviti određena pitanja. Potrebno je razmisliti o potencijalnoj koristi za korisnike, upravitelje zbirki i institucije, odrediti kada je pravo vrijeme da se započne s digitalizacijom, odrediti razumni budžet i rok do kada bi projekt trebao biti gotov, te odlučiti je li digitalizaciju bolje provoditi izvan ili unutar institucije (Northeast Document Conservation Center). Na odabir mjesta digitalizacije utječu faktori, kao što su iznos predviđenih financija za projekt digitalizacije, vremenski rok do kada bi projekt trebao biti gotov, posjedovanje potrebne opreme i stručnjaka. Prije same digitalizacije, određuje se grupa stručnjaka koja će odrediti kriterije za odabir gradiva za digitalizaciju na temelju različitih priručnika i smjernica. Ponekad nije potrebno sačuvati sve gradivo, već samo ono koje ima dugoročnu vrijednost. Primjerice, popisi dolazaka studenata na predavanja profesoru su potrebni samo za vrijeme trajanja tog semestra pa se oni neće digitalizirati i čuvati za kasnije, dok završni i diplomski radovi imaju dugoročnu vrijednost, pa će se oni digitalizirati i kasnije dati na korištenje. Tijekom odabira gradiva za digitalizaciju, ono se prvo predlaže, zatim procjenjuje i na kraju se određuju prioriteti. Kod odabira gradiva, potrebno je postaviti određena pitanja koja će pomoći kod određivanja koje je gradivo potrebno najprije digitalizirati. Najvažniju ulogu ovdje imaju vrijednost gradiva i fizičko stanje u kojem se ono nalazi. Dakle gradivo koje bi se najprije trebalo digitalizirati je ono koje ima visoku vrijednost te bi se zato i često koristilo, ali je i u jako lošem stanju, pa je digitalizaciju potrebno što prije provesti (Stančić 2009: 15-32).

Bilo bi dobro kada bi sve institucije s jednakom ulogom, primjerice, sve knjižnice ili svi arhivi, zajedno surađivale u procesu digitalizacije, kako se isti materijal ne bi nepotrebno digitalizirao više puta i kako bi se stvorila zajednička digitalna zbirka sa svim potrebnim istraživačkim izvorima (Smith 1999: 9).



Slika 1. Matrica za donošenje odluka prilikom odabira građe za digitalizaciju

Izvor: Smjernice za odabir građe za digitalizaciju

2.1.2. Digitalizacija gradiva

Kako će se izvoditi sama digitalizacija gradiva ovisi o vrsti tog gradiva, bilo ono tekstualno, slikovno, zvučno, video ili trodimenzionalno. Tekstualno gradivo može se digitalizirati i prepisivanjem, no taj je postupak veoma dugotrajan i skup i koristi se samo u slučajevima kada se trebaju digitalizirati stari rukopisi s požutjelim stranicama i izbijeljenim tekstom, često s bilješkama na marginama. Kada bi se takvi dokumenti skenirali, dobivena slika bila bi veoma loša i tradicionalno optičko prepoznavanje znakova ne bi bilo moguće, jer bi bilo prepuno grešaka. Prepisivanje se koristi i kada je potrebno dobiti potpuno točan i pretraživ prijepis nekog dokumenta.

U svim ostalim slučajevima, za digitalizaciju tekstualnog gradiva koriste se skeneri (snimači) i digitalni fotoaparati. Skener čita sliku i tako predstavlja „oko“ računala. On pretvara fizičke slike u slikovne datoteke koje računalo može obrađivati.[1] Postoje dvije vrste skenera s obzirom na tijek procesa skeniranja: koračni i protočni skeneri. Kod koračnih skenera potrebna je ljudska intervencija, jer se oni sami ne mogu pomicati po stranici koja se skenira i ne mogu sami okretati stranice. Protočni skeneri imaju uvlakač listova, pa mogu sami okretati stranice, te je skeniranje na takvim skenerima puno brže nego kod koračnih skenera. Kod odabira skenera potrebno je obratiti pozornost na njegovu brzinu, razlučivost, dinamički raspon i polje skeniranja, s obzirom na vrstu i veličinu gradiva koje želimo skenirati. Kada se radi o skeniranju par stranica brzina skenera nam ne igra preveliku ulogu, ali ako dokument koji se želi skenirati ima primjerice 400 stranica, brzina skenera bit će veoma važna. Dakako, što je brzina i kvaliteta skenera veća, to će i njegova cijena biti viša. Ako nam je cilj digitalizacije dobiti slike stranica teksta, dokument se može skenirati u boji, odnosno koristeći spektar sive boje, a dobivenim slikama potrebno je pridružiti metapodatke, kako bi bile pretražive. Kada se skenirani tekst kasnije želi provući kroz program za optičko prepoznavanje znakova, potrebno ga je skenirati u crno bijeloj tehnici ili u boji, u rezoluciji od minimalno 300 dpi[2], kako bi dobivena slika bila dobre kvalitete za kasniji OCR.

Za digitalizaciju slikovnog gradiva koriste se skeneri s visokom razlučivošću ili digitalni fotoaparat, a kako bi se dobio kvalitetan rezultat, potrebno je skenirati u visokoj razlučivosti:

[1] „Skeneri djeluju tako da puste svjetlost objekt ili dokument koji se digitalizira i usmjere reflektirano svjetlo (obično kroz niz zrcala i leća) na fotoosjetljivi element. U većini skenera, senzorski medij je elektronički integrirani krug osjetljiv na svjetlost – CCD senzor (engl. charged coupled device). Fotografije osjetljive na svjetlost raspoređene duž CCD senzora pretvaraju svjetlost u elektronske signale koji se zatim obrađuju u digitalnu sliku. (http://preservationtutorial.library.cornell.edu/technical/technicalB-02.html, 08. 04. 2019)

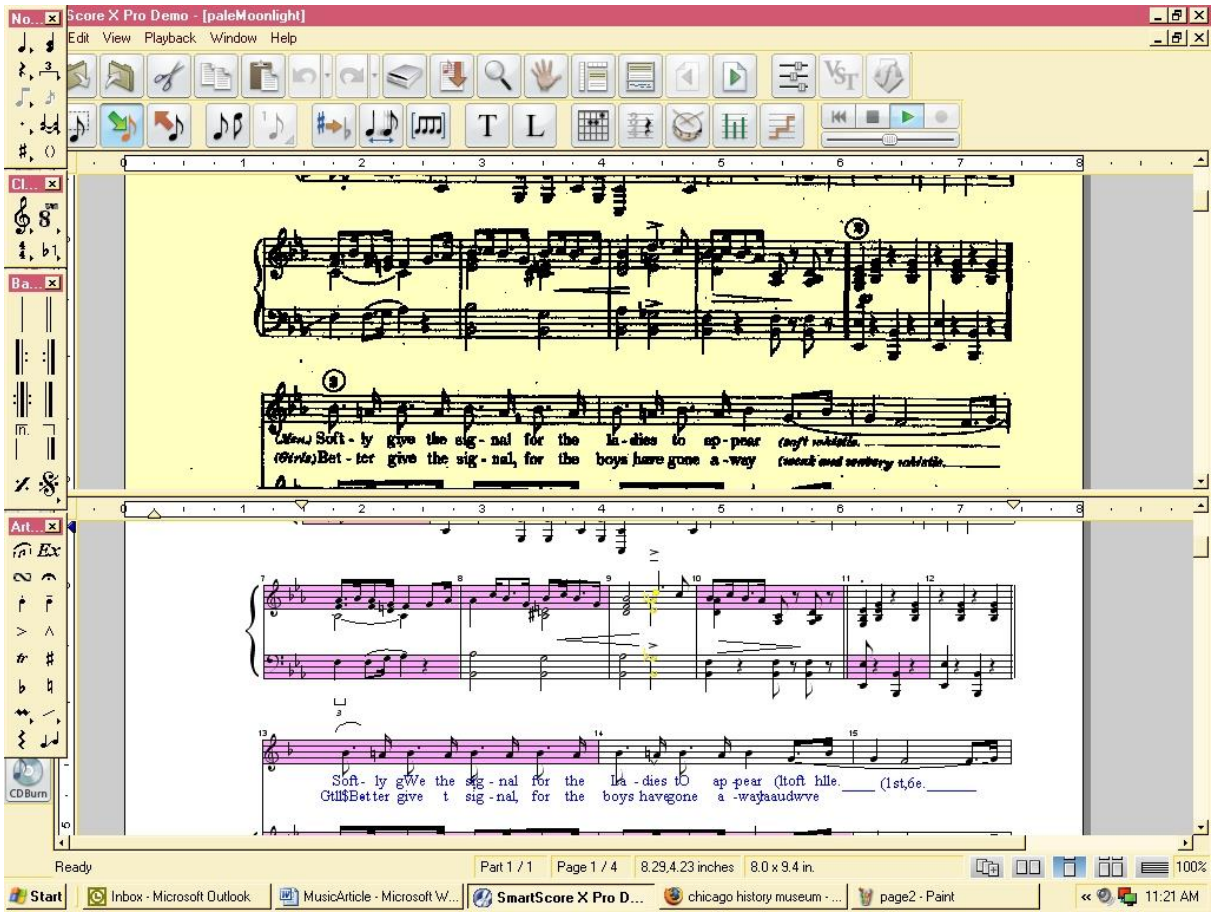
[2] dpi (engl. dots per inch)- broj točka po kvadratnom inču

za male slike koristi se min 600 dpi, a za velike slike 300 dpi. Razlučivost ne smije biti prevelika, jer onda zauzima previše memorije, te nije prikladna za slanje. Kvaliteta digitalne slike ovisi o rezoluciji, bitnoj dubini točke i boja. Preporuka je da se nakon digitalizacije izrade tri verzije digitalizirane slike: prva mora biti najsličnija originalu, u boji i visoke razlučivosti, bez korištenja komprimiranja, druga s korištenjem manje nijansi boji ili u spektru sivih tonova, komprimirana, i treća mala, identifikacijska, koja se onda koristi kao vizualna referenca ili veza na prethodne slike.

Za digitalizaciju zvučnog gradiva zvučni izlaz uređaja za reprodukciju audio sadržaja povezuje se s računalom koji u sebi sadržava zvučnu karticu i program za prihvat i obradu zvučnog signala. Digitalizacija zvuka sastoji se od uzorkovanja i kvantizacije. Digitalizacija video gradiva zapravo se sastoji od digitalizacije slike i zvuka jer je video zapravo niz brzo promijenjenih slika koje ljudsko oko prepoznaje kao neprekinuto gibanje, uz zvuk. Digitalni video zapis zauzima puno memorije, pa se zato nakon digitalizacije primjenjuje postupak komprimiranja, kako bi datoteka zauzimala manje mjesta. Za digitalizaciju 3D objekata najčešće se koriste posebni skeneri. Kod jednostavnijih objekata koriste se i obični skeneri ili fotoaparati, no tada se gubi plastičnost objekata (Stančić 2009: 33-70).

2.1.3. Obrada i kontrola kvalitete

U fazi obrade i kontrole kvalitete kod skeniranog tekstualnog gradiva tekst se provlači kroz program za optičko prepoznavanje teksta. Sam postupak OCR-a detaljnije će se objasniti u sljedećem poglavlju. Postoje i programi za prepoznavanje notnih zapisa – OMR programi (engl. Optical Music Recognition). Specifičnost notnog zapisa jest to da on ima dvije dimenzije zapisa: horizontalni, odnosno vremenski tijek i vertikalni, odnosno istovremena događanja, pa dolazi do problema kod razdvajanja objekata.



Slika 2. OMR

Izvor: https://journal.code4lib.org/articles/84

Nakon skeniranja slikovnog i zvučnog gradiva mora se provesti kontrola kvalitete dobivene digitalne slike, odnosno zapisa, te obraditi u nekom od programa za obradu slike ili zvuka (Stančić 2009: 71-94).

2.1.4. Zaštita gradiva u elektroničkoj okolini

Zaštita gradiva važna je jer štiti gradivo od neovlaštenog pristupa, korištenja, kopiranja i distribuiranja i dokazuje autentičnost gradiva. U tu se svrhu koriste razni mehanizmi za zaštitu sustava i samog gradiva. Kod zaštite sustava važno je upravljanje razinama pristupa, kako bi korisnicima bili dostupni samo javni podaci, a tajni ostali skriveni. Zato je bitno dobro čuvati sve lozinke, postaviti antivirusne programe i vatrozid (engl. firewall). Neke od metoda za zaštitu gradiva su šifriranje simetričnim ili javnim ključem, digitalni potpisi, digitalni certifikati i digitalni vodeni žigovi (Stančić 2009: 95-111).

2.1.5. Pohrana i prijenos digitalnog gradiva

Nakon digitalizacije digitalizirano gradivo mora biti svima dostupno, pa se zato pohrana i prijenos gradiva do korisnika zajedno proučavaju. Kada je riječ o odabiru sustava za pohranu, važno je obratiti pozornost na dugovječnost medija, trajnost medija, kapacitet, cijenu, prihvaćenost te vrstu sustava (izravan ili poluizravan) (Stančić 2009:113). Mora se i odrediti vrsta institucije, prema količini digitaliziranog gradiva koje stvara. Ovdje se ne radi samo o digitaliziranom gradivu koje nastaje nakon procesa digitalizacije, već i gradivu koje se odmah stvara u elektroničkom obliku. Sustavi za pohranu dijele se na izravne, poluizravne, hijerarhijske, neizravne, sustave za mrežnu pohranu i mreže za pohranu. Kod svih vrsta sustava obavezno je imati dvije odvojene sigurnosne kopije, pohranjene na dvije vrste medija, od kojih bi se jedna trebala čuvati na nekom drugom mjestu (Stančić 2009: 113-138).

2.1.6. Pregled i korištenje digitalnog gradiva

Tijekom procesa digitalizacije potrebno je predvidjeti i odrediti na koje će se sve načine digitalizirano gradivo pregledavati i koristiti. Hoće li korisnici gradivo moći samo pregledavati ili će se ono moći i ispisivati na pisačima, te hoće li gradivo biti dostupno samo lokalno ili i putem Interneta. Ako se gradivo može ispisivati, mora se obratiti pozornost na vrstu gradiva, je li tekstualno i slikovno, te njegova kvaliteta, s obzirom na to, hoće li se ispisivati u crno-bijeloj tehnici, sivoj, ili u boji (Stančić 2009: 139-140).

2.1.7. Održavanje digitalnog gradiva

Na kraju uvodnog dijela bitno je istaknuti da digitalizacija ne jamči trajnost digitaliziranog gradiva. Zato je važno da se sve digitalizirano gradivo održava, jer gradivo vrlo brzo zastarijeva i prestaje biti dostupno korisnicima, te u tom slučaju cijeli postupak digitalizacije nema smisla. Često je medij na kojemu se nalazi digitalno gradivo nestabilan, jer se brzo može pojaviti neki bolji i napredniji medij, pa će medij sa zapisanim gradivom zastarjeti i prestati se koristiti. Uzmimo kao primjer diskete (engl. floppy disk), koje su se nekada koristile za zapisivanje gradiva. Danas je jako teško pronaći računalo koje ima disketnu jedinicu za čitanje i pisanje disketa, pa gradivo koje je pohranjeno na nekoj disketi uopće ne možemo otvoriti i pročitati. Osim trajnosti medija, potrebno je paziti da su podaci uvijek kodirani u čitljivim formatima, te je zato potrebno neko vrijeme paralelno koristiti stariji i noviji softver prije potpunog prebacivanja na novi, iako tako može doći do malih promjena u datoteci. Održavanje je iznimno važno kod gradiva koje je izvorno nastalo u elektroničkom

obliku, jer ono nema svoj originalni analogni oblik, pa ga je nemoguće povratiti i ponovno digitalizirati (Stančić 2009: 141-157).

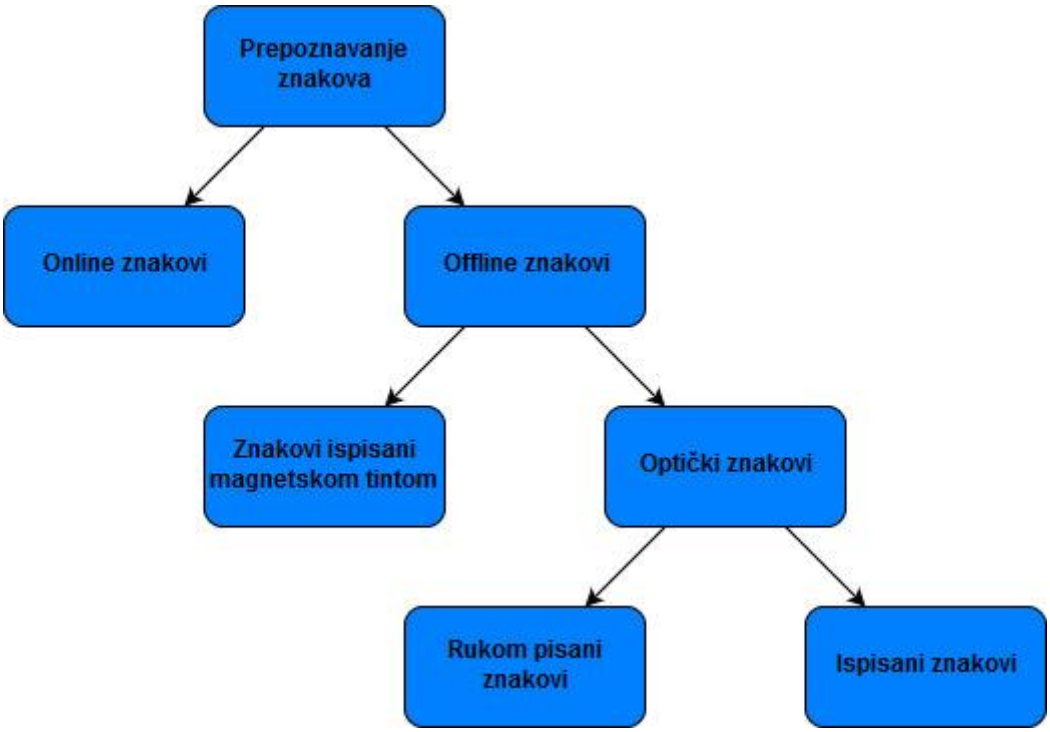
3. Optičko prepoznavanje znakova

Za razliku od ljudskog mozga koji može vrlo lako prepoznati i razlikovati tekst i znakove iz slike, samo računalo nije dovoljno sposobno da bi vidjelo i prepoznalo sve dostupne informacije na slici. Ono mora vrlo precizno raščlaniti sliku na sve njene komponente, kako bi ju uspješno prebacilo u odgovarajući programski kod u određenom programskom jeziku. Zbog toga moramo koristiti programe koji koriste posebnu tehnologiju za prepoznavanje znakova. Optičko prepoznavanje znakova (engl. OCR – Optical Character Recognition) je tehnologija, pomoću koje se rukopisi, tiskani ili printani tekstovi i dokumenti, snimljeni u digitalnom obliku, pretvaraju u tekstualne dokumente (npr. Word, ASCII, HTML), koji se mogu obrađivati i pretraživati. Digitalnu verziju koja nastane nakon OCR-a računalo može čitati bez potrebe ručnog unosa teksta. Budući da je poznato da je optičko prepoznavanje znakova nekoliko desetaka puta brže nego ručno tipkanje teksta, ušteda vremena kod velikih dokumenata ili knjiga je značajna.

Prepoznavanje znakova može se izvoditi online i offline. Dalbir i Singh (2015) online prepoznavanje znakova opisuju kao proces prepoznavanja rukom pisanog teksta u trenutku pisanja tog teksta snimljenog digitalizatorom koji prati pomicanje olovke. Rukopis se snima i pohranjuje u digitalnom obliku, te se za to najčešće koristi posebna olovka u kombinaciji s elektroničkom površinom. Prilikom pomicanja olovke, dvodimenzionalne koordinate uzastopnih točaka pohranjuju se. Online način prepoznavanje znakova kod rukopisa najčešće daje bolje rezultate nego offline način, budući da se prilikom online načina pohranjuju neke dodatne informacije, poput smjera i brzine, te broja i redoslijeda pokreta prilikom pisanja.

Optičko prepoznavanje znakova je vrsta offline prepoznavanja znakova. To znači da program skenira i prepoznaje statične slike znakova koji su ranije bili ispisani, bilo strojem ili rukom. Nakon skeniranja i prije samog prepoznavanja potrebna je i dodatna obrada kako bi rezultati bili što točniji. Kod optičkog prepoznavanja teksta, slika se dobiva pomoću optičkih sredstava, odnosno skenera i fotoaparata. Osim optičkog prepoznavanja znakova, postoji i magnetsko prepoznavanje znakova, odnosno MCR (engl. Magnetic Character Recognition) ili MICR (engl. Magnetic ink character recognition). Kod magnetskog prepoznavanja znakovi se ispisuju magnetskom tintom koje onda uređaj za čitanje prepoznaje zahvaljujući jedinstvenom

magnetskom polju svakog znaka. MCR se uglavnom koristi u bankama, kako bi se olakšala obrada i provjera autentičnosti čekova i drugih dokumenata.



Slika 3. Vrste prepoznavanja znakova

Prema: https://bit.ly/2ViQYQA

OCR programi međusobno se razlikuju prema cijeni, točnosti, brzini, značajkama i jezicima koje prepoznaju. Postoje komercijalni i besplatni OCR programi. Koji program ćemo odabrati ovisi o tome za što nam je potreban OCR i koliki nam je budžet i opseg gradiva koje treba digitalizirati. Recimo, ako neka knjižnica treba digitalizirati veću količinu knjiga u kratkom vremenskom roku i ima dovoljan budžet za to, odabrat će dobar i skup komercijalni program kako bi rezultat nakon OCR-a bio što točniji. Ali zato će privatni korisnik koji treba digitalizirati određenu knjigu prije odabrati besplatni program koji možda neće dati najbolje rezultate kao neki skupi program, pa će zato biti potrebno uložiti dodatno vrijeme na ručnom ispravljanju grešaka. Prema Nieldu, neki od danas najboljih komercijalnih OCR programa su:

1. Nuance OmniPage Ultimate,
2. Abbyy FineReader,
3. Adobe Acrobat Pro DC,
4. Readiris,

5. Google Drive.

Nuance OmniPage jedan je od prvih OCR programa koji se počeo koristiti na privatnim računalima. Danas se najčešće koristi u malim i velikim tvrtkama u kojima je potrebno obraditi velike količine papira. Neke od tvrtki koje ga koriste su: Coca-Cola, Microsoft, HP i Amazon. Unutar programa moguće je namjestiti automatsko slanje na unaprijed zadanu e-mail adresu, ili više njih, nakon što se tekst prepozna. Za manje tvrtke, kojima je *Ultimate* opcija preskupa, postoji i opcija *Standard* koja ima nižu cijenu, ali ne nudi jednak broj mogućnosti kao opcija *Ultimate*. Prepoznavanje je moguće na više od 120 jezika.

Abbyy FineReader jedan je od najpoznatijih OCR programa kojeg mnoge tvrtke koriste već više od dvadeset godina. Neki od njihovih korisnika su Samsung i Fujitsu. Osim što omogućuje prepoznavanje tekstova i prebacivanje u druge formate, tekstove je moguće i uspoređivati i komentirati. Posljednja 14. verzija podržava prepoznavanje tekstova sa 192 jezika, a za njih 48 postoji i ugrađena provjera pravopisa. Abbyy također nudi aplikacije za mobilne telefone.

Adobe Acrobat Pro DC je dobro rješenje za tvrtke koje već koriste Adobe alate, poput Photoshopa ili Adobe aplikacija. Adobe Acrobat DC sve dokumente pohranjuje u oblak (engl. cloud) kako bi bili dostupni sa svih računala unutar neke tvrtke, što označuje i kratica DC (engl. Document Cloud). Pro verzija omogućuje komentiranje i usporedbu različitih dokumenata, te specijalizirani alat za skeniranje tablica.

Readiris jedan je od najbržih OCR programa koji nudi veliki broj značajki. Unutar dokumenata moguće je staviti digitalne potpise, sigurnosne zaštite, vodene žigove i komentare. U slučaju da korisnik nije zadovoljan programom nakon mjesec dana korištenja, moguć je povrat novca. Verzija Readiris PDF 17 nudi prepoznavanje 38 jezika, dok ReadIris Pro 17 i Corporate nude prepoznavanje 138 jezika.

Pomoću Google Drive-a moguće je .JPEG, .PNG, .GIF i PDF datoteke pretvoriti u pretražive dokumente. Dokument mora biti točno orijentiran, biti u što većoj rezoluciji i njegova veličina ne smije biti veća od 2 MB, kako bi se dobio dobar rezultat. Također se preporučuje korištenje čestih fontova. Google Drive omogućuje prepoznavanje znakova s tekstova pisanih na čak 225 jezika. Google Drive nudi i aplikaciju za Android pametne telefone, u kojoj je moguće prepoznati tekstove direktno nakon fotografiranja mobitelom.

Većina boljih OCR programa prilično je skupa i za malog privatnog korisnika najčešće neisplativa, ali dosta njih pruža besplatan početni period korištenja, koji najčešće traju 7 ili 15 dana (engl. free trial). Tako je moguće i isprobati par programa, prije donošenja konačne odluke koji program odabrati.

Osim komercijalnih OCR programa, na Internetu je moguće pronaći velik broj besplatnih programa za optičko prepoznavanje znakova. Takvi programi često imaju mnoštvo grešaka i njihovi rezultati nisu uvijek precizni. Neki od besplatni programa su:

1. FreeOCR,
2. Microsoft Office Document Imaging (MODI),
3. Microsoft OneNote,
4. SimpleOCR.

FreeOCR je program koji je jednostavan i lak za korištenje, te je jedan od najtočnijih besplatnih programa. Program koristi Tesseractov OCR stroj kojim trenutačno upravlja Google. FreeOCR radi samo na Windowsima, a zadnja je verzija izdana 2015. godine. Nakon što se tekst prepozna, program ga prebaci u obradiv Word dokument. Za neke manje i jednostavne projekte ovaj program je odličan jer mu je brzina i točnost prilično dobra. Za neke kompliciranije projekte ipak nije najpogodniji jer je prejednostavan i ne nudi naknadnu obradu teksta, a često dolazi i do preklapanja linija i stupaca (Sharma 2017).

Još jedan program koji je moguće koristiti samo na operacijskom sustavu Microsoft Windows, je Microsoft Office Document Imaging. Još jedno od ograničenja je to da program može prepoznavati znakove unutar TIFF formata, pa ukoliko imamo neki drugi format, potrebno ga je konvertirati u TIFF. Ako koristimo verziju Worda iz 2010. godine ili stariju, MODI je već uključen u nju, a ako je verzija koju koristimo novija, potrebno je instalirati SharePoint Designer 2007. Korištenje je prilično jednostavno (Matthews 2017).

Microsoft OneNote je još jedan Microsoftov program unutar kojeg je moguće upotrijebiti OCR. Taj program podržava sve formate (PNG, JPG, BMP ili TIFF), a opcija OCR funkcionira jednostavnim odabirom funkcije Kopiranja teksta iz slike. No ipak postoje neka ograničenja, pa je tako nemoguće prepoznati znakove unutar tablice ili stupca. Zato se koristi za prepoznavanje slika s jednostavnim tekstualnim sadržajima (Sharma 2017).

SimpleOCR je program koji je nešto između besplatnih i komercijalnih programa. On je besplatan za prepoznavanje strojno pisanih tekstova, dok je za opciju prepoznavanje rukopisa

potrebno platiti. Moguće je preuzeti besplatnu verziju prepoznavanja rukopisa, no samo na 14 dana. Ovaj je program dostupan samo na Windows platformi. Unutar programa nalazi se alat za provjeru pravopisa i ispravljanje pogrešaka pri pretvaranju slike u tekstualni dokument (Ilindra 2018).

Iako se besplatni OCR programi mogu činiti kao najisplativija opcija, ponekad i nije baš tako. Budući da takvi programi često daju rezultate pune grešaka, morat ćemo uložiti dodatno vrijeme u njihovo ispravljanje, a kako ne nude korisničku podršku, neke stvari ćemo morati dugo tražiti na različitim forumima i za savjete pitati druge korisnike. Tako da ćemo kod nekih većih projekata ipak trebati uložiti više novca u bolji OCR program kako bi se ispoštovao zadani rok i dobio prihvatljiv rezultat.

Tablica 1. Usporedba različitih OCR programa s financijskog i operativnog aspekta (2019.)

Ime	Cijena	Operacijski sustavi
Google Drive	0-299,99 $	Windows, Mac OS X
Nuance OmniPage Ultimate	Standard 149,99 $ Ultimate 449,99 $	Windows, Mac OS X, Linux
Abbyy FineReader	Standard 199 € Corporate 299 €	Windows, Mac OS X, Linux, BSD
Adobe Acrobat Pro DC	Od 1.750 kn	Windows, Mac OS X
Readiris	PDF 49 $ Pro 99 $ Corporate 199 $	Windows, Mac OS X
FreeOCR	besplatan	Windows
Microsoft Office Document Imaging	besplatan	Windows
Microsoft OneNote	besplatan	Windows
SimpleOCR	besplatan	Windows

3.1. Povijest OCR-a

Prepoznavanje znakova pripada području prepoznavanja uzoraka, te su neke tehnike i pojmovi preuzeti iz prepoznavanja uzoraka i obrade slika. Međutim, upravo je prepoznavanje

znakova pomoglo da prepoznavanje uzoraka i analiza slika postanu zrelija područja znanosti i inženjerstva (Eikvil 1993).

Početke prepoznavanja znakova možemo pronaći još 1870. godine kada je Amerikanac C.R. Carey izumio retinalni skener. To je bio sustav koji je sadržavao mozaik fotoćeliju, te je služio za prijenos slika. Prve su verzije mogle raditi samo na jednom fontu odjednom. Sljedeći izum važan za optičko prepoznavanje teksta bio je Nipkow disk, koje je izumio poljsko-njemački inženjer 1884. godine. To je bio sekvencijski skener koji je pokazao mogućnost pretvorbe slike u električni signal (Britannica). Koristio se u prvim mehaničkim televizorima, te je bio važan za razvoj moderne televizije i strojeva za čitanje. Početkom 20. stoljeća istraživač A. M. Turing pokušao je napraviti stroj za pomoć slijepim i slabovidnim osobama, koji bi koristio OCR, no u tome nije uspio. Prve modernije verzije OCR-a pojavile su se tek 1940-ih kada su se razvila prva digitalna računala. Najraniji OCR sustavi nisu bili računala, nego mehanički uređaji koji su mogli prepoznati određene znakove, ali su bili veoma spori i imali su puno grešaka.

Elektronička obrada podataka postala je važno područje tijekom tehnološke revolucije pedesetih godina. Za unos podataka koristile su se bušene kartice, a količina podataka koje je trebalo obraditi, svakim je danom sve više rasla, pa je trebalo pronaći efikasno i financijski isplativo rješenje. M. Sheppard 1951. godine izumio je stroj koji je mogao čitati glazbene zapise. On je mogao prepoznati 23 znaka, te se smatra jednim od najranijih modernih OCR strojeva. Tehnologija za strojno čitanje je tijekom pedesetih godina 20. stoljeća dovoljno napredovala, pa su tako OCR strojevi tada postali i komercijalno dostupni. Prva komercijalna instalacija OCR sustava ostvarila se 1954. godine u tvrtki Reader's Digest u New Yorku. Taj je sustav služio za pretvaranje podataka o prodaji i narudžbama iz rukom pisanog oblika u bušene kartice koje su se onda unosile u odjelne računalne jedinice.

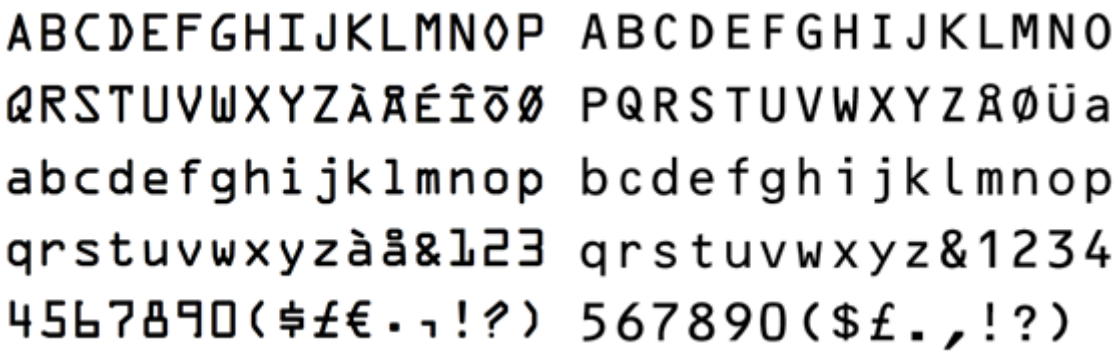
OCR strojevi dijele se u tri grupe:

1. Prva generacija: strojevi nastali od 1960. do 1965. godine,
2. Druga generacija: strojevi nastali od sredine 60-ih godina do početka 70-ih,
3. Treća generacija: strojevi nastali od sredine 70-ih godina do 1980. godine.

Strojevi prve generacije bili su dosta ograničeni i mogli su prepoznavati samo znakove određenih veličina, oblika i fonta. Znakovi su bili posebno dizajnirani za strojeve za čitanje, te nisu izgledali prepoznatljivo. S vremenom su strojevi mogli čitati i veći broj fontova, a broj

fontova bio je ograničen metodom prepoznavanja koja se koristila. Kod metode prepoznavanja uzoraka, stroj je uspoređivao sliku znaka s prototipskim slikama tog znaka u bazi određenog fonta.

Strojevi druge generacije su, osim strojno tiskanih znakova, mogli prepoznati i neke rukom pisane znakove: brojeve i par slova i simbola. Najpoznatiji stroj druge generacije bio je IBM 1287, koji je bio predstavljen 1965. godine na Svjetskom sajmu u New Yorku. Tijekom šezdesetih godina postalo je jasno kako postoji potreba da se provede standardizacija OCR fontova da bi se olakšao postupak prepoznavanja znakova. Tako je 1966. godine Američki nacionalni institut za standarde (engl. American National Standard Institute- ANSI) sastavio skup znakova nazvan OCR-A, dok je Europska udruga proizvođača računala (engl. European Computer Manufacturers Association – ECMA) sastavila svoj skup znakova nazvan OCR-B. Američki je font bio osmišljen kako bi olakšao optičko prepoznavanje, te je zato bio veoma stiliziran, dok je europski bio prirodniji (Eikvil 1993).



Slika 4. Usporedba OCR-A i OCR-B fonta

Izvor: http://www.identifont.com/differences?first=OCR-A&second=OCR-B+(BT)

Brzi razvoj hardvera (engl. hardware) doveo je do pada cijena OCR strojeva treće generacije i do poboljšavanja njihovih performansi. Iako su se tada pokušavali proizvesti napredni OCR strojevi koji bi mogli prepoznavati rukopise i dokumente lošije kvalitete, značajnu ulogu su i dalje imali jednostavniji strojevi. Oni su se koristili u kombinaciji s pisaćim strojevima jer se na njima koristila samo nekolicina fontova. Tekst bi se tipkao na stroju, te bi se tada ubacio u računalo pomoću OCR stroja, te bi se na računalu radile daljnje izmjene. Godine 1978. u prodaju je pušten prvi stroj za pretvaranje knjiga i drugih tiskanih materijala u sintetički govor. Tijekom 90-ih godina došlo je do još većeg napretka prepoznavanja znakova, zahvaljujući početku primjene neuronskih mreža, te razvoju novih alata i metoda. Istraživači

su razvili složene OCR algoritme, a obrada slike i prepoznavanje uzoraka uspješno su kombinirane s metodama umjetne inteligencije. Danas se koriste bitno razvijenije metode prepoznavanja znakova, razvoju kojih su doprinijeli moderniji i precizniji skeneri i fotoaparati. No još uvijek je ostalo mnogo mjesta za napredak, osobito u području prepoznavanja rukopisa.

3.2. Upotreba OCR-a

Intenzivni istraživački napor na području OCR-a nije bio uvjetovan samo primjenom u simulaciji ljudskog čitanja, već i zbog učinkovitosti svoje primjene u automatskoj obradi velikih količina papira, prijenosa podataka u strojeve i web sučelja u papirnate dokumente. OCR tehnologija primjenjuje se u različitim područjima: financijama, obrazovanju, zdravstvu, pravnim ustanovama i vladinim agencijama. Mi ju koristimo u svakodnevnom životu, a da često toga nismo niti svjesni.

Važna uporaba OCR-a u bankarstvu je kod obrade i provjere čekova i uplatnica. Ček ili uplatnica ubaci se u stroj u kojem sustav čita količinu novca koja se uplaćuje i prebacuje. U tom postupku nije potreban čovjek, te se time ubrzava cijeli postupak i samim time smanjuje čekanje u redu. Za tiskane čekove i uplatnice ta je tehnologija potpuno razvijena i vrlo rijetko nastaju greške, no i za rukom ispunjene čekove i uplatnice u načelu je točna.

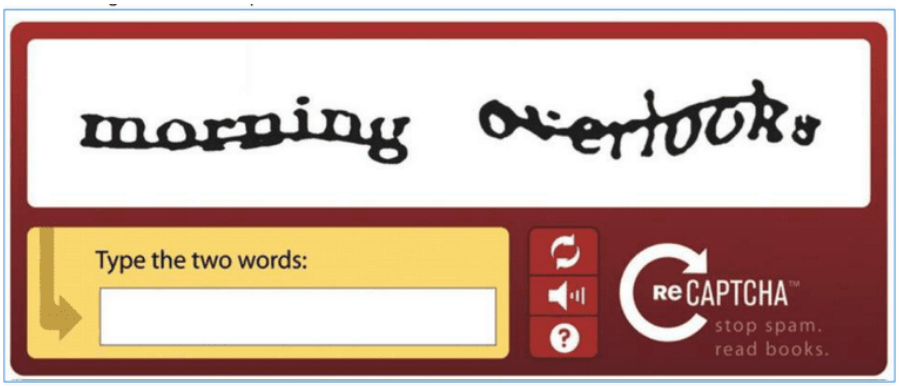
U obrazovanju OCR tehnologija može se koristiti kod obrade velike količine ispitnog materijala. Ta je tehnologija najlakše primjenjiva kod ispita s pitanjima s ponuđenim odgovorima, gdje ispitanik mora zacrniti polje kraj točnog odgovora. Osim toga, OCR se primjenjuje i kod izrade digitalnih repozitorija na fakultetima i drugim obrazovnim ustanovama. To su zbirke sastavljene od knjiga, monografija, članaka, zbornika, istraživanja, teza, disertacija i prezentacija. One se sastavljaju kako bi se svi materijali prikupili na jednom mjestu, te kako bi bili široko dostupni svima kojima su potrebni.

U zdravstvu se OCR tehnologija primjenjuje kako bi se riješio, ili barem smanjio, problem velike količine dokumentacije. Pomoću OCR tehnologije, važni podaci se ekstrahiraju iz obrazaca koje ispunjavaju pacijenti, te se pohranjuju u digitalne baze podataka, kako bi bili lako dostupni i pretraživi u svakome trenutku (Verma, Arora, Verma 2016:186-190).

OCR tehnologija iznimno je bitna i za pomoć slijepim i slabovidnim osobama. Pomoću nje slijepe i slabovidne osobe mogu pisani tekst na računalu preslušati. To je posebno važno za osobe koje još ne znaju Brailleovo pismo i stoga ne mogu čitati knjige pisane tim pismom.

Policija koristi OCR tehnologiju kod praćenja prometa prepoznajući znakove s registarskih tablica. Ta se tehnologija koristi za naplaćivanje cestarine, praćenje kretanja prometa i pojedinaca. Tako policija može vidjeti u koje doba su velike gužve na određenim cestama, te preusmjeriti promet kako bi se smanjio broj eventualnih prometnih nesreća.

Još jedna primjena OCR tehnologije je CAPTCHA. Captcha (engl. Completely Automated Public Turing test to tell Computers and Humans Apart) je način autentifikacije koji se koristi na različitim internetskim stranica, poput blogova, foruma i webmail servisa. Ona služi za sprječavanje napada zlonamjernih softvera za zlouporabu osobnih podataka. Funkcionira tako da korisnik mora upisati tekst koji se vidi na izobličenim tekstualnim slikama. Taj se tekst najčešće sastoji od brojeva i slova različite veličine i različitih fontova, a pozadina je često šarana. Captcha test je i jednostavan za rješavanje ljudima, ali ga zato trenutna softverska tehnologija ne može riješiti (Azaid, Jain 2013).



Slika 5. CAPTCHA

Izvor: https://www.lifewire.com/what-is-a-captcha-test-2483166

3.3. Faze optičkog prepoznavanja znakova

Da bi se točno razumjelo kako optičko prepoznavanje znakova funkcionira, potrebno je taj proces podijeliti u odvojene faze kroz koje OCR softver prolazi, a to su:

1. prethodna obrada,
2. analiza slike, odnosno segmentacija,

3. prepoznavanje znakova,
4. naknadna obrada.

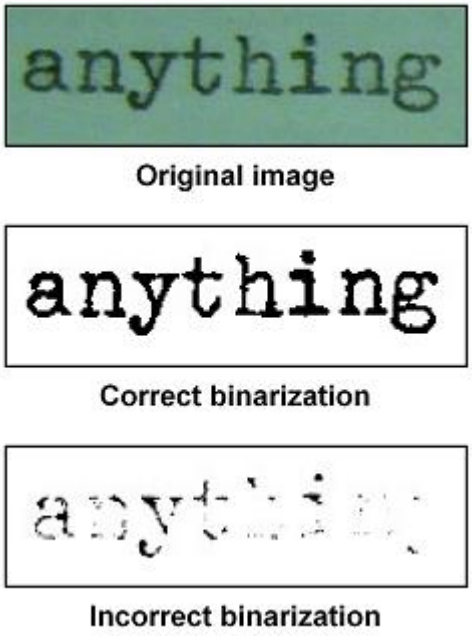
3.3.1. Prethodna obrada

Da bi program za optičko prepoznavanje znakova prepoznao tekst, dokument je potrebno skenirati i dodatno obraditi, kako bi skenirana slika bila što čišća, a završni rezultat što točniji. Sam postupak skeniranja slike već je prethodno objašnjen u poglavlju 2.1.2. Digitalizacija gradiva. Skener čita sliku i šalje ju dalje računalu na obradu. On prolazi kroz dokument s lijeva na desno (odnosno s desna na lijevo za arapski i hebrejski jezik) i odozgo prema dolje, snimajući piksel po piksel kako bi na kraju stvorio sliku.

Nakon skeniranja slike potrebno je eliminirati neželjene šumove na slici, ali bez gubljenja značajnih informacija. Prvo se provodi postupak binarizacije. Iako je prethodno navedeno kako je za najbolje rezultate optičkog prepoznavanja teksta dokument potrebno skenirati u crno-bijeloj tehnici, Vynckier ističe kako to nije uvijek pravilo. Kod dokumenata koji imaju pozadinu u boji, prilikom skeniranja u crno-bijeloj tehnici, skener neće moći uspješno razlikovati tekst od pozadine jer im boje nisu dovoljno kontrastne. Taj se problem ponekad može riješiti tako da se podesi svjetlina kako bi se pozadina dovoljno razlikovala od teksta. No u slučajevima kada imamo crni ili tamni tekst ispisan na jednako tamnoj pozadini ili svijetli tekst na svijetloj pozadini, ni to neće pomoći, jer će se prilikom mijenjanja svjetline cijelog dokumenta, osim svjetline pozadine, osvijetliti i sami tekst, pa opet nećemo imati dovoljan kontrast za kasnije čitanje teksta. Time možemo izgubiti dijelove teksta, ali i uvesti „buku“ koja kasnije smeta kod prepoznavanja znakova (Vynckier 2017).

Kod binarizacije se koristi filter praga (engl. threshold filter): to je prag koji određuje koji će se dijelovi dokumenta pobojati bijelo, a koji crno, odnosno, pikseli, čija je svjetlina veća od praga, postat će bijeli, a pikseli, čija je svjetlina manja od praga, postat će crni. Filter praga može biti fiksan iznos kod dokumenata s visokim kontrastom i jednoličnom pozadinom, no kod dokumenata s visokom razinom kontrasta potrebno je koristiti određene metode za određivanje praga (Eikvil 1993:12). Najbolje se slike dobivaju korištenjem metoda kod kojih je moguće mijenjati prag kroz dokument, prilagođavajući se svjetlini i kontrastu. To je potrebno kada je pozadina šarena ili kad sva slova u tekstu nisu iste boje. Kod korištenja takvih metoda, dobivena slika zauzima veću memoriju.

U nekim slučajevima čak ni binarizacija nije dovoljna, pa se uz nju koristi i zaglađivanje boje, kod koje se boja piksela zamjenjuje prosjekom piksela koji okružuju početni piksel. Time se zaglađuju razlike u intenzitetu i kasnije se dobiva bolja čitljivost. (Vynckier 2017)



Slika 6. Usporedba rezultata neispravne i ispravne binarizacije

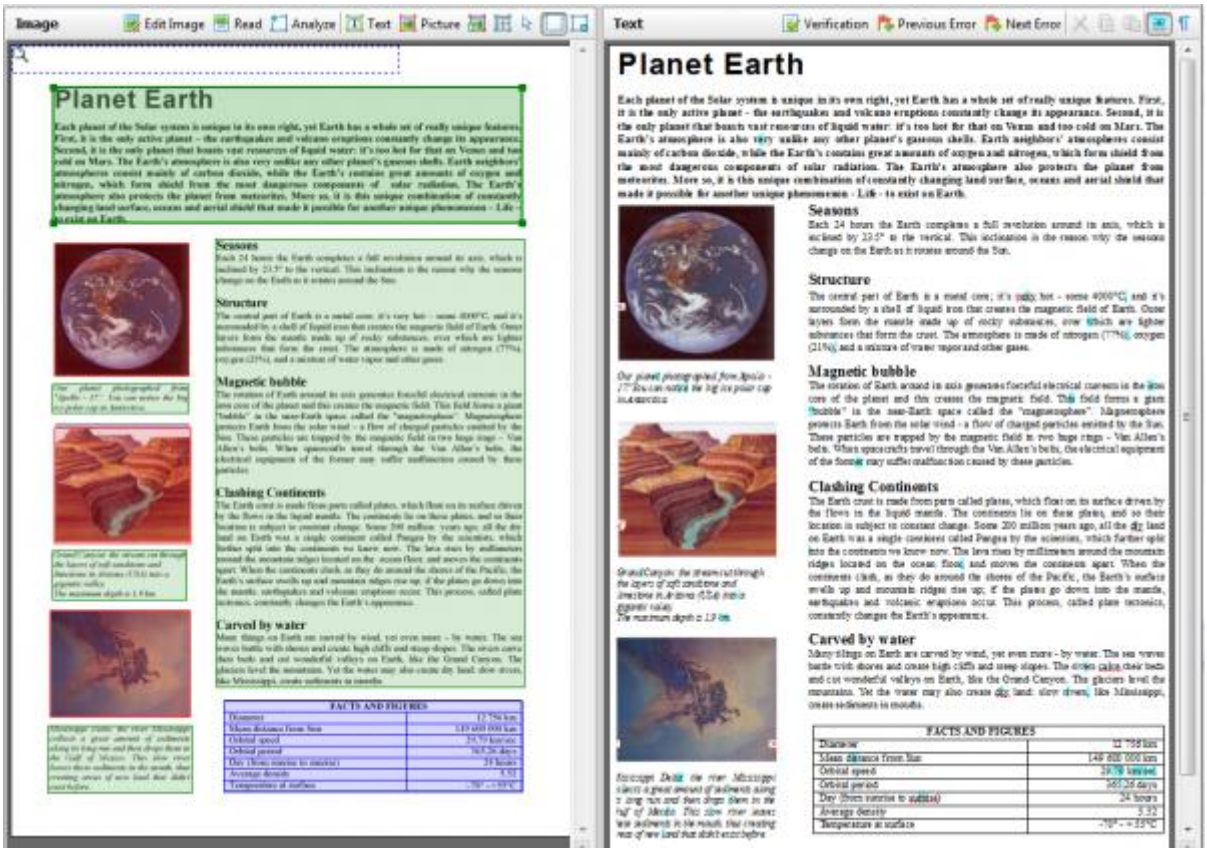
Izvor: https://www.nicomsoft.com/optical-character-recognition-ocr-how-it-works/

Ovisno o rezoluciji skenera, moguće je da dobivena slika sadržava nečistoće. Neki znakovi tako mogu biti razlomljeni, imati rupe ili biti zamrljani. Kako bi se dobio što točniji rezultat nakon cijelog postupka OCR-a, potrebno je popuniti praznine i rupe unutar znakova, te suziti širinu crta. Osim toga, potrebno je provesti proces normalizacije, odnosno izjednačiti veličinu, nagnutost i rotaciju stranica i linija teksta. Za to se koristi varijanta Hough transformacije za otkrivanje izvrtanja. No rotacija pojedinačnih znakova moguća je tek nakon prepoznavanja znakova, jer je tek tada moguće odrediti kut rotacije. Nakon ove faze dobiva se slika bolje kvalitete, spremna za sljedeću fazu.

3.3.2. Analiza stranice − segmentacija

Analiza slike, odnosno segmentacija, proces je klasifikacije, gdje se dokument dijeli na homogene zone. Svaka zona smije sadržavati samo jednu vrstu informacije, bila to slika, tekst ili tablica (Abdulwahhab Hamad, Kaya 2016:246). Primjerice, kod optičkog prepoznavanja znakova s osobne iskaznice dio slike s fotografijom vlasnika odvaja se od dijela na kojem su

napisani njegovi podaci, poput imena i prezimena. Ovaj postupak program može provesti automatski, no često može doći do pogrešaka kod označavanja zona teksta, tablica i slika. Korisnik može sâm ručno označiti dijelove povlačeći pravokutnike preko dijelova koje želi označiti, no taj postupak zahtijeva puno vremena. Zato je najbolje rješenje pustiti program da automatski provede analizu stranice i onda naknadno ispraviti greške, ako ih bude.



Slika 7. Analiza stranice

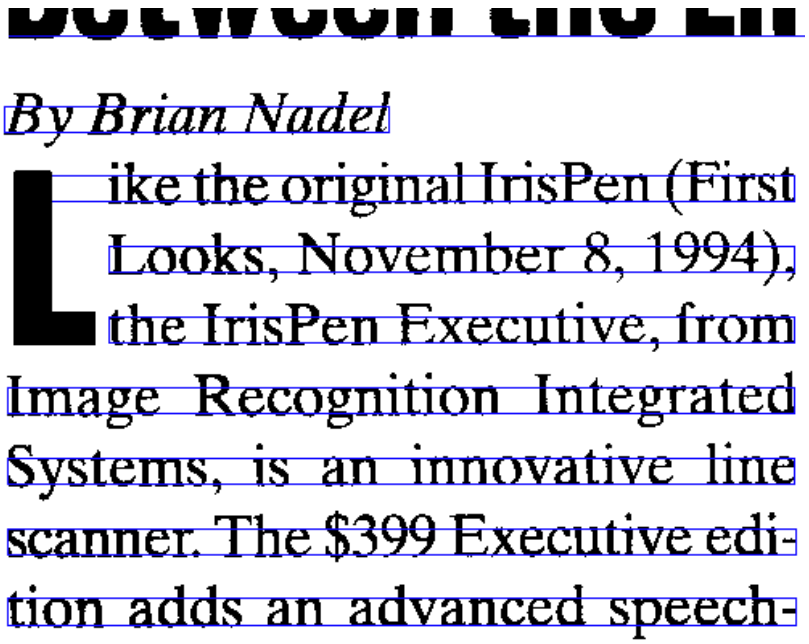
Izvor: https://abbyy.technology/en:features:ocr:document_analysis

Ako slika nije točnog pravokutnog oblika, program će to detektirati i točno ju označiti. Ako cijelo vrijeme imamo dokumente s jednakim ili sličnim izgledom, poput osobnih iskaznica, moguće je i spremiti u programu predložak za označavanje, kako bi ubrzali cijeli postupak. Kod analize azijskih pisama potrebna je posebna analiza zbog njihovih znakova, jednako kao i kod arapskog i hebrejskog pisma, jer je kod njih smjer pisanja s desna na lijevo. Točnost cijelog optičkog prepoznavanja znakova uvelike ovisi o točnosti postupka segmentacije, stoga je potrebno da sve zone budu pravilno označene.

Nakon analize stranice, OCR program dijeli stranicu na odvojene zone. Postupak segmentacije dijeli se na tri faze:

1. segmentacija linija,
2. segmentacija riječi,
3. segmentacija znakova.

Prvo se područje teksta odvaja u zasebne linije. Linije teksta moraju biti dovoljno odvojene, kako bi tekst bio čitljiv čovjeku. Ovaj postupak prilično je jednostavan za provedbu, no problemi se pojavljuju u slučaju kada se znakovi iz dviju linija dodiruju ili preklapaju. Probleme kod segmentacije linije rade uvećana početna slova na početku poglavlja, koja pokrivaju nekoliko linija teksta, no današnji OCR programi ipak ih uspješno otkrivaju.



Slika 8. Segmentacija linija

Izvor: http://www.how-ocr-works.com/OCR/line-segmentation.html

Nakon segmentacije linija slijedi segmentacija riječi. Riječi se uvijek odvajaju bijelim prostorom između njih, koji je unutar teksta s lijevim, centralnim i desnim poravnavanjem uvijek iste duljine, no kod obostranog poravnavanja može biti različite duljine u svakom redu. U arapskom pismu razmak između riječi uvijek je jednake duljine, ali se zato neka slova izdužuju, kako bi tekst bio poravnan.

Poslije segmentacije riječi, riječ se dalje dijeli na znakove, odnosno slova, brojeve, interpunkcijske i druge specijalne znakove. Ovisno o fontu, znakovi unutar jednog teksta mogu zauzimati isti ili različiti prostor. Fontovi se tako dijele na fiksne i proporcionalne fontove. Kod fiksnih fontova svaki znak zauzima jednak prostor. To ne znači da su svi znakovi jednake širine, već da je prostor znaka uvijek jednake veličine (znak + prazan prostor oko njega). Dokumenti tiskani na starim pisaćim mašinama pisani su fiksnim fontovima. Kod proporcionalnih znakova znakovi zauzimaju različit prostor, ovisno o njihovoj širini. Tako će recimo slovo „i“ zauzimati puno manji prostor nego slovo „m“. To je primjerice slučaj kod fonta Times New Roman, kojim je pisan ovaj rad.

Znak se najčešće sastoji od jedne cjeline, no postoje i neki znakovi koji su sastavljeni od više dijelova. Tako se recimo navodni znak „“ sastoji od dva dijela, a znak za postotak „%“ od čak tri (Radošević 1996:21). Tijekom segmentacije znakova može doći do problema, ako je jedan znak prepolovljen u dva dijela ili ako se dva znaka dodiruju. Do segmentacija znakova ne dolazi kod OCR softvera, koji čitaju cijele riječi, a ne zasebne znakove. Oni koriste neuronske mreže, koji djeluju po uzoru na ljudski mozak. Kod takvih programa slika cijele riječi se uspoređuje s riječima. Ta je tehnika slična prepoznavanju govora. U usporedbi s prepoznavanjem zasebnih znakova, prepoznavanje cijelih riječi je puno lakše jer je puno lakše prepoznati loše otisnutu ili isprekidanu riječ, nego loše otisnut znak (Vynckier 2017).

3.3.3. Prepoznavanje znakova

Nakon što su se svi znakovi izdvojili, potrebno ih je odvojeno prepoznati. Radošević (1996) izdvaja dvije glavne metode prepoznavanja znakova:

1. prepoznavanje na temelju predložaka,
2. prepoznavanje na temelju svojstava oblika.

Postupak prepoznavanja na temelju predložaka provodi se tako da se svaki odvojeni znak uspoređuje s gotovim predlošcima, pohranjenim u bazi podataka. Ta bi metoda bila uspješna kada bi svi koristili samo jedan font, te kada bi svi imali identičan rukopis. Upravo zato su i stvoreni fontovi OCR-A i OCR-B. Tadašnji OCR programi bili su istrenirani da prepoznaju upravo te fontove, te je rezultat bio prilično točan. Kako bi se znak mogao usporediti s njegovim predloškom, potrebno je provesti postupak normalizacije. Nakon uspoređivanja dobiva se stupanj sličnosti izražen u postupcima, koji nam govori koji predložak je najsličniji našem znaku. Razvijanjem OCR programa omogućeno je uspoređivanje znakova s većim

brojem sličnih fontova, kao što su Times i Helvetica. No i dalje se mogu pojaviti novi fontovi, koje program neće znati prepoznati. Budući da znamo da postoji preko 50.000 različitih fontova, a rukopisa beskonačno mnogo, ova metoda je korisna samo kod tekstova koji su pisani često korištenim fontovima (Woodford 2018).

Slika 9. Usporedba različitih fontova korištenih u Wordu

Puno kompleksnija metoda je prepoznavanje na temelju svojstava oblika, poznata i kao ICR (engl. Intelligent Character Recognition). Znakovi se prepoznaju zahvaljujući njihovim značajkama. Umjesto da se prepoznaje cjelokupni znak, određuju se njegove pojedinačne komponente (nagnute linije, spojevi, zaobljeni dijelovi).

Slika 10. Prepoznavanje na temelju svojstava oblika: komponente slova A

Izvor: https://www.explainthatstuff.com/how-ocr-works.html

Kada se unutar OCR programa koristi više OCR strojeva, dobivaju se različite varijante za pojedini znak ili riječ i onda je potrebno „glasanje“. „Glasanje“ će se provesti uz upotrebu određenih baza podataka i algoritama, kako bi se odredio ispravan znak ili riječ. Ako program nije potpuno siguran o kojem je znaku riječ, napravit će se numerička procjena vjerojatnosti da određena slika zapravo predstavlja određeni znak. Primjerice, kod loše otisnutog slova „o“

u riječi „kos“, program neće biti siguran radi li se o slovu „o“, „e“, ili „c“. Procjenom vjerojatnosti dobit će postotci vjerojatnosti za svako od tih slova:

- slovo "o" – 95%,
- slovo "c" – 82%,
- slovo "e" – 65%.

Time će se riječ ispravno prepoznati kao riječ „kos“. Ponekad kod jako loše otisnutih znakova postoji mogućnost da će vjerojatnost biti veća za neki neispravni znak (na našem primjeru, ako se unutar slova „o“ pojavila mrlja, postotak za slovo „e“ mogao bi biti veći nego za slovo „o“). U takvim se slučajevima koriste instalirani rječnici za jezik kojim je pisan tekst, kako bi se provjerilo nalazi li se dobivena riječ u njemu. Tako slovo „kes“ ne postoji u hrvatskom rječniku, pa će se ipak odabrati slovo „o“ kao ispravan znak u riječi „kos“. Odnos između instaliranih rječnika i algoritama i hipotezama dosta je kompliciran, a softverske tvrtke ne otkrivaju kako se oni zajedno integriraju (Holley 2009).

Većina današnjih OCR programa koriste metodu prepoznavanja na temelju svojstava oblika, i oni prepoznaju znakove neovisno o fontu kojim je tekst pisan. Takvi programi nazivaju se i Omnifont OCR programi.

3.3.4. Naknadna obrada

Nakon faze prepoznavanja znakova, potrebno je ponovno sastaviti sve znakove kako bi se dobio cjeloviti tekst, te provjeriti tekst kako bi se ispravile moguće pogreške. Postupak ponovnog sastavljanja teksta naziva se grupiranje. Grupiranje može stvarati probleme, ako skenirani tekst nije poravnan, pa je teško izdvojiti redove teksta. Tada je potrebno popraviti nagib teksta. Razdvajanje riječi unutar teksta također može biti kompliciranо jer je potrebno odrediti koliki broj razmaka je potrebno dodijeliti prepoznatim znakovima (Radošević 1996:24).

Nakon sastavljanja teksta program upućuje na znakove koji su mu neprecizni, te nudi opciju ručnog ispravljanja mogućih grešaka. Postoje dvije metode pomoću kojih program otkriva greške. Prva proučava kako je niz znakova poredan jedan za drugim. Primjerice, nakon točke pretpostavka je da sljedeće slovo mora biti veliko. Jednako tako može se odrediti koja slova u zadanom jeziku ne mogu slijediti jedno za drugim. Tako se recimo slovo „ć“ nikada neće naći ispred slova „č“, pa ako se prepozna ta kombinacija, program će javiti grešku. Druga metoda

je učinkovitija i zahtijeva korištenje rječnika. Riječ, za koju postoji mogućnost da je pogrešna, provjerava se u rječniku, i ako ona u rječniku ne postoji, javlja se greška. Riječ se tada ispravlja u najsličniju riječ, za koju postoji najveća vjerojatnost da je točna. No postoji mogućnost da iako riječ ne postoji u rječniku, da je ona točna, pa će se riječ pogrešno prepraviti u drugu. Zato je i veoma važno da se u OCR programu označi točan jezik kojim je pisan tekst, ili više njih, ako je tekst pisan na više jezika. I dalje je moguće da će i nakon provjere i ispravljanja grešaka koje provede program, ostati grešaka. Ako nam je potrebno da tekst bude 100% točan, morat ćemo sami provjeriti tekst i ispraviti moguće greške, što je veoma iscrpljujuć i zahtjevan posao, te zahtijeva puno vremena i visok stupanj koncentracije (Eikvil:21-22).

Tablica 2. Faze optičkog prepoznavanja znakova

Faza	Opis	Postupci
Prethodna obrada	Proces dobivanja slike i poboljšavanja njene kvalitete.	Skeniranje, binarizacija, zaglađivanje boje, uklanjanje šumova, normalizacija, rotacija.
Analiza stranice	Podjela slike na njene sastavne dijelove.	Segmentacija linija, segmentacija riječi, segmentacija znakova .
Prepoznavanje znakova	Svrstavanje svakog zasebnog znaka u posebnu kategoriju.	Prepoznavanje na temelju predložaka, prepoznavanje na temelju svojstava oblika.
Naknadna obrada	Poboljšavanje točnosti OCR rezultata.	Grupiranje, ispravljanje grešaka: automatski i ručno.

4. Najčešće greške kod optičkog prepoznavanja znakova

Ako sve faze OCR-a nisu uspješno provedene ili je izvornik u veoma lošem stanju, pojavit će se greške u rezultatu. Čak i ako je izvornik u odličnom stanju i ako smo cijeli postupak digitalizacije i OCR pravilno proveli, rezultat nikada neće biti 100 % točan. Neke greške se često pojavljuju, pa ćemo tako nabrojiti neke od najčešćih grešaka u tekstovima dobivenim optičkim prepoznavanjem znakova.

Prva greška je odbijanje. Može se dogoditi da određene znakove program ne može prepoznati, pa će se ti znakovi zamijeniti znakom „~“. Druga greška je zamjena, što znači da program

pogrešno prepoznaje određeni znak. Do zamjene dolazi kod slova koja su slična po obliku i strukturi, kao što su „h“ i „b“ ili „c“, „e“ i „o“. Ponekad će pogrešno prepoznat znak stvoriti novu riječ, koja doista postoji u rječniku, pa se tijekom kasnije provjere ta pogreška neće ispraviti, već ju prepoznati i ispraviti može jedino čovjek. Recimo unutar riječi „lak“ slovo „a“ može biti pogrešno kao slovo „u“, ali budući da riječ „luk“ postoji u hrvatskom jeziku, program neće posebno istaknuti ovu riječ. Ovisno o softveru, neki OCR programi, kada nisu sigurni u određeni znak, radije će staviti znak „~“ nego staviti pogrešan znak, za razliku od drugih koji će radije staviti bilo koji znak, nego neodređeni „~“. To ovisi o "pragovima sigurnosti" korištenim u OCR motorima. Neki radije nude relativnu sigurnost za svoje rezultate, što bi značilo da se u njihovim rezultatima češće pojavljuje neodređeni znak, ali zato i manje grešaka, dok drugi preferiraju rezultat bez neodređenog znaka koji će biti upitne točnosti. Ove se greške pojavljuju kada se na znaku pojavi mrlja ili kada dijelovi znaka nisu dobro otisnuti.

Sljedeća moguća greška je greška veliko-malo slovo. Tako se slovo koje bi trebalo biti veliko, primjerice prvo slovo vlastite imenica ili riječi na početku rečenice, može slučajno zamijeniti malim slovom, ili pak slovo koje bi trebalo biti malo, zamijeni velikim. Još neke od čestih grešaka su one vezane uz razmake. Tako se recimo dva odvojena znaka ili dvije riječi mogu pogrešno prepoznati kao jedan spojen znak ili riječ, ili se pak jedan znak ili riječ može pogrešno prepoznati kao dva znaka ili riječi (Vynckier 2017).

Do pogrešnog spajanja može doći ako se radi o tamnoj fotokopiji, ili ako se između dva odvojena znaka pojavi mrlja, a do pogrešnog odvajanja, ako se radi o svijetloj fotokopiji. Osim toga bitno je da je razmak između znakova konstantan, jer ako se razmak između dva znaka unutar teksta slučajno smanji, program ih može pogrešno prepoznati. Česte greške su i one s interpunkcijskim znakovima. Ponekad se mrlje mogu prepoznati kao točka, zarez ili navodni znak, ali i obrnuto, pa će se na mjestima gdje bi se trebao nalaziti razmak pojaviti interpunkcijski znak, ili će mjesto na kojem bi trebao biti interpunkcijski znak biti prazno (Eikvil 1993). Kod pretraživanja greške nemaju jednaku važnost, pa tako recimo ako se u riječi koja se želi pretražiti pojavi greška odbijanja, zamjene ili razdvajanja, ta će se riječ jako teško pronaći unutar teksta. No budući da pretraživanje unutar word ili pdf dokumenata nije osjetljivo na velika i mala slova, lako će se pronaći željenu riječ.

Tablica 3. Vrste grešaka

(Vynckier 2017)

Očekivani rezultat	Rezultat	Vrsta greške	Objašnjenje
riječ	rij~~	Odbijanje	Znak nije uspješno prepoznat, zamijenjen je znakom „~“.
	rlječ	Zamjena	Znak je pogrešno prepoznat.
	rijEč	Veliko-malo slovo	Malo slovo je zamijenjeno velikim.
napisati riječ	napisatiriječ	Spajanje	Nedostaje razmak između riječi.
riječ	ri ječ	Odvajanje	Riječ je razdvojena.
nova riječ	nova.riječ	Interpunkcijski znak	Umjesto razmaka stavljena je točka.

Osim navedenih pogrešaka, dolazi i do grešaka kod same klasifikacije dijelova stranice. Slika se ponekad može pogrešno označiti kao tekst, pa tada bude poslana na OCR. Kao rezultat dobit će se niz nesuvislih znakova, a kod završnog spajanja svih dijelova umjesto slike stajat će pogrešno prepoznat tekst. To se događa, ako se unutar slike nalazi neko slovo ili broj, ili oblik koji nalikuje nekom znaku. Jednako je tako moguće da se tekst pogrešno prepozna kao slika, pa ni ne bude poslan na OCR. To se događa kada su pozadina teksta i slika jednake boje ili kada se tekst nalazi odmah uz sliku, kao npr. potpis autora slike ili fotografije.

Greške se mogu pojaviti i kod prepoznavanja i analize tablice. Program može pogrešno prepoznati sliku ili tekst kao tablicu, ako se unutar slike pojavljuju ćelije ili ako su riječi unutar teksta organizirane kao tablica. Mogu se pojaviti greške i ako su neke ćelije spojene ili ako je tekst unutar ćelija pisan različitim fontovima. Ako predobrada nije dobro provedena, program može pogrešno prepoznati mrlje ili sjenu kao tekst, pa se na završnom rezultatu može pojaviti grupa nesuvislih znakova (Andrianov 2009).

5. Točnost OCR-a

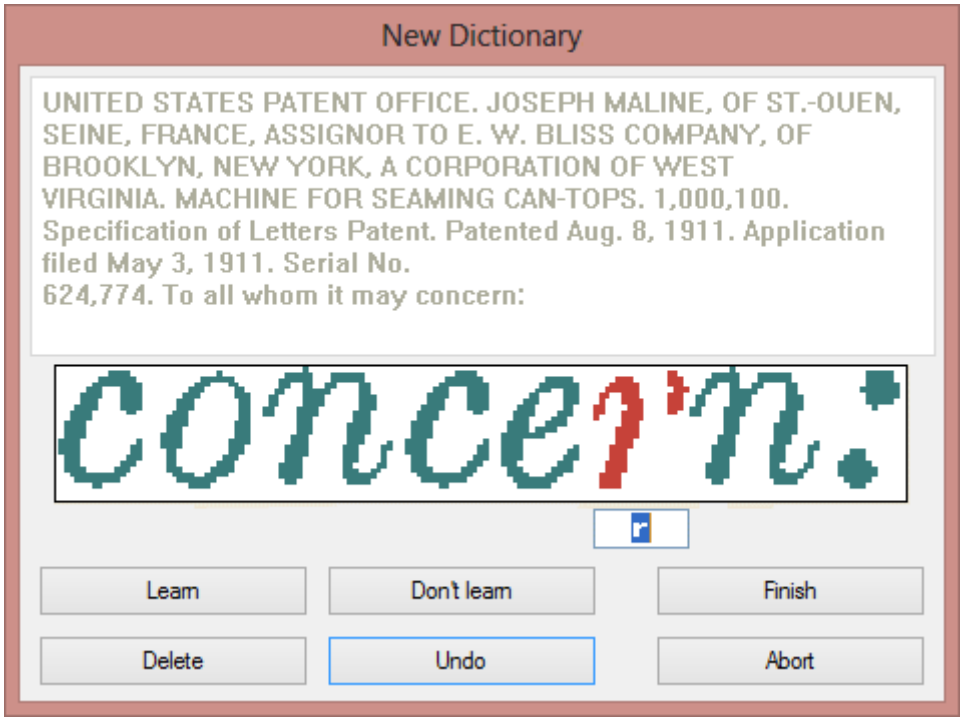
Na točnost rezultata optičkog prepoznavanja znakova utječu brojni faktori. Krenut ćemo od samog izvornika, njegove složenosti, kvalitete i stanja u kakvom se on nalazi. Ako se izvornik sastoji od samog teksta, pisanog jednim fontom, točnost OCR-a može biti vrlo visoka, no ako

se uz tekst pojavljuje velik broj fontova ili stupaca i ilustracija, možemo očekivati da će točnost biti nešto niža. Kod materijala iz 19. ili 20. stoljeća pisanih latiničnim oblikom pisma, jednim od često korištenih fontova, rezultat OCR-a trebao bi biti prilično točan (Stančić). No ukoliko je izvornik star i loše očuvan, mogu se pojaviti brojne smetnje, poput mrlja, poderotina i presavinutih stranica. U knjigama iz knjižnica često možemo pronaći bilješke koje su zapisivali njeni korisnici, koji također mogu smetati kod OCR i rezultirati loše pročitanim tekstom. No i sam otisak teksta može biti nekonzistentan i nejednoličan, a korišten font zastario. Osim toga, kod pisaćih strojeva, batić za slovo koje se često upotrebljava, može biti istrošen, pa će to slovo biti loše otisnuto, ili boja u printeru može biti pri kraju, pa će sva slova biti veoma svijetla. Tada je velika vjerojatnost da softver neće moći prepoznati pojedine znakove ili će ih pogrešno prepoznati. Kako bi se to izbjeglo, za početak je potrebno detaljno razraditi sam plan digitalizacije i skeniranja materijala, te pametno odabrati uređaj kojim ćemo skenirati materijal, te program za OCR. Osim toga, kod većih projekata potrebno je izvršiti pokusno skeniranje, kako bi bili sigurni da će se skeniranjem dobiti kvalitetne slike. Točnost OCR-a uvelike ovisi i o podešavanju rezolucije. I ovdje je potrebno napraviti testiranje prije samog projekta, tako da se par stranica skenira u različitim rezolucijama. Nakon toga potrebno je usporediti rezultate OCR-a kako bi se odabrala točna rezolucija. Najčešće se uzima rezolucija od 400 dpi, ali kod nekih izvornika i manje ili veće rezolucije mogu dati bolje rezultate.

Kod starijih i izblijedjelih materijala punih mrlja, iznimno je važno pravilno provesti fazu prethodne obrade. Njome ćemo potpuno ili barem djelomično riješiti problem kontrasta, te ćemo izvesti korekciju boje, čišćenje slike uklanjanjem buke i smeća, ispraviti zakrivljenost linija teksta, podijeliti sliku na dvije stranice, ispraviti zakrivljenost slike i provesti binarizaciju.

Važan segment poboljšavanja točnosti OCR programa je treniranje programa. Tu opciju imaju samo neki OCR programi i ona se odvija nakon faze prepoznavanja znakova. Treniranje programa odvija se tako da se u skočnom prozoru korisniku istaknu znakovi za koje program nije u potpunosti siguran jesu li točni. Program nikada ne može 100% znati je li dobiveni znak točan, već samo može biti siguran ili nesiguran u dobiveni rezultat. On daje razinu pouzdanosti između vrijednosti 0 i 9. Korisnik tada treba potvrditi interpretaciju OCR programa za znak za koji je program nesiguran, ako je ona točna, ili ju ispraviti i onda potvrditi. Nakon toga taj se znak pohranjuje u bazu podataka. To je iznimno korisno za stare fontove ili za stare i oštećene materijale. Sve te oblike softver pamti i koristit će za slične

znakove u budućnosti (Holley 2009). Osim nejasnih znakova, pomoću treniranja program može naučiti i neke nove znakove. To mogu biti novi znakovi u matematici ili recimo oznake valuta. Tako možemo istrenirati program da bitmapu ✆ prepoznaje kao „tel“ te u njegovu bazu podataka pohraniti znak za autorsko pravo „©“ i znak za registrirani zaštitni znak „®“. Što više pohranjenih znakova program ima unutar svoje baze podataka, to je on bolji, stoga im je u cilju da što više korisnika koristi opciju treniranja.



Slika 11. Treniranje OCR programa

Izvor: http://www.how-ocr-works.com/accuracy/accuracy.html

Iako je proces treniranja veoma koristan, jer unaprijed sprječava greške i povećava brzinu i preciznost sustava, on zahtjeva puno vremena i duboku koncentraciju korisnika, pa se zato ne koristi kod velikih projekata kod kojih je jasno definiran rok. Treniranje nije moguće za neke azijske jezike, kao što su japanski ili kineski. U tim jezicima ne postoje slova, nego simboli, odnosno ideogrami kojih je veoma mnogo. Primjerice, kineski jezik sadrži preko 40.000 različitih simbola. U OCR programe ideograme je potrebno utipkati ručno. (Vynckier)

Točnost OCR-a računa se uspoređivanjem teksta dobivenog korištenjem OCR programa s potpuno točnim tekstom, a rezultat se izražava kao postotak. Prilikom računanja točnosti OCR programa treba obratiti pozornost na tri ključna faktora:

- stopa prepoznavanja (udio ispravno prepoznatih znakova),
- stopa odbijanja (udio znakova koje program nije uspio prepoznati),
- stopa pogrešaka (udio pogrešno prepoznatih znakova) (Eikvil 1993).

OCR program koji bi davao rezultate s točnošću od 100% ne postoji. Današnji OCR programi daju rezultate s prosječnom točnošću od 99,95% (Stančić). To bi značilo da se na 1000 znakova pojavljuje 5 grešaka. To dakako vrijedi za materijale koji su u odličnom stanju i kod kojih nema puno mrlja ili poderotina. Kod starijih materijala točnost OCR-a puno je niža.

Tanner, Muñoz i Ros u svome članku pišu kako točnost od 99,98% ili više mogu imati samo tekstovi pisani 1950. godine i na dalje, dok je za tekstove pisane između 1900. i 1950. godine stopa točnosti u prosjeku 95%. Svi tekstovi pisani prije 1900. imaju puno nižu stopu točnosti, pa se tako točnost od 85% ili više za tekst pisan prije 1900. godine može smatrati odličnim rezultatom (Tanner, Muñoz, Ros 2009).

Osim starosti, na točnost OCR rezultata može utjecati i vrsta gradiva. Na istraživanju Programa digitalizacije novina Nacionalne knjižnice Australije, prilikom OCR-a novina iz razdoblja 1803.-1954. godine točnost je bila 71%. Općenito točnost OCR-a kod digitalizacije novina puno je niža nego kod knjiga. Zato se za stare novine uzima da je :

- Dobra OCR točnost = 98-99% točna (1-2% OCR-a nije točno),
- Prosječna točnost OCR-a = 90-98% točna (2-10% OCR-a nije točno),
- Loša OCR točnost = točna ispod 90% (više od 10% OCR-a nije točno).

Uzima se da, ako je točnost OCR-a 90% ili više, OCR se isplati, no ako točnost padne ispod 90%, više se isplati ručno prepisati tekst. Točnost od 90% značila bi da se na 100 znakova pojavljuje 10 grešaka, a takav tekst bilo bi teško za pročitati, a proces ispravljanja grešaka dugotrajan (Holley 2009). Iako se smatra da su najveći izazovi za optičko prepoznavanje znakova rukom pisani tekstovi, točnost prepoznavanja pisanih slova još je niža nego točnost rukopisa (Shahi, Ahlawat, Pandey 2012).

6. Digitalizacija i optičko prepoznavanje znakova iz povijesnih tekstova

Povijesni tekstovi čine ogromnu riznicu povijesnih informacija, sačuvanih u knjižnicama, arhivima i muzejima diljem svijeta. No za digitalizaciju i optičko prepoznavanje znakova oni predstavljaju velik izazov, jer postupak digitalizacije i OCR nije tako jednostavan, kao za

današnje tekstove, pisane standardiziranim jezicima i tiskane modernim printerima. Probleme, koji se javljaju prilikom digitalizacije povijesnih tekstova, možemo podijeliti na grafičke i leksičke te grafemske.

6.1. Grafički problemi prilikom OCR-a starih tekstova

Za početak, sam izvornik predstavlja problem kod skeniranja i OCR-a. Prije samog skeniranja ponekad se mora prvo očistiti knjiga od prašine, a u pojedinim slučajevima i restaurirati. Za takve poslove unajmljuju se posebne tvrtke, jer je to iznimno delikatan posao i nestručna bi osoba mogla napraviti veliku štetu. Stare knjige vrlo su osjetljive i s njima se mora oprezno baratati, a prilikom skeniranja ne smiju se otvoriti do 180 stupnjeva. Zato se kod takvih knjiga provodi nedestruktivno skeniranje, odnosno, kako bi se maksimalno zaštitio izvornik koriste se posebni skeneri u obliku slova V. Stranice često okreće robot kako ne bi došlo do oštećivanja stranica ljudskom rukom. Papiri, korišteni u prošlosti, često nisu bili bijeljeni, pa su već na početku bili tamniji nego današnji papiri, a s vremenom bi još dodatno promijenili boju i počeli se raspadati, ako nisu bili čuvani u optimalnim uvjetima. Kod OCR-a takvo gradivo predstavlja problem, jer kontrast između pozadine i teksta nije velik. Često se događalo da su stari tekstovi tijekom vremena mijenjali vlasnike i mjesto na kojem su se nalazili, pa bi ponekad došlo do gubljenja stranica, ili dijelova knjiga, ili su se zbog cenzure šarali dijelovi teksta. Kasnije su se radile rekonstrukcije izgubljenih dijelova, ali takvi tekstovi nisu bili identični nekadašnjem originalu, jer su se dijelovi mogli samo pogađati. Otisnuti tekst često je mutan ili loše otisnut, te s vremenom izblijedi. Ponekad tinta s jedne stranice zna prijeći na sljedeću, ako se dovoljno ne osuši (Hauser 2007).

Sljedeći problem su fontovi korišteni u starim tekstovima. Današnji OCR softveri često loše prepoznaju stare fontove. Slova u takvim fontovima često su veoma slična, pa ih je teško razlikovati. U starim fontovima često su korištene i ligature, odnosno dva spojena slova, pa OCR program takva slova može pogrešno prepoznati kao jedno. Kod starih tekstova razmaci između slova često su nekonzistentni (Pirker, Wunzinger). Nekada je izrada fontova bila prava umjetnost, te su se čak tekstovi pisani istim fontovima znali uvelike razlikovati. Još jedna posebnost takvih tekstova bila su velika početna slova (tzv. inicijali) u uvodu u odlomak ili poglavlje nacrtana poput umjetničkog djela. Slika je podsjećala na slovo, ali je često bila šarena i sadržavala je ljude ili životinje. Takvo slovo program prepoznaje kao sliku, pa u konačnom rezultatu OCR-a nedostaje jedan znak. Zato se mora paziti kod naknadne obrade,

kako program ne bi pogrešno ispravio sljedeće slovo u veliko početno i kako u konačnom rezultatu riječi ne bi imala na početku dva velika slova.

Slika 12. Primjeri početnih slova (B, L, E) korištenih u povijesnim tekstovima

Izvor: http://www.how-ocr-works.com/OCR/line-segmentation.html

6.2. Leksički i grafemski problemi prilikom OCR-a starih tekstova

U leksičke i grafemske probleme ulaze sve varijacije pravopisa, morfološke promjene i varijacije, zastarjeli vokabular i posebni skup znakova i kratice (Hauser). Pravopisne varijacije dijele se na dijakronijske i sinkronijske. Dijakronija proučava povijesni razvoj određenih lingvističkih pojava i jezičnog sustava u cjelini, dok sinkronija označava stanje nekoga jezika onako kako postoji u nekom vremenskom trenutku (Hrvatska enciklopedija). Sinkronijske varijacije često su vezane uz dijalekte koji uzrokuju lokalne varijacije. Uspješnost naknadne obrade uvelike ovisi o leksikonu svih oblika riječi i leksikonu povijesnih varijacija pravopisa (Springmann, Najock, Morgenroth, Schmid, Gotscharek, Fink 2014). U današnje vrijeme jezici su standardizirani, pa nema toliko pravopisnih varijacija.

Sljedeći problem vezan je uz morfološke promjene i varijacije, te zastarjeli vokabular. Morfološke promjene označavaju promjene u strukturi riječi. Riječi korištene u povijesnim tekstovima po svojoj strukturi uvelike se razlikuju od danas korištenih riječi, a neki oblici riječi više se ne koriste. U starim tekstovima pojavljuju se mnoge riječi koje se danas više ne koriste. Iako se neke riječi više ne koriste u svakodnevnom govoru, one i dalje postoje u

rječnicima, dok nekih više nema niti u rječnicima. Budući da OCR programi kod provjere koriste suvremene rječnike i gramatičke strukture, a u sebi najčešće nemaju inkorporirane povijesne rječnike, zastarjele riječi i njihovi oblici bit će označeni kao greške. Osim toga, događa se da povijesni tekstovi često ne slijede specifične pravopisne strukture i pravila, pa tako jednake riječi unutar jednog teksta mogu biti različito napisane. (Pirker, Wurzinger)

U povijesnim tekstovima pojavljuju se i znakovi i slova koja se danas više ne koriste, niti ne postoje u tim jezicima. Osim znakova, pojavljuju se i kratice, koje danas više nemaju značenje, koje su nekada imale.

7. Razlike između staroslavenskog i suvremenog ruskog jezika

Prilikom optičkog prepoznavanja znakova OCR program u svojoj bazi najčešće sadržava samo suvremene rječnike za provjeru dobivenog teksta. Velik broj starih ruskih knjiga pisan je na staroslavenskom jeziku ili nekom starijem obliku ruskog jezika te se tako uvelike razlikuje od tekstova pisanih suvremenim ruskim jezikom. Prilikom OCR-a starih ruskih knjiga, ako program ne nudi prepoznavanje na temelju staroslavenskog, program će imati problema sa svim riječima, oblicima riječi, te pravopisom koji odstupa od današnjeg standardiziranog ruskog jezika.

Staroslavenski je najstariji slavenski književni jezik, nastao u 9. stoljeću, no pisani spomenici iz tog razdoblja nisu sačuvani. Svrha njegova nastajanja bila je širenje kršćanstva[3], te se koristio kod prevođenja propovijedi i liturgijskih knjiga s novogrčkog jezika. Za sastavljanje staroslavenskog bili su odgovorni braća Konstantin: Ćiril i Metod. Staroslavenski je imao dva pisma: glagoljicu i ćirilicu. Glagoljicu je sastavio Konstantin-Ćiril[4], dok se za ćirilicu ne zna točno. Budući da djela iz 9. Stoljeća nisu sačuvana, ne zna se koje je pismo bilo prvo

[3] v. (Čelić 2008: 206, 214): „Iako su pojmovi u kazalu poredani ovim slijedom: pismenost, vjera, država …, slijed je činjenično, tj. uzročno-posljedično, obrnut. Pismenost je stavljena na prvo mjesto jer korespondira s gramatikama. No, pismenosti u Slavenâ ne bi bilo u ovome obliku da nije bilo istaknuto pitanje vjere, odabira konkretne konfesije, a to je, pak, bio državno-gospodarski problem koji će, u oba slučaja (hrvatskome i ruskome) riješiti politika. (…) Područje koje su naselila ruska plemena nije pretpostavljalo otprije razvijenu jezičnu i pismenu kulturu; stoga i istočnoslavenski znanstvenici početak pismenosti smještaju tek u 9. stoljeće , odnosno navode proces pokrštavanja kao glavni uzrok opismenjavanja.“
Čelić, Željka (2008): „Latinski metajezik – matrix slavenskih gramatika. Utjecaj latinskoga na hrvatski i istočnoslavenske jezike, prikazan jezičnim nazivljem, opisom glasova i oblika u hrvatskome i istočnoslavenskim jezicima. Zagreb: doktorska disertacija

[4] v. (Ačimović 2018: 12): „Ipak, ta brojna istraživanja s vremenom su rezultirala spoznajom da je glagoljica autorsko djelo, rezultat individualnoga čina, da ju je vjerojatno stvorio kršćanin, filolog, poliglot iz grčkoga kulturnog ozračja. Većina se stručnjaka danas slaže da je to bio Konstantin-Ćiril.“
Ačimović, Alma (2018) Upotreba ćirilice na istočno- i južnoslavenskom prostoru (od postanka do suvremenih azbuka, sociopolitički pogled). Zagreb: diplomski rad.

(Popović, 1983: 3-4). Od 12. stoljeća u staroslavenski počinju ulaziti živi slavenski jezici, pa tako dolazi do češko-moravske, panonske, ruske, hrvatske, srpske, bugarske, makedonske i vlaške ili rumunjske redakcije.

Ćirilica je nastala na temelju grčkog uncijalnog pisma te se na početku sastojala od 45 slova. Iz grčkog alfabeta ćirilica je preuzela 26 nepromijenjenih slova. Neka slova nastala su kombiniranjem grčkih slova, dok je podrijetlo ostatka slova nepoznanica (Damjanović 2003: 29-30). Tijekom različitih reforma, od kojih su najznačajnije bile reforma Petra I. u 17. stoljeću i sovjetska reforma u 20. stoljeću, neka slova su bila izbačena, a neka dodana.

Кириллица		Греческое уставное письмо	Кириллица		Греческое уставное письмо
Буквы и их название	Цифровое значение		Буквы и их название	Цифровое значение	
А –аз	1	Α	Х –хер	600	Χ
Б –буки			Ѡ –омега*	800	Ω
В –веди	2	Β	Ц –цы	900	
Г –глаголь	3	Γ	Ч –червь	90	
Д –добро	4	Δ	Ш –ша		
Є –есть**	5	Ε	Щ –ща		
Ж –живете			Ъ –ер		
Ѕ –зело*	6	Ϛ	Ы –еры		
З –земля**	7	Ζ	Ь –ерь		
І –и*	10	Ι	Ѣ –ять*		
Н –иже**	8	Η	Ю –ю		
К –како	20	Κ	Ꙗ –(и)я**		
Л –люди	30	Λ	Ѥ –(и)е**		
М –мыслете	40	Μ	Ѧ –юс малый*		
N –наш**	50	Ν	Ѫ –юс большой*		
О –он	70	Ο	Ѩ –йотов. юс малый*		
П –покой	80	Π	Ѭ –йотов. юс большой*		
Р –рцы	100	Ρ	Ѯ –кси*	60	Ξ
С –слово	200	C	Ѱ –пси*	700	Ψ
Т –твердо	300	Τ	Ѳ –фита*	9	Θ
Оу –ук**	400		Ѵ –ижица*		Υ
Ф –ферт	500	Φ			

* Буквы, исключённые впоследствии из русского алфавита
** Буквы, у которых изменилось начертание

Slika 13. Staro ćirilično pismo

Izvor: http://genobooks.narod.ru/Azbuka/Azbuka.htm

Današnje se rusko pismo sastoji od 33 slova: а, б, в, г, д, е, ё, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ъ, ы, ь, э, ю, я. U starim ruskim knjigama tako možemo naći slova koja se danas više ne koriste. To su slova „ѱ“, „ѯ“, „ѡ“, „ѵ“, „ѧ“, „ѳ“, „і“ i „ї“. Ukoliko OCR program nema opciju prepoznavanja starijeg oblika ruskog pisma, kao grešku označit će sva slova koja više ne postoje u suvremenom ruskom pismu ili će ih pogrešno prepoznati kao drugo najsličnije slovo.

Osim u pismu, postoje razlike i u gramatici. U staroslavenskom je jeziku osim nominativa, genitiva, dativa, akuzativa i instrumentala, padežā koji i danas postoje u ruskom jeziku, postojao i oblik vokativa (rus. звательный падеж). Osim jednine i množine, u staroslavenskom je postojala i dvojina koja se kasnije prestala upotrebljavati u ruskom jeziku. Imenice koje u suvremenom ruskom imaju samo oblik jednine, kao što je riječ krv (rus. кровь) u staroslavenskom su imali i oblik množine: rus. кровь – крови. U suvremenom ruskom postoje 3 vrste deklinacije imenica, dok je u staroslavenskom postojalo 5-6 vrsta. U suvremenom ruskom postoji samo 3 vremena: prošlo, sadašnje i buduće, dok su staroslavenskom postojala 4 tipa prošlog (aorist, imperfekt, perfekt i pluskvamperfekt), jedno sadašnje i 3 tipa budućeg vremena (jednostavno buduće, složeno buduće I i složeno buduće II) (Popović 1983:23-60).

8. Istraživanje

U istraživačkom dijelu ovog rada provela se analiza dvaju različitih načina i programa za optičko prepoznavanje znakova iz starih ruskih knjiga. Cilj istraživanja bio je usporediti mogućnosti koje programi nude, te točnost dobivenih rezultata. Programi odabrani za istraživanje su Abbyy FineReader (komercijalni program) i Transkribus (nekomercijalni program).

Za primjer stare ruske knjige odabrana je prva ruska gramatika Mihaila Vasiljeviča Lomonosova, napisana 1755. godine. *Ruska gramatika* (rus. *Россійская грамматика /Россїйская грамматика*) predstavlja početak proučavanja povijesti ruskog jezika, te se smatra jednim od najvažnijih djela u povijesti ruske filologije. Iako je to prva gramatika pisana ruskim jezikom za Ruse, neke gramatičke osnove bile su već prethodno postavljene. U ruskom društvu 18. stoljeća postojala je potreba za gramatikom koja bi odražavala stvarnu jezičnu situaciju u društvu i koja bi doprinijela organizaciji jezika. U gramatici je trebalo postaviti norme gramatičke strukture ruskog jezika i njegove stilistike, dati skup pravila

ruskog književnog jezika tog vremena, uzimajući u obzir stilske značajke različitih riječi, njihove gramatičke i fonetske oblike i varijante.

U *Gramatici* je Lomonosov napravio detaljnu analizu ruskog jezika, te zaključio kako je ruski, osim ostalim slavenskim jezicima, srodan i latinskom, grčkom i njemačkom jeziku. Osim toga, prvi je istaknuo podjelu ruskog jezika na tri dijalekta (Ivanov). Knjiga je prvi put objavljena 1757. godine, a kasnije je bila tiskana još nekoliko puta (1765., 1771., 1777., i 1784. godine). U gramatici Lomonosova jasno su formulirani glavni aspekti proučavanja gramatičke strukture ruskog jezika: formalni, funkcionalni i stilski, te su i kasnije gramatike konstruirane na gotovo jednak način.

Knjiga je podijeljena u 6 poglavlja, odnosno „uputa“:

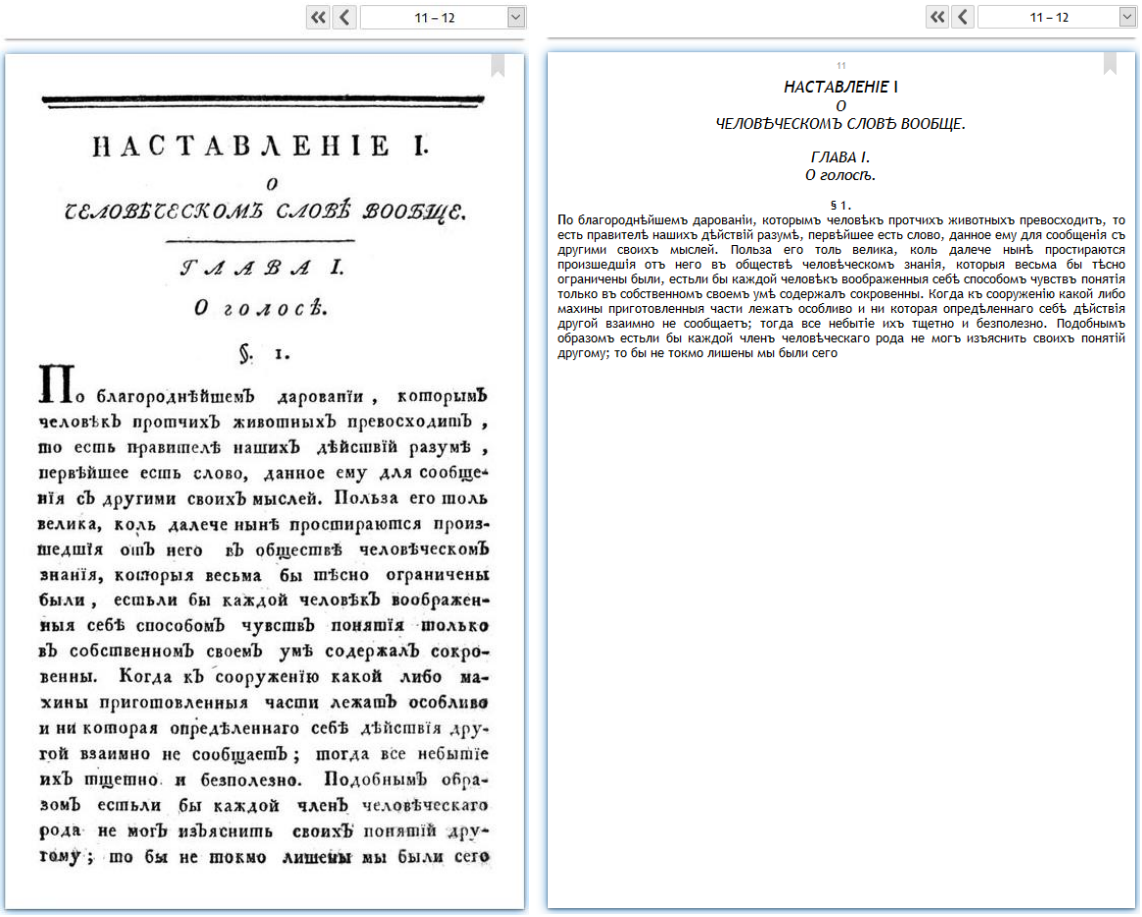
1. O ljudskoj riječi općenito (rus. О человеческом слове вообще),
2. O čitanju i pravopisu ruskom (rus. О чтении и правописании российском),
3. O imenima (rus. О имени),
4. O glagolu (rus. О глаголе),
5. O pomoćnim vrstama riječi (rus. О вспомогательных или служебных частях слова),
6. O tvorbi dijelova riječi (rus. О сочинении частей слова).

Lomonosov je u svojoj gramatici jasno odvojio ruski jezik od staroslavenskog, te je istaknuo razlike između ta dva jezika. Ovom se gramatikom tako prestaju navoditi i obrađivati neki oblici staroslavenskog jezika, poput dvojine, vokativa, određenih nastavaka u različitim padežima, pomoćnog glagola u prošlom vremenu, te vremena imperfekt i aorist.

Iako se gramatika tako odmaknula od prethodnog staroslavenskog jezika, pisani se tekst u toj knjizi i dalje uvelike razlikuje od današnjeg suvremenog ruskog jezika. Lomonosov u knjizi piše pravopisom korištenim prije reforme te koristi riječi koje danas više ne postoje u suvremenom ruskom jeziku. Osim toga, Lomonosov koristi neka slova koja više ne postoje u suvremenom ruskom pismu: „ѣ“ i „ї“. U knjizi Lomonosov spominje neke primjere iz drugih jezika, te možemo pretpostaviti kako će OCR programi imati poteškoća u čitanju tih riječi.

Budući da Lomonosovljeva *Gramatika* već postoji u digitalnom obliku, nećemo se baviti i samim skeniranjem knjige, nego ćemo koristiti postojeću knjigu u digitalnom obliku, jer je objavljeno digitalno izdanje u veoma dobrom stanju, te je kontrast između pozadine i teksta prilično dobar, a mi ne bismo mogli dobiti bolji rezultat. Kada bismo ju sami skenirali, bili bi

nam potrebni posebni skeneri u obliku slova V kako bismo sačuvali stari izvornik i dobili dobar rezultat skeniranja. Knjiga se nalazi na stranici Znanstvene pedagoške knjižnice K. D. Ušinski, te osim knjige u slikovnom obliku, moguće je ovdje pronaći knjigu i u tекstualnom obliku. To će nam biti iznimno korisno kod prepoznavanja znakova pomoću Transkribusa i kod računanja točnosti dobivenih rezultata.



НАСТАВЛЕНІЕ I.
О
ЧЕЛОВѢЧЕСКОМЪ СЛОВѢ ВООБЩЕ.

ГЛАВА I.
О голосѣ.

§. 1.

По благороднѣйшемъ дарованіи, которымъ человѣкъ протчихъ животныхъ превосходитъ, то есть правителѣ нашихъ дѣйствій разумѣ, первѣйшее есть слово, данное ему для сообщенія съ другими своихъ мыслей. Польза его толь велика, коль далече нынѣ простираются произшедшія отъ него въ обществѣ человѣческомъ знанія, которыя весьма бы тѣсно ограничены были, естьли бы каждой человѣкъ воображенныя себѣ способомъ чувствъ понятія только въ собственномъ своемъ умѣ содержалъ сокровенны. Когда къ сооруженію какой либо махины приготовленныя части лежатъ особливо и ни которая опредѣленнаго себѣ дѣйствія другой взаимно не сообщаетъ; тогда все небытіе ихъ тщетно и безполезно. Подобнымъ образомъ естьли бы каждой членъ человѣческаго рода не могъ изъяснить своихъ понятій другому; то бы не токмо лишены мы были сего

11

НАСТАВЛЕНІЕ I
О
ЧЕЛОВѢЧЕСКОМЪ СЛОВѢ ВООБЩЕ.

ГЛАВА I.
О голосѣ.

§ 1.

По благороднѣйшемъ дарованіи, которымъ человѣкъ протчихъ животныхъ превосходитъ, то есть правителѣ нашихъ дѣйствій разумѣ, первѣйшее есть слово, данное ему для сообщенія съ другими своихъ мыслей. Польза его толь велика, коль далече нынѣ простираются произшедшія отъ него въ обществѣ человѣческомъ знанія, которыя весьма бы тѣсно ограничены были, естьли бы каждой человѣкъ воображенныя себѣ способомъ чувствъ понятія только въ собственномъ своемъ умѣ содержалъ сокровенны. Когда къ сооруженію какой либо махины приготовленныя части лежатъ особливо и ни которая опредѣленнаго себѣ дѣйствія другой взаимно не сообщаетъ; тогда все небытіе ихъ тщетно и безполезно. Подобнымъ образомъ естьли бы каждой членъ человѣческаго рода не могъ изъяснить своихъ понятій другому; то бы не токмо лишены мы были сего

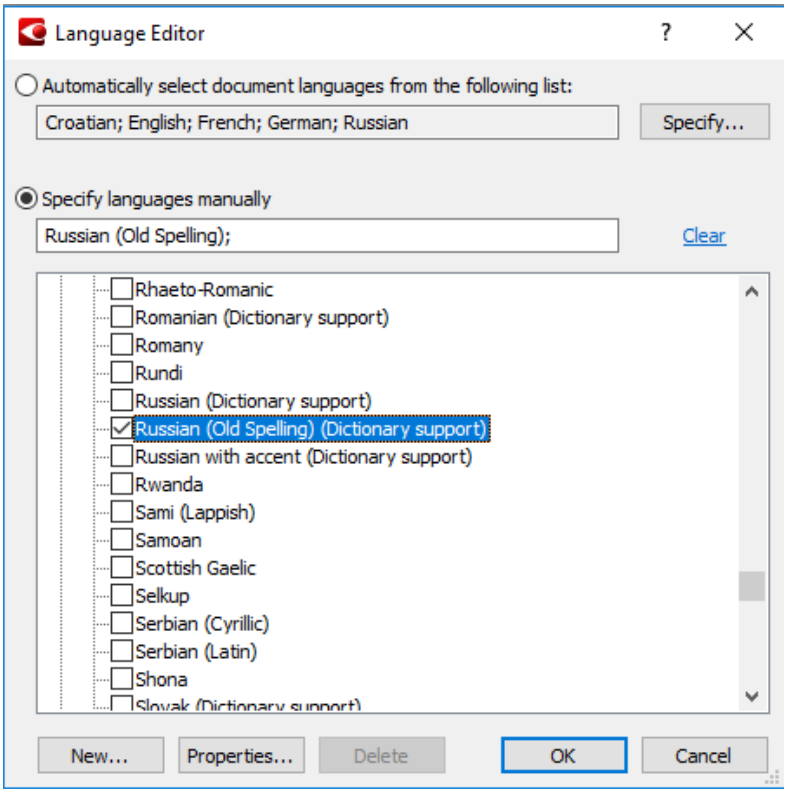
Slika 14. *Ruska gramatika* u slikovnom i tekstualnom obliku

Izvor: http://elib.gnpbu.ru/text/lomonosov_rossiyskaya-grammatika_1788/go,10;fs,1/

8.1. Abbyy FineReader

Kao što je već prethodno spomenuto, Abbyy FineReader jedan je od najčešće korištenih programa za optičko prepoznavanje znakova. Program je razvila ruska firma Abbyy 1993. godine, kada je razvijen prvi omnifont sustav u Rusiji. Program je komercijalan, te ga koriste neke od najvećih tvrtki u svijetu.

Abbyy FineReader nudi optičko prepoznavanje znakova iz tekstova pisanih na 192 jezika, a ono što je bilo važno za ovo istraživanje, osim ruskog jezika, nudi i prepoznavanje starih oblika ruskog jezika. Moguće je odabrati i više jezika istovremeno, ako je tekst pisan na dva ili više jezika. Jezik prepoznavanja bira se kod samog ulaska u program, ali ga je moguće i naknadno promijeniti ili dodati još neki.

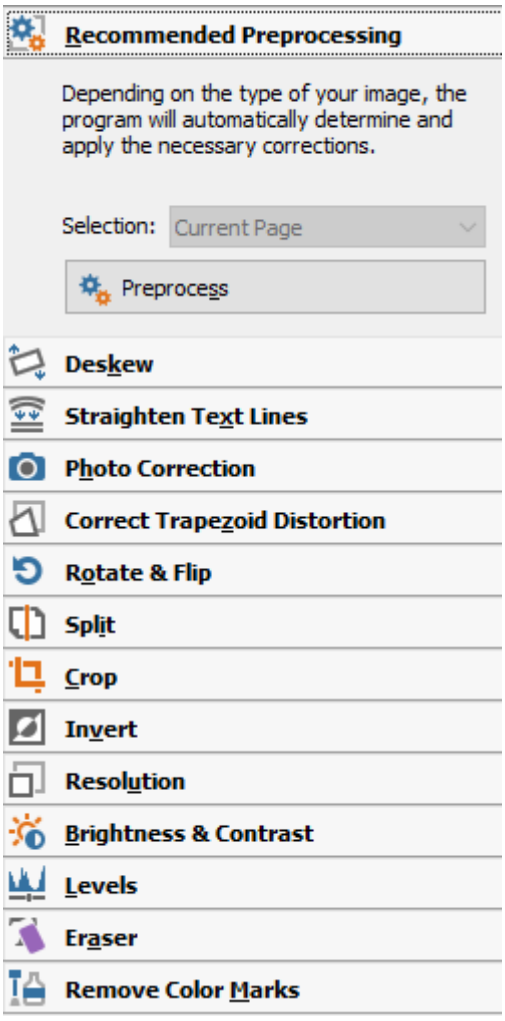


Slika 15. Odabir jezika u OCR programu Abbyy FineReader

Nakon otvaranja dokumenata koji se žele provući kroz OCR unutar Abby FineReader-a, stranice teksta moguće je urediti kako bi se dobio točniji finalni rezultat. Opcije koje program nudi su:

1. Recommended Preprocessing – program sam provodi potrebne ispravke,
2. Deskew – ispravljanje zakrivljenosti stranice, koja najčešće nastaje kod skeniranja debelih knjiga,
3. Straighten Text Lines – ispravljanje linija teksta,
4. Photo Correction – ispravljanje slike, pomoću uklanjanja geometrijske distorzije i zamućenja nastalih pomicanjem slike, smanjenja ISO buke i izbjeljivanja pozadine,
5. Correct Trapezoid Distortion – ispravljanje trapezoidne distorzije,
6. Rotate / Flip – zakretanje stranice,

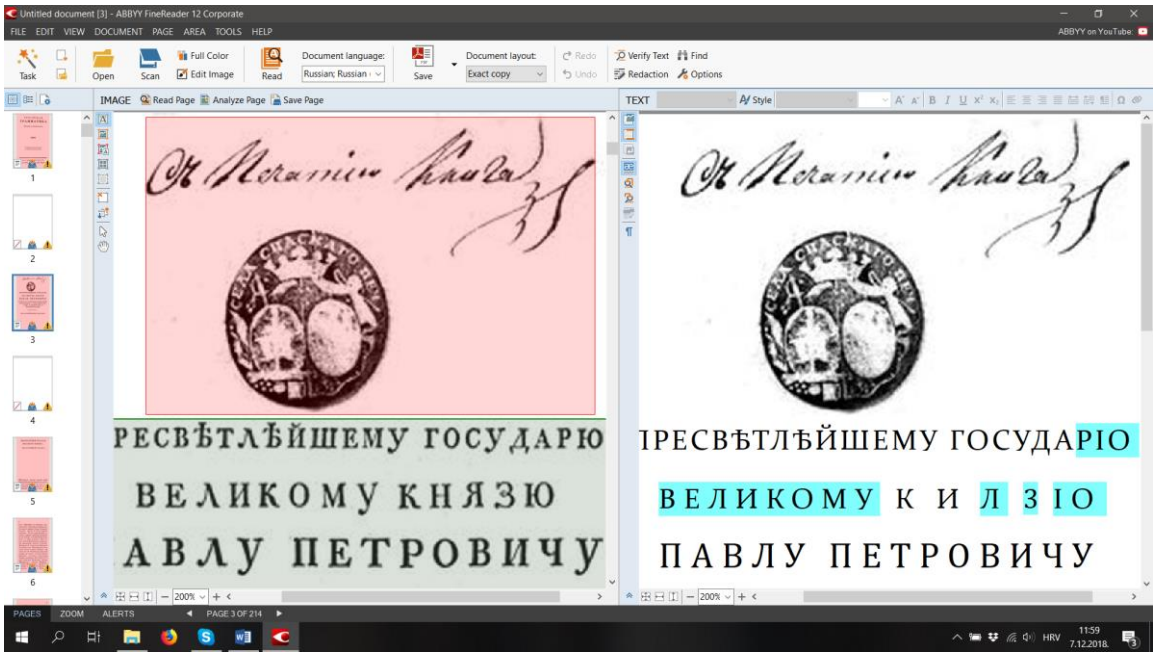
7. Split – odvajanje dijelova slike, koje se koristi kada su zajedno skenirane dvije stranice,
8. Crop – rezanje rubova slike, ako oni ne sadržavaju nikakve informacije,
9. Invert – invertiranje boje slike, koje se koristi kada je tekst svijetle boje ispisan na tamnoj podlozi,
10. Resolution – namještanje optimalne rezolucije,
11. Brightness / Contrast – podešavanje svjetline i kontrasta,
12. Levels – podešavanje razina boja slika promjenom intenziteta sjena, svjetla i polutonova,
13. Eraser – brisanje svih mrlja i buke koja bi kasnije mogla smetati kod OCR i zbuniti program,
14. Remove Color Marks – program prepoznaje i briše oznake u boji, ukoliko one prekrivaju tekst, jer bi kasnije mogle smetati kod OCR-a.



Slika 16. Opcije uređivanja skenirane slike

Sve opcije moguće je pokrenuti na trenutnoj stranici, svim neparnim, svim parnim ili na svim stranicama. Skeniranu Lomonosovljeva gramatiku nije bilo potrebno dorađivati, jer je već bila dobro obrađena. Stranice i linije teksta su poravnate, kontrast između teksta i pozadine je dobar, te nije bilo mrlja na stranicama.

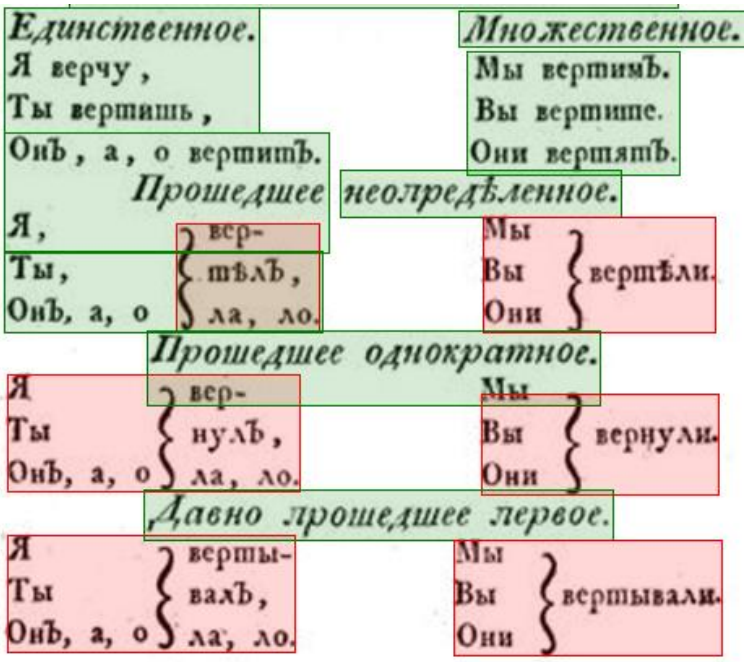
Nakon obrade trebalo je provesti analizu slike, odnosno segmentaciju. Ručna segmentacija zahtijevala bi previše vremena, stoga smo pustili program da ju automatski provede. Za analizu svih 214 stranica Lomonosovljeve gramatike programu Abbyy FineReader trebalo je nešto više od pola minute. Nakon provedene analize sve su stranice podijeljene na dijelove te pobojane različitim bojama, ovisno o vrsti informacije: tekst je pobojan zelenom bojom, slika crvenom te tablica plavom bojom. Program je učinio određene greške prilikom segmentacije. Na prvoj stranici uz sličicu nalazi se rukom pisan tekst koji je program pogrešno prepoznao kao sliku. Do te je greške došlo jer se rukopis nalazi jako blizu sličice, pa je program sve zajedno prepoznao i označio kao sliku. Za prepoznavanje rukopisa bilo bi potrebno koristiti poseban program, jer obični OCR programi nisu dovoljno napredni za prepoznavanje rukopisa.



Slika 17. Pogreška kod prepoznavanja rukopisa

Greška koja se pojavila na jako puno mjesta, bila je kod prepoznavanja drugog dijela vitičaste zagrade, koji se koristi kôd označavanja nečega što vrijedi za sve pojmove. Program je tu

zagradu i riječi uz nju prepoznao i označio kao sliku. U tom se slučaju taj dio teksta ne bi kasnije provukao kroz OCR, nego bi bio ostavljen kao slika i ne bi bio pretraživ. Osim toga, još su neki manji dijelovi teksta bili pogrešno prepoznati kao slika, ali samo na par mjesta.



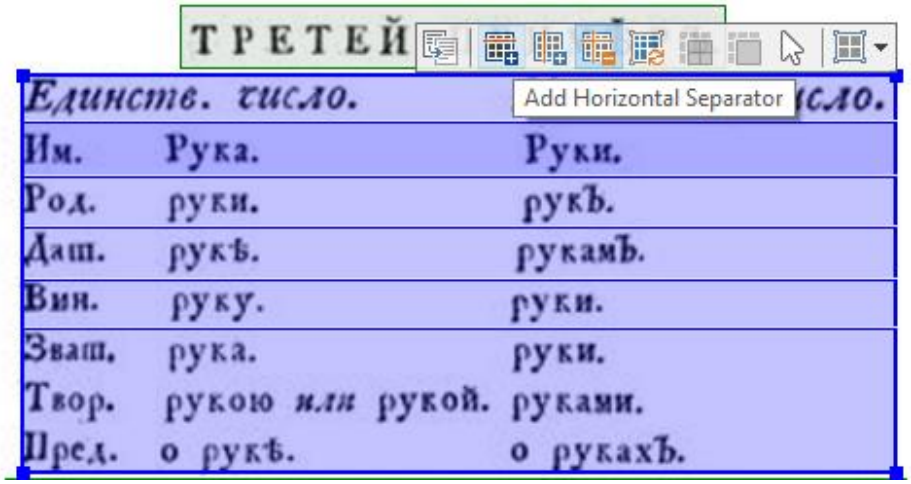
Slika 18. Pogreška kod prepoznavanja vitičaste zagrade

Knjiga sadržava mnogo tablica, no redovi i stupci u njoj nisu odvojeni pregradama, stoga je program imao problema kod analize. U nekim ih je slučajevima pravilno prepoznao kao tablicu i odvojio u redove, dok ih je u drugim prepoznao samo kao tekст.

Slika 19. Pogreška kod prepoznavanja tablice

U primjeru vidimo slučaj kada je program prvi red i desni stupac prepoznao kao tekst, dok je dva lijeva stupca prepoznao kao tablicu.

Nakon što je program prošao sve stranice i obavio segmentaciju linija, bilo je potrebno ispraviti greške. To je vrlo jednostavno napraviti, no potrebno je uložiti vrijeme da se prođu sve stranice. Pogrešno prepoznate dijelove moguće je obrisati te ručno označiti tekst, sliku ili tablicu. Kod označavanja tablice, potrebno ju je još odvojiti u stupce i redove. U nekim tablicama nije moguće odvojiti stupce, jer pojedine riječi prelaze u sljedeći stupac, pa crta koja odvaja stupce ne bi bila ravna.

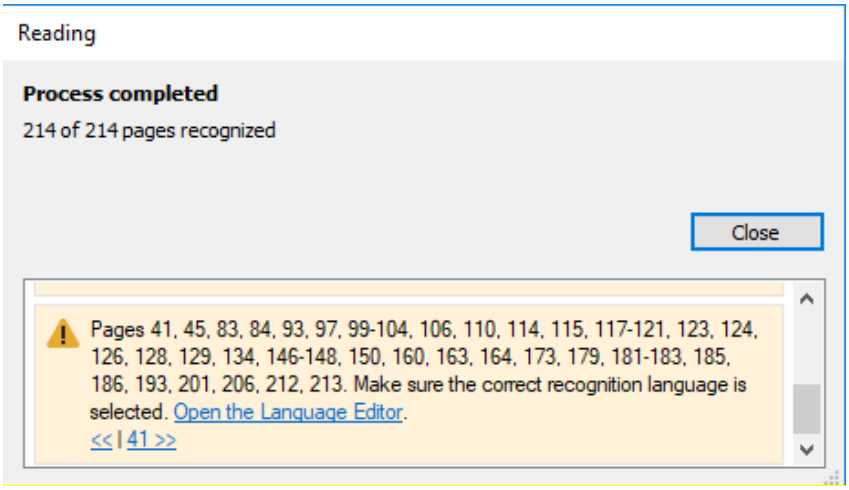


Slika 20. Odvajanje tablice u stupce i redove

Nakon segmentacije linija slijedi segmentacija riječi i znakova, te prepoznavanje znakova, koje program sâm provodi. Programu su bile potrebne 4 minute za prepoznavanje znakova sa svih 214 stranica. Abbyy FineReader, kao i drugi OCR programi, funkcionira tako da nakon što odvoji znakove unutar riječi, on ih klasificira. U tom se postupku koriste sljedeće vrste klasifikatora:

- rasterski klasifikator,
- klasifikator svojstava,
- klasifikator kontura,
- klasifikator strukture,
- klasifikator razlikovanja značajki,
- klasifikator razlikovanja strukture. (Abbyy technology portal)

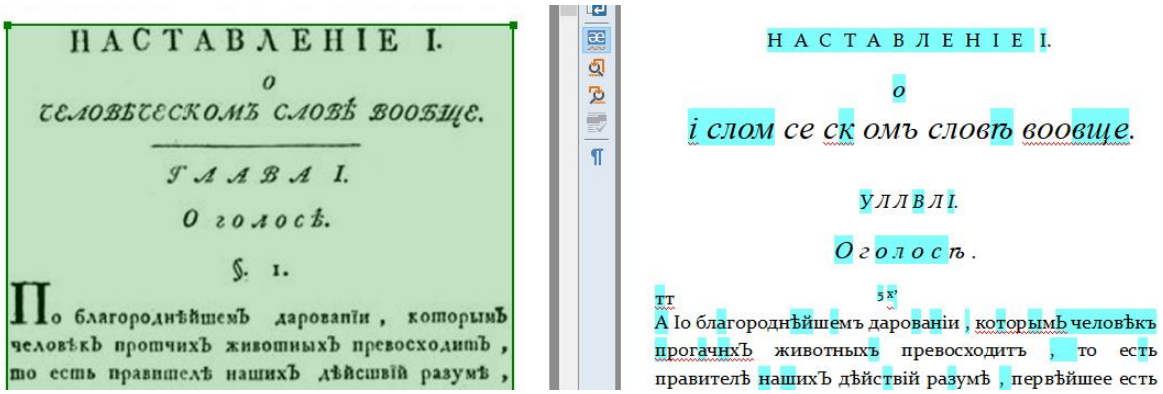
Program je upozorio i da se na nekim stranicama nalaze riječi koje nisu na jeziku namještenom za prepoznavanje, te da je potrebno točno označiti jezike za prepoznavanje. To se događalo na stranicama gdje su korišteni primjeri iz drugih jezika, pa ih program nije mogao prepoznati, jer je za jezik prepoznavanja bio postavljen ruski. Na tim stranicama mogu se očekivati pogreške kod prepoznavanja riječi.



Slika 21. Javljanje greške zbog pojavljivanja riječi na drugim jezicima

Nakon tog postupka, na desnoj strani dobiva se tekst kojeg je program prepoznao. Prepoznati tekst moguće je proći i ispraviti sve greške, a program će znakove za koje nije u potpunosti siguran i za koje sumnja da su možda pogrešno prepoznati, označiti tirkiznom bojom. Greške možemo ispravljati tako da sami mišem kliknemo na pogrešno prepoznati znak na desnoj stranici ili da nas program automatski vodi kroz znakove za koje sumnja da nisu točni, te tako za svaki označimo da je točan ili da unesemo točan. Taj se postupak naziva verificiranje teksta (engl. verify text). Postupak ispravljanja grešaka zahtijeva jako puno vremena i visok stupanj koncentracije, stoga se koristi samo kada je potrebno dobiti isključivo točan tekst.

Moguće je primijetiti da Abbyy FineReader loše prepoznaje znakove iz naslova poglavlja kada je tekst pisan drugačijim fontom, te kurzivom. Osim toga, program ima problema i s prepoznavanjem početnih slova odjeljka kada su povećana.

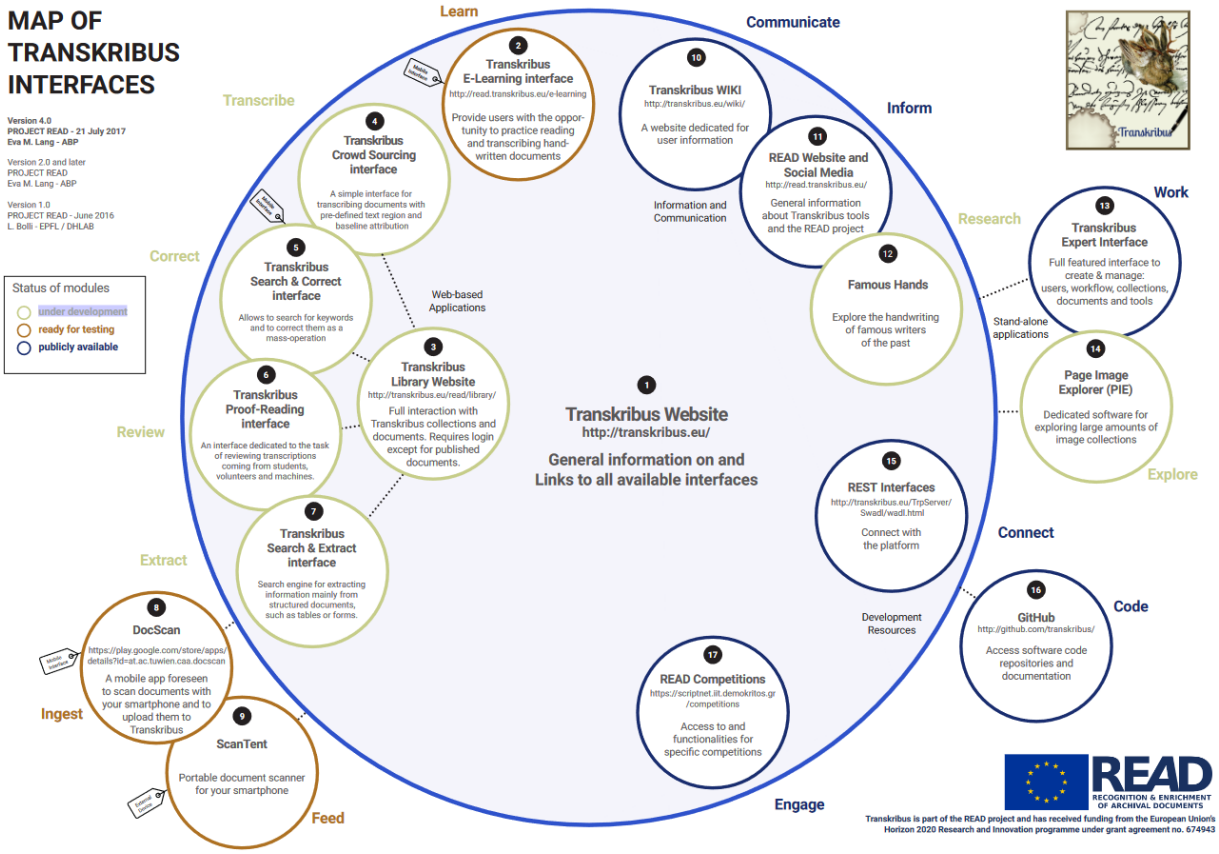


Slika 22. Greške kod prepoznavanja znakova pisanih drugačijim fontom

Abby FineReader ima opciju prepoznavanja znakova uz pomoć treniranja na bazi uzoraka i to se koristi kada je tekst pisan ukrasnim fontom, kada tekst sadrži neobične znakove, poput matematičkih simbola ili kada se radi o dokumentu loše kvalitete koji se sastoji od više od 100 stranica.

8.2. Transkribus

Transkribus je platforma za automatsko prepoznavanje, transkripciju i pretraživanje povijesnih dokumenata. To je dio projekta READ (engl. Recognition and Enrichment of Archival Documents), financiranog od strane Europske Unije. Projekt READ bavi se istraživanjem prepoznavanja rukopisa (engl Handwritten Text Recognition – HTR), osobito povijesnih djela. Cilj projekta je ubrzavanje i olakšavanje procesa prepoznavanja znakova iz povijesnih tekstova, osobito kada se radi o ogromnim količinama teksta. Projekt je u potpunosti otvorenog karaktera, te su sve publikacije, istraživanja i softveri dostupni svima online. Transkribus se sastoji od ekspertnog alata (Transkribus), web sučelja (http://transkribus.eu/) i nekoliko usluga u oblaku. Neki njegovi dijelovi još su uvijek u razvoju ili ih je moguće i testirati.



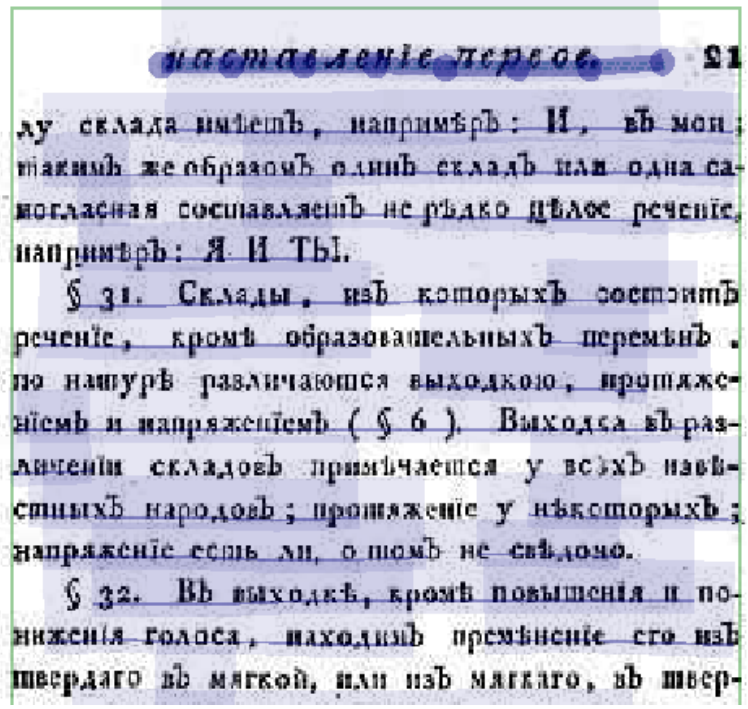
Slika 23. Mapa sučelja unutar Transkribusa

Izvor: https://read.transkribus.eu/wp-content/uploads/2017/07/Interfaces_Map_v4.0.pdf

Transkribus je prvenstveno namijenjen korisnicima koji se bave digitalizacijom tiskanih ili rukom pisanih dokumenata, odnosno znanstvenicima humanističkih znanosti, arhivarima i informatičarima. Svim korisnicima dostupna su detaljna uputstva za korištenje Transkribus alata te video upute.

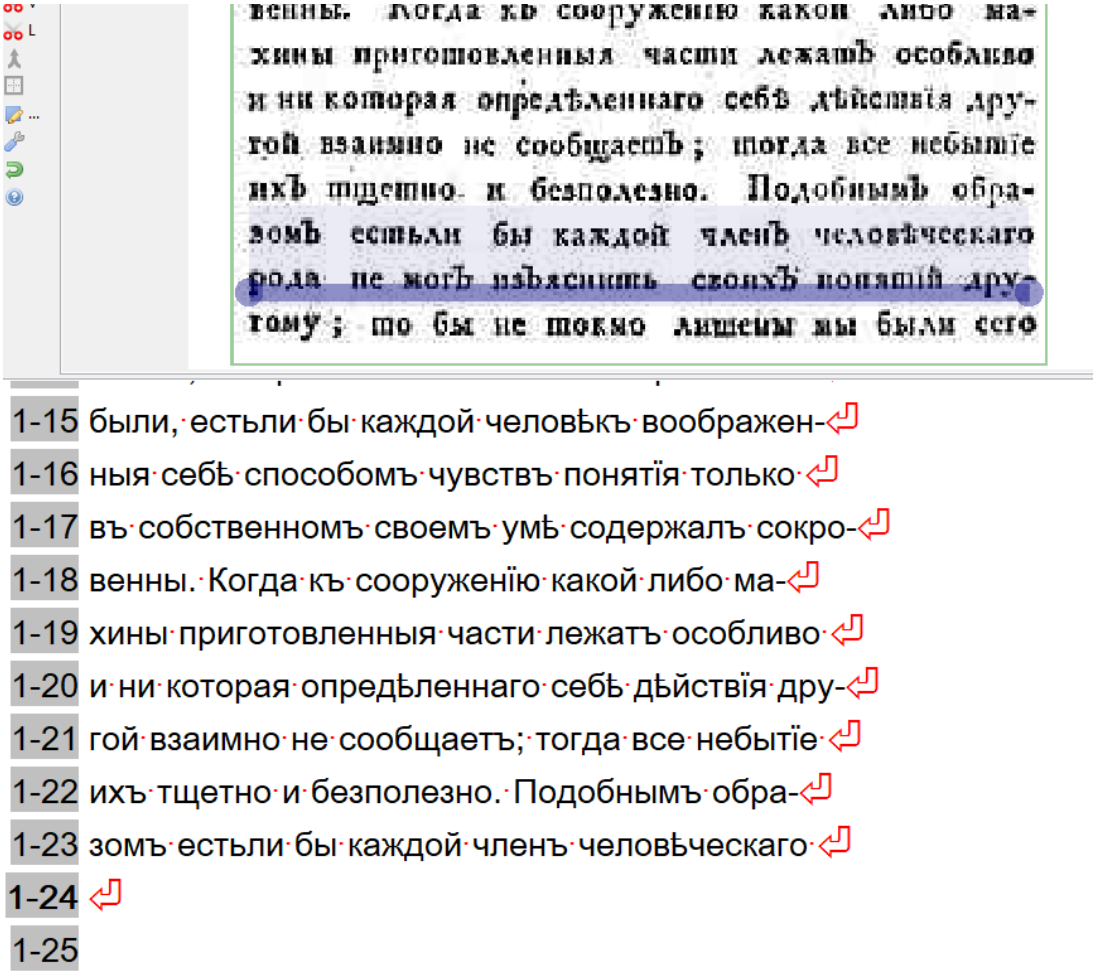
Kako bi krenuli s korištenjem Transkribusa, prvo je bilo potrebno registrirati se na web stranici te preuzeti i instalirati program. Svaki put prilikom pokretanja programa potrebno je ponovno ući u sustav. Budući da smo u istraživanju koristili samo Lomonosovljevu gramatiku koja je pisana strojem, a ne rukom, za treniranje smo koristili samo prvih 30 stranica knjige. Kada bi se radilo o većem opusu i o rukom pisanim tekstovima, bila bi potrebna puno veća količina teksta. Nakon toga, trebalo je provesti analizu stranica, odnosno podijeliti tekst u odvojene linije. Na prvih par stranica pustili smo da program automatski provede segmentaciju, no već na šestoj stranici, na kojoj se nalazi puno više teksta nego na prethodnim, program je počeo preskakati pojedina slova i cijele redove. Zato je bilo

jednostavnije ručno označiti linije dijelova stranice. Prvo je potrebno označiti radi li se o tekstu ili tablici. Unutar teksta potrebno je povući linije svakog reda, a u tablici podijeliti redove i retke. Taj proces zahtijevao je dosta vremena, te bi za veći broj stranica to bio uistinu dugotrajan proces.



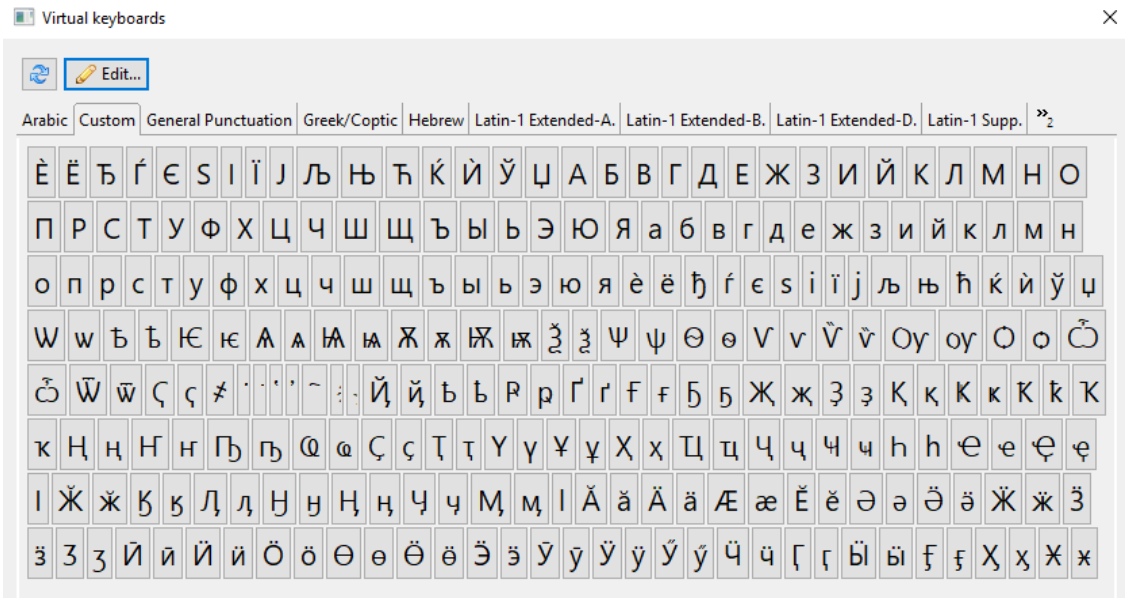
Slika 24. Pogrešna segmentacija unutar Transkribusa

Zatim je trebalo unijeti transkripciju svakog retka, kako bi program imao potpuno točan tekst za treniranje modela koji će se kasnije koristiti za prepoznavanje znakova. U Transkribusu se pritiskom na redak koji se trenutačno upisuje, automatski označava taj isti redak u originalnom tekstu kako bi se spriječila greška upisivanja pogrešnog retka. Originalni tekst moguće je i povećati kada je potrebno prepisati sitna slova, ili smanjiti, kako bi se vidio tekst cijelog retka. Budući da je Lomonosovljeva gramatika stara knjiga i otisak nije najbolji, kod prevelikog povećavanja teksta gubi se oštrina i definicija znakova, pa je ponekad teško pročitati tekst sa slike.



Slika 25. Transliteracija unutar Transkribusa

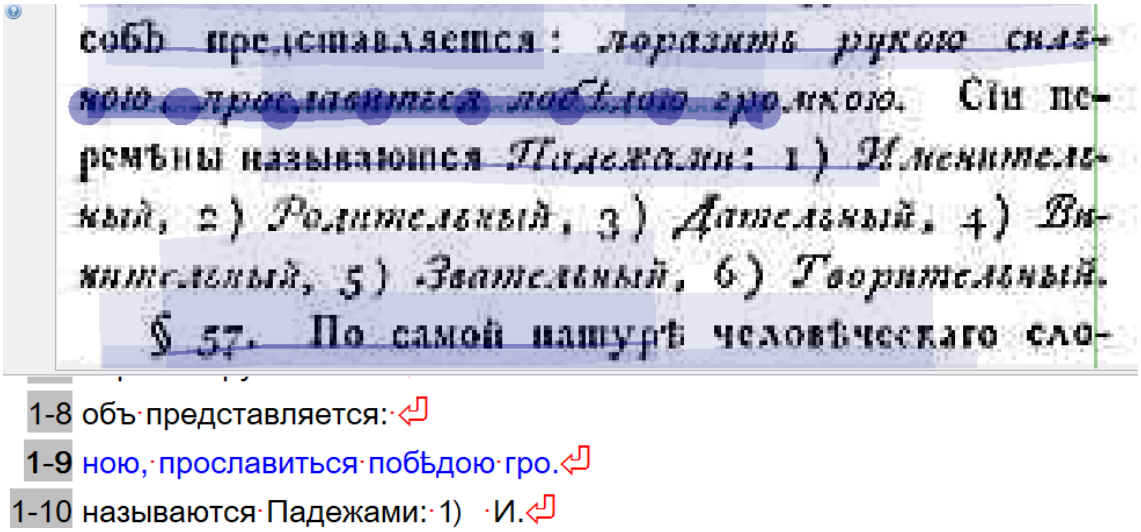
Budući da je Transkribus posebno namijenjen za prepoznavanje znakova iz povijesnih tekstova u kojima se često koriste stara slova koja se danas više ne koriste, za transliteraciju se nudi posebna virtualna tipkovnica. U njoj se već nalaze jezici poput arapskog, grčkog, hebrejskog i latinskog, ali je moguće napraviti i tipkovnicu prilagođenu našim potrebama u koju ćemo unijeti slova i znakove koji su nam potrebni. Slova i znakovi se dodaju upisujući njihov unikod, a kako se ne bi morao svaki kôd posebno upisivati, unesen je raspon znakova korištenih u ruskom i staroslavenskom pismu: U+0400-U+04FF.



Slika 26. Virtualna tipkovnica

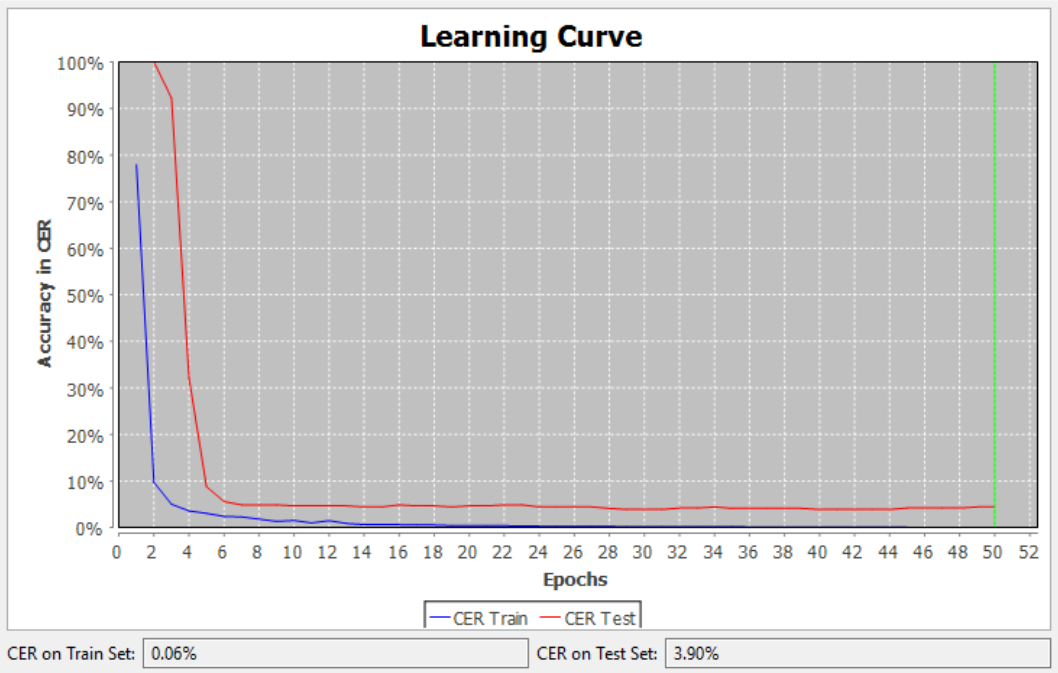
Kao što je prethodno spomenuto, *Ruska gramatika* već postoji u digitalnom obliku, tako da ju nije bilo potrebno transliterirati, odnosno ručno ju prepisivati, ali je bilo potrebno ispraviti pojedine greške, te kopirati redak po redak.

Nakon što je prvih 30 stranica bilo segmentirano i transliterirani tekst unesen, bilo je potrebno kontaktirati s djelatnicima Transkribusa kako bi njihov tim stručnjaka krenuo s treniranjem modela. Već sljedeći dan model je bio spreman za korištenje. Isproban je na sljedećoj stranici *Gramatike*. Za početak je bilo potrebno ponovno provesti analizu stranice gdje smo se susreli s istim problemom kao i na samome početku: program je preskakao pojedine riječi ili cijele retke teksta. Pretpostavljeno je da će to uzrokovati probleme kod sljedećeg koraka, jer će kod prepoznavanja znakova program tako preskočiti dijelove teksta koji nisu prepoznati u prethodnom koraku. To se i dogodilo. Rezultat takvog prepoznavanja znakova naravno ne bi bio zadovoljavajući, jer bi u tekstu nedostajali cijeli dijelovi, te bi naknadno dodavanje dijelova koji nedostaju, zahtijevalo više vremena nego da se cijeli tekst ručno prepiše.



Slika 27. Greška prilikom analize stranice i prepoznavanja znakova

Za dobivanje preciznijeg rezultata, bilo je potrebno ponovno ručno provesti analizu stranice. Nakon analize, ponovno smo pokrenuli prepoznavanje znakova iz teksta. Prilikom pokretanja prepoznavanja moguće je odabrati metodu prepoznavanja: HTR (CITIab), koja je preporučena i najefektivnija, ili OCR iz Abbyy FineReadera koji je već korišten u prethodnom poglavlju. Nakon odabira HTR metode, bilo je potrebno odabrati i model. Pritiskom na tu tipku otvara se novi prozor u kojem su ponuđene različite metode za različite uzorke, a između ostalog i istrenirani model. Odabirom izrađenog modela pojavljuje se kratak opis, te krivulja učenja modela.



Slika 28. Krivulja učenja HTR modela

Na dijagramu x-os predstavlja točnost u CER-u (engl. Character Error Rate), odnosno postotak znakova koje je program pogrešno prepisao. Krivulja će na početku krenuti s vrlo visoke razine, te će se postupno snižavati kako će trening napredovati i model se poboljšavati. Y-os označava epohe na koje je proces treniranja podijeljen, te se nakon svake epohe radi procjena. Primjer procesa treniranja podijeljen je u 50 takvih epoha. Na dijagramu se može vidjeti plava i crvena krivulju. Plava krivulja predstavlja napredak treninga, a crvena napredak procjena na testnom setu. Program će se prvo trenirati na setu za treniranje, te nakon toga testirati na setu za testiranje. Ispod dijagrama nalazi se postotak pogrešno prepoznatih znakova tijekom treniranja i testiranja. Na izrađenom primjeru vidi se kako je model vrlo točan, jer je prvi postotak pogrešaka prilikom treniranja samo 0,06%, dok je postotak pogrešaka kod testiranja 3,9% (READ). Nakon pokretanja modela već je na prvi pogled moguće primijetiti da model doista uspješno funkcionira. Nakon što je proces prepoznavanja završen, dokument je moguće prikazati (engl. export) u formatima pdf, tei, docx ili u Excelu, ako se radi o tablici, tako da na mail adresu bude poslan link s kojeg je moguće preuzeti datoteku.

8.3. Rezultati

Točnost dobivenih rezultata optičkog prepoznavanja znakova pomoću programa Abbyy FineReader i Transkribus izračunata je na 30 stranica Lomonosovljeve *Ruske gramatike*: od 31. do 60. stranice, jer je prvih 30 stranica korišteno za treniranje modela u Transkribusu. Za računanje točnosti korišteni su ISRI analitički alati razvoj kojih je pokrenut na Institutu za informacijske znanosti na Sveučilištu Nevada u Las Vegasu, SAD. Moguće ih je besplatno preuzeti, a pokrenuti samo u operativnom sustavu Linux. Za dobivanje rezultata uspoređeni su tekstovi dobiveni od Abbyy FineReader-a i Transkribusa s potpuno točnim tekstom. Svaku stranicu teksta potrebno je pohraniti u .txt formatu kao zasebnu utf-8 datoteku. Stranica točnog teksta sa stranicom teksta dobivenog optičkim prepoznavanjem teksta uspoređuje se naredbom *accuracy*, te je u istraženom primjeru ta naredba izgledala ovako:

Accuracy tocan31.txt abbyy31.txt a31

Naredbom *accuracy* pokreće se mjerenje točnosti, *tocan31.txt* označava potpuno točan tekst s 31. stranice, a *abbyy31.txt* označava tekst dobiven korištenjem Abbyy FinReader OCR-a, te se tako stvara datoteka *a31* u kojoj se pohranjuje izvještaj o usporedbi tih dviju stranica. Isti je postupak primijenjen i na ostalih 29 stranica. Za dobivanje točnosti programa Transkribus korištena je ista naredba *accuracy*:

Accuracy tocan31.txt transkribus31.txt t31

Ta je naredba također primijenjena i na ostalih 29 stranica. Tako su dobiveni izvještaji točnosti za svaku stranicu i za svaki od korištenih programa. Kako bi alat izmjerio ukupnu točnost trebalo je objediniti sve te datoteke. Za to se koristi naredba *accsum*. Kod računanja točnosti svih 30 stranica dobivenih Abbyy FineReaderom cijela naredba izgledala je ovako:

Accsum a31 a32 a33 a34 a35 a36 a37 a38 a39 a40 a41 a42 a43 a44 a45 a46 a47 a48 a49 a50 a51 a52 a53 a54 a55 a56 a57 a58 a59 a60 > abbyy

Za računanje točnosti Transkribusa potrebno je objediniti sve datoteke, dobivene naredbom *accuracy*:

Accsum t31 t32 t33 t34 t35 t36 t37 t38 t39 t40 t41 t42 t43 t44 t45 t46 t47 t48 t49 t50 t51 t52 t53 t54 t55 t56 t57 t58 t59 t60 > transkribus

Nakon izvršenja tih naredbi nastale su datoteke *abbyy* i *transkribus*, u kojima se nalazio izvještaj točnosti svih 30 stranica. Izvještaj točnosti sastoji se od 6 dijelova. U prvom je dijelu napisano od koliko znakova se sastoji potpuno točan tekst, broj grešaka u tekstu koji je provjeravan, te točnost tog teksta. U drugom se dijelu navodi broj odbijenih znakova, sumnjivih i pogrešnih oznaka, te se pokazuje koliko bi porasla točnost kada bi se sumnjivi znakovi ispravili. U trećem se dijelu navodi broj označenih i neoznačenih pogrešaka. Četvrti dio prikazuje točnost prema klasi znakova. U petom se dijelu navode sve pogreške unutar teksta, a zadnji dio nudi popis svih znakova potpuno točnog teksta, s brojem pojavljivanja, brojem pogrešaka u testiranom tekstu, te točnošću za taj znak.

Izvještaj točnosti programa Abby FineReader pokazao je da je točnost dobivenog teksta samo 56,48%:

33.509 znakova
14.584 pogrešaka
56,48% točnost

Od ukupno 33.509 znakova, njih čak 14.584 pogrešno je prepoznato. U izvještaju je moguće pročitati točnost svakog pojedinačnog znaka:

Tablica 4. Izvještaj točnosti programa Abbyy Fine Reader

Count	Missed	% Right		Count	Missed	% Right		Count	Missed	% Right	
157	53	66.24	{<\n>}	9	6	33.33	{I}	77	23	70.13	{§}
4697	670	85.74	{ }	1	1	0.00	{N}	11	11	0.00	{◌̀}
1	1	0.00	{!}	3	3	0.00	{V}	23	23	0.00	{◌́}
5	4	20.00	{(}	1	1	0.00	{[}	19	19	0.00	{◌̂}
82	35	57.32	{)}	14	14	0.00	{a}	1	1	0.00	{ξ}
872	165	81.08	{,}	2	2	0.00	{b}	1	1	0.00	{ψ}
42	24	42.86	{-}	6	6	0.00	{c}	14	14	0.00	{Ї}
330	77	76.67	{.}	4	4	0.00	{d}	35	23	34.29	{А}
23	16	30.43	{0}	12	12	0.00	{e}	30	20	33.33	{Б}
62	42	32.26	{1}	5	5	0.00	{g}	52	23	55.77	{В}
32	20	37.50	{2}	3	3	0.00	{h}	32	15	53.12	{Г}
36	18	50.00	{3}	11	11	0.00	{i}	21	3	85.71	{Д}
32	13	59.38	{4}	5	5	0.00	{l}	53	19	64.15	{Е}
33	4	87.88	{5}	1	1	0.00	{m}	12	1	91.67	{Ж}
24	5	79.17	{6}	5	5	0.00	{n}	15	10	33.33	{З}
24	7	70.83	{7}	2	2	0.00	{o}	69	26	62.32	{И}
27	10	62.96	{8}	4	4	0.00	{r}	7	6	14.29	{Й}
25	8	68.00	{9}	5	5	0.00	{s}	21	9	57.14	{К}
129	27	79.07	{:}	5	5	0.00	{t}	23	6	73.91	{Л}
122	30	75.41	{;}	7	7	0.00	{u}	28	10	64.29	{М}
1	1	0.00	{?}	2	2	0.00	{v}	51	18	64.71	{Н}
								41	13	68.29	{О}

59	34	42.37	{П}	363	130	64.19	{б}	552	135	75.54	{у}
59	15	74.58	{Р}	1042	248	76.20	{в}	7	4	42.86	{ф}
77	27	64.94	{С}	515	123	76.12	{г}	310	80	74.19	{х}
30	4	86.67	{Т}	751	233	68.97	{д}	56	16	71.43	{ц}
13	10	23.08	{У}	1788	457	74.44	{е}	368	127	65.49	{ч}
11	10	9.09	{Ф}	260	61	76.54	{ж}	136	45	66.91	{ш}
8	6	25.00	{Х}	391	95	75.70	{з}	87	20	77.01	{щ}
5	1	80.00	{Ц}	1418	404	71.51	{и}	1323	1128	14.74	{ъ}
13	7	46.15	{Ч}	294	89	69.73	{й}	594	134	77.44	{ы}
9	4	55.56	{Ш}	791	167	78.89	{к}	413	124	69.98	{ь}
2	0	100.00	{Щ}	966	199	79.40	{л}	278	61	78.06	{ю}
9	1	88.89	{Ъ}	746	212	71.58	{м}	749	156	79.17	{я}
5	0	100.00	{Ы}	1735	342	80.29	{н}	19	13	31.58	{і}
5	0	100.00	{Ь}	2552	498	80.49	{о}	398	398	0.00	{ї}
1	0	100.00	{Э}	633	230	63.67	{п}	13	9	30.77	{Ѣ}
10	4	60.00	{Ю}	1108	200	81.95	{р}	476	153	67.86	{ѣ}
19	3	84.21	{Я}	1396	265	81.02	{с}	1	0	100.00	{Ѳ}
1632	300	81.62	{а}	1483	426	71.27	{т}	30	0	100.00	{<FEFF>}

Može se primijetiti kako je program najviše problema imao s prepoznavanjem posebnih znakova, poput grčkih slova „ξ“ i „ψ“ i slova latinice, no to je i za očekivati, jer u ruskom suvremenom alfabetu ta slova ne postoje. Prilično niska točnost je i kod prepoznavanja nekih znamenki. Prosječna točnost velikih slova ćirilice je 60,49 %, a malih 68,12%. Slova sa stopostotnom točnošću su velika slova „Щ“, „Ы“, „Ь“ i „Э“, dok su i veliko i malo slovo „ї“ svaki puta bili pogrešno prepoznati. Razlog tomu jest to, da slovo „ї“ ne postoji u današnjem ruskom alfabetu.

Izvještaj točnosti Transkribusa pokazao je da je točnost teksta dobivenog Transkribusom 97,60%:

33.509 znakova
805 pogrešaka
97,60% točnost

I ovdje je posebno zanimljiva točnost svakog pojedinačnog znaka.

Tablica 5. Izvještaj točnosti programa Transkribus

Count	Missed	%	Right								
157	144	8.28	{<\n>}	23	2	91.30	{0}	25	1	96.00	{9}
4697	21	99.55	{ }	62	2	96.77	{1}	129	1	99.22	{:}
1	1	0.00	{!}	32	3	90.62	{2}	122	1	99.18	{;}
5	2	60.00	{(}	36	1	97.22	{3}	1	1	0.00	{?}
82	0	100.00	{)}	32	6	81.25	{4}	9	6	33.33	{I}
872	12	98.62	{,}	33	5	84.85	{5}	1	1	0.00	{N}
42	15	64.29	{-}	24	1	95.83	{6}	3	2	33.33	{V}
330	6	98.18	{.}	24	1	95.83	{7}	1	1	0.00	{[}
				27	5	81.48	{8}	1	1	0.00	{]}

14	12	14.29	{a}	15	2	86.67	{З}	260	0	100.00	{ж}
2	2	0.00	{b}	69	5	92.75	{И}	391	8	97.95	{з}
6	6	0.00	{c}	7	7	0.00	{Й}	1418	16	98.87	{и}
4	3	25.00	{d}	21	1	95.24	{К}	294	1	99.66	{й}
12	6	50.00	{e}	23	2	91.30	{Л}	791	6	99.24	{к}
5	4	20.00	{g}	28	0	100.00	{М}	966	9	99.07	{л}
3	2	33.33	{h}	51	6	88.24	{Н}	746	3	99.60	{м}
11	10	9.09	{i}	41	2	95.12	{О}	1735	25	98.56	{н}
5	5	0.00	{l}	59	2	96.61	{П}	2552	13	99.49	{о}
1	1	0.00	{m}	59	2	96.61	{Р}	633	2	99.68	{п}
5	4	20.00	{n}	77	1	98.70	{С}	1108	6	99.46	{р}
2	2	0.00	{o}	30	7	76.67	{Т}	1396	12	99.14	{с}
4	1	75.00	{r}	13	4	69.23	{У}	1483	13	99.12	{т}
5	5	0.00	{s}	11	5	54.55	{Ф}	552	2	99.64	{у}
5	4	20.00	{t}	8	1	87.50	{Х}	7	2	71.43	{ф}
7	7	0.00	{u}	5	0	100.00	{Ц}	310	2	99.35	{х}
2	2	0.00	{v}	13	3	76.92	{Ч}	56	4	92.86	{ц}
77	0	100.00	{§}	9	0	100.00	{Ш}	368	11	97.01	{ч}
11	11	0.00	{◌̀}	2	1	50.00	{Щ}	136	23	83.09	{ш}
23	23	0.00	{◌́}	9	9	0.00	{Ъ}	87	1	98.85	{щ}
19	19	0.00	{◌̂}	5	0	100.00	{Ы}	1323	4	99.70	{ъ}
1	0	100.00	{ξ}	5	5	0.00	{Ь}	594	3	99.49	{ы}
1	1	0.00	{ψ}	1	1	0.00	{Э}	413	5	98.79	{ь}
14	13	7.14	{Ï}	10	1	90.00	{Ю}	278	4	98.56	{ю}
35	7	80.00	{А}	19	0	100.00	{Я}	749	4	99.47	{я}
30	11	63.33	{Б}	1632	7	99.57	{а}	19	19	0.00	{ї}
52	4	92.31	{В}	363	3	99.17	{б}	398	1	99.75	{ї}
32	4	87.50	{Г}	1042	6	99.42	{в}	13	13	0.00	{Ђ}
21	0	100.00	{Д}	515	9	98.25	{г}	476	4	99.16	{ђ}
53	7	86.79	{Е}	751	2	99.73	{д}	1	0	100.00	{Ѳ}
12	3	75.00	{Ж}	1788	10	99.44	{е}	30	0	100.00	{<FEFF>}

Iako se i ovdje vidi kako slova latinice imaju dosta nisku točnost, neka slova su ipak bila prepoznata. Da je na prvim stranicama teksta, na kojim je bio treniran model, bilo više slova latinice, točnost bi bila još i veća. Zato je važno da se za treniranja modela u Transkribusu uzme uzorak u kojem se pojavljuju svi znakovi koji će se pojavljivati u tekstu na kojem ćemo koristiti model za prepoznavanje. Prosječna točnost velikih slova ćirilice ovdje je 73,88%, a malih 94,64%. Mala slova se puno češće pojavljuju nego velika, stoga je i njihova točnost veća, jer je Transkribus imao veći uzorak za treniranje modela. Slova sa stopostotnom točnošću su „Д“, „М“, „Ц“, „Ш“, „Ы“, „Я“ i „ж“, dok slova „Й“, „Ъ“, „Ь“, „Э“ i „Ђ“ nisu niti jednom točno prepoznata.

Usporedbom rezultata Abbyy FineReadera i Transkribusa može se zaključiti kako je Transkribus puno bolji u prepoznavanju znakova. U istraženom i analiziranom slučaju običan OCR program poput Abbyy FineReadera nije dovoljno uspješan u prepoznavanju znakova. Točnost bi bila veća kada bi se unutar programa koristila mogućnost treniranja. Za digitalizaciju starih ruskih knjiga zato je puno bolje koristiti programe poput Transkribusa koji se uz pomoć istreniranog modela može koristiti na svim knjigama koje su pisane istim fontom i na istom jeziku. Iako je segmentacija i transliteracija unutar Transkribusa zahtijevala nešto više vremena nego prepravljanje grešaka kod segmentacije u Abbyy FineReaderu, rezultati su puno točniji i naknadno ispravljanje grešaka je minimalno. Također, kako raste broj obrađenih stranica tako Transkribus ima sve povoljniji odnos korektno automatski prepoznatoga teksta i utrošenog vremena za korigiranje.

9. Dostupnost ruskih knjiga u digitalnom obliku

Mnoge ruske institucije rade na tome da sve svoje gradivo prebace u digitalni oblik kako bi ono bilo očuvano i šire dostupno. Događaj koji je skrenuo veliku pažnju na važnost očuvanja knjiga, bio je požar u knjižnici Instituta za znanstvene informacije o društvenim znanostima Ruske akademije znanosti (rus. Институт научной информации по общественным наукам, РАН). Do požara je došlo 30. siječnja 2015. godine kada je vatra zahvatila gotovo 2 tisuće kvadratnih metara knjižnice. Kao posljedica vatre, a i vode kojom su pokušali ugasiti vatru, uništeno je oko 5 milijuna i 700 tisuća knjiga, odnosno 20% knjižničnog fonda. Neke od njih bile su pripremljene za otpis ili su imale kopiju u nekoj drugo knjižnici, ali više od 2 milijuna knjiga bili su jedinstveni primjerci (INION RAN).

Kako više ne bi došlo do tako velikih gubitaka u Rusiji se sve više radi na digitalizaciji, osobito starih knjiga, te se pokreću razni projekti digitalizacije. Na Internetu je moguće pronaći veliku količinu ruske literature unutar digitalnih knjižnica i na stranicama različitih projekata. Knjige je najčešće moguće čitati online ili ih je moguće preuzeti, najčešće u PDF formatu. Moguće je pronaći i transliterirane knjige, no češće se radi o skeniranim knjigama, pretraživim ili nepretraživim. Na nekim je stranicama moguće pronaći i zvučne (audio) knjige.

Projektima digitalizacije, provedenim unutar Ruske državne knjižnice (rus. Российская государственная библиотека) i Državne javne povijesne knjižnice Rusije (rus. Государственная публичная историческая библиотека России), stvoreni su fondovi

digitalnih knjiga, koje je inače moguće pronaći u njihovoj knjižnici, a nakon digitalizacije njihovo je korištenje omogućeno putem Interneta. Manji dio knjiga unutar digitalne zbirke dostupan je svima, dok je veći dio knjiga dostupan samo korisnicima tih knjižnica.

Unutar nekih projekata digitalizacije udružio se veći broj knjižnica kako bi stvorili zajednički fond. Tako su stvorene:

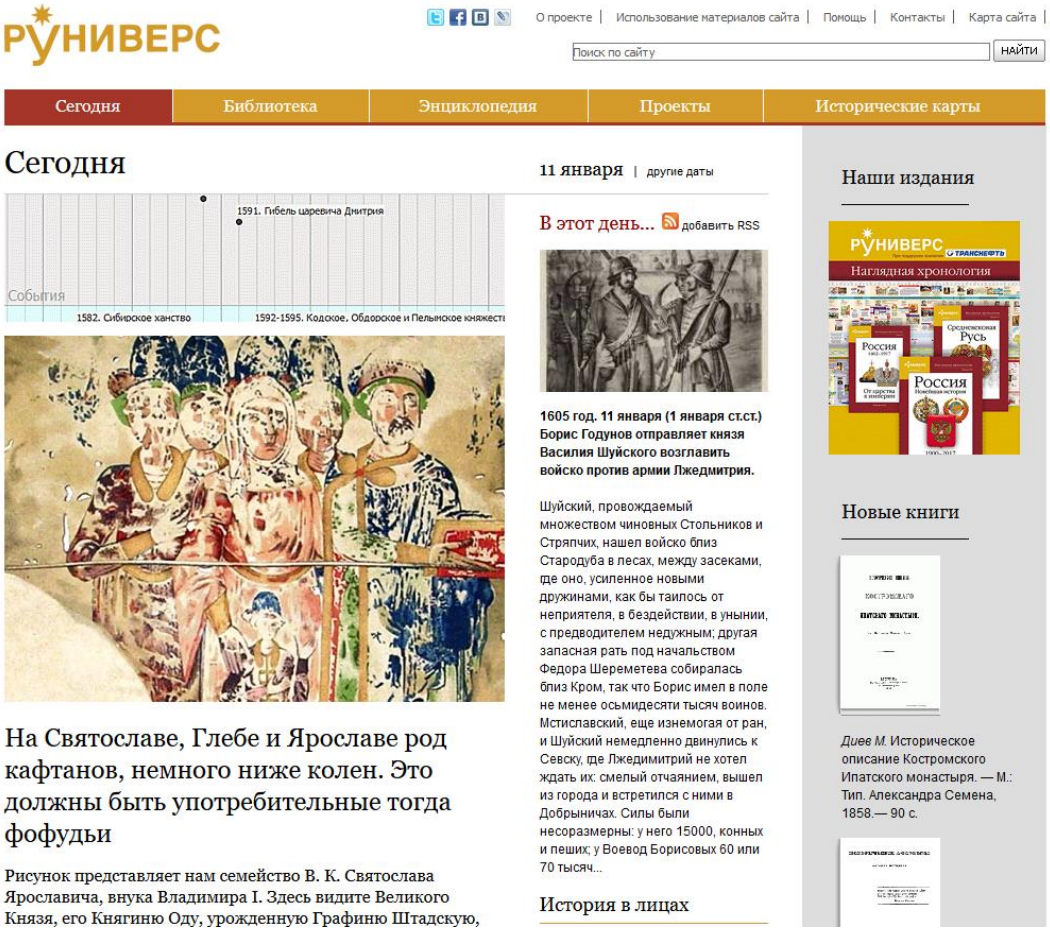
- Nacionalna elektronička knjižnica (rus. Национальная электронная библиотека: НЭБ),
- Nacionalna elektronička dječja knjižnica (rus. Национальная электронная детская библиотека),
- Runivers (rus. Руниверс),
- Elektronska knjižnica ImWerden (rus. Электронная библиотека ImWerden),
- Knjižnica Maksima Moškova (rus. Библиотека Максима Мошкова),
- Elektronska knjižnica Aldebaran (rus. Электронная библиотека Альдебаран),
- Mreža elektroničkih knjižnica Vivaldi (rus. Сеть электронных библиотек Vivaldi),
- CyberLeninka (rus. КиберЛенинка).

Projekt Nacionalne elektroničke knjižnice pokrenut je 2004. godine od strane različitih ruskih knjižnica uz potporu Ministarstva kulture Ruske Federacije. Danas ona sadržava više od 4 i pol milijuna elektroničkih dokumenata, originali kojih se nalaze u 92 različite knjižnice, muzeja, arhiva i obrazovnih ustanova. Većina knjiga je dostupna u digitalnom obliku, dok su neke još uvijek nedostupne zbog autorskih prava. Stranica sadržava i elektronički katalog ruskih knjižnica, kako bi korisnici mogli naći u kojoj se knjižnici nalaze izvornici (Nacional'naja Elektronnaja Biblioteka).

Još jedan važan projekt digitalizacije ruskih knjiga je Runivers (rus. Руниверс), neprofitni projekt posvećen ruskoj povijesti i kulturi, kojemu je cilj osigurati slobodan pristup primarnim izvorima, knjigama i tekstovima koji se nalaze u najvećim knjižnicama i državnim arhivima, a koji su do sada bili dostupni samo posjetiteljima u desetak najvećih ruskih knjižnica. Digitalizaciju podržava ruska tvrtka Transneft. Stranica sadržava faksimile[5] s više od 3 tisuće svezaka ruskih knjiga iz povijesti i filozofije, te enciklopedija, atlasa i memoara, objavljenih tijekom 19. i početkom 20. stoljeća. Na stranici je moguće pronaći i jedinstvenu zbirku s

[5] Reprodukcija koja je po obliku, dimenzijama, boji i svim pojedinostima nalik na original. Kao faksimili reproduciraju se i umnožavaju rukopisi, autografi, potpisi, crteži, grafički listovi, dokumenti, iluminacije, slike u bojama, stare i rijetke knjige, kodeksi i sl.

gotovo 4 tisuće karata i preko 20 tisuća ilustracija i fotografija. Runivers je zamišljen kao elektronička enciklopedija, u kojoj su svi materijali uneseni u zajedničku bazu podataka, a uz pomoć koje će njeni korisnici moći pronaći digitalizirane knjige, dokumente, slike i fotografije na temu koja ih zanima (Bondarenko 2012).



Slika 29. Stranica Runivers

Izvor: https://www.runivers.ru/

Još jedna digitalna knjižnica vrijedna spominjanja je InWerden. Osnovana 2000. godine, veliku pažnju posvećuje sakupljanju djelâ ruske književnosti, osobito 18. i 19. stoljeća. Stranicu održava i uređuje Andrej Nikitin-Perenski. Na stranici je moguće pronaći i neka veoma rijetka djela, kao što su knjige i časopisi izdani tijekom 20-ih i 30-ih godina 18. stoljeća. Jedna od zanimljivosti ove stranice je da se na njoj mogu naći i audio snimke u kojima autori čitaju svoja vlastita djela. ImWerden knjižnica je najveća ruska zbirka audio zapisa s autorskim pravima na Internetu. Osim audio snimki, sadržava i video snimke čitanja djela, te video snimke, intervjue i dokumentarce. Glavni formati su PDF, MP3 i AVI (Nikitin-Perenski).

Neke stranice služe za objedinjavanje digitalnih knjiga koje se nalaze na drugim internetskim stranicama kako bi se korisnicima olakšala potraga za željenom knjigom. Takva stranica je Gbooks. To je stranica koja prikuplja stare knjige i časopise objavljene do 1920. godine. Glavni cilj projekta je prikupiti i organizirati poveznice na e-knjige koje mogu biti zanimljive studentima povijesti, geografije, etnografije i drugih srodnih disciplina u jednom katalogu. Knjige su dostupne zahvaljujući većem projektu tvrtke Google, Google Books, predstavljenom 2004. godine, te zahvaljujući pomoći različitih knjižnica. Kvaliteta digitalnih knjiga nije uvijek najbolja i često se događa da neka stranica nedostaje ili redoslijed stranica nije točan, a ilustracije i karte su loše skenirane, no kvaliteta se ipak poboljšala u protekle dvije godine. To često ovisi i o standardu knjižnice koja im je ustupila materijale. Knjige je moguće pretraživati uz pomoć kataloga, prema autoru ili nazivu djela. Odabirom željene knjige, stranica prebaci korisnika na stranicu na kojoj se nalazi knjiga, najčešće u pdf formatu, ili nešto rjeđe, u DjVu. Ako je neka knjiga objavljenja više puta u više izdanja, u slučaju da je svako novo izdanje bilo proširivano, prvo se ponudi ono zadnje, no ako su sva izdanja jednaka, ponudit će se ono s najboljom kvalitetom (Gbooks).

Manji broj ruskih knjiga u digitalnom obliku moguće je pronaći i na stranicama stranih projekata, kao što su:

1. Projekt Gutenberg,
2. Europeana,
3. World digital library.

9.1. Problem autorskih prava

Uz digitalizaciju i stvaranje elektroničkih knjižnica usko je vezan i problem očuvanja prava autora i izdavača. Prema Igoru Glihi, pojam „‘autorsko pravo’ ima dvojako značenje. Njime se označuje autorsko pravo u objektivnom, ali i u subjektivnom smislu. U objektivnom smislu ima uže i šire značenje. U objektivnom užem smislu pojam autorsko pravo označuje skup pravnih pravila kojima se uređuju pravni odnosi glede intelektualnih tvorevina s književnoga, znanstvenoga i umjetničkoga područja, a u širem smislu pojam autorsko pravo obuhvaća i pravna pravila kojima se uređuju srodna (susjedna, koneksna) prava, kao što su pravo umjetnika izvođača, pravo proizvođača fonograma, pravo proizvođača videograma, pravo organizacija za radiodifuziju, pravo proizvođača baza podataka. Pojam autorsko pravo označuje i autorsko pravo u subjektivnom smislu, tj. kao

subjektivno pravo koje daje neposrednu i najveću privatnopravnu vlast glede autorskog djela“ (Gliha 2000: 5-6). Postoje knjige kod kojih su autorska prava istekla, no postoje i knjige čiji se autori ili nositelji autorskih prava mogu protiviti prijenosu rezultata njihovog intelektualnog vlasništva u elektronički oblik. Knjižnice bi trebale omogućiti neometan pristup informacijama, no istovremeno moraju voditi brigu o tome da se autorska prava ne krše. Uz nove tehnologije, djela književnosti i umjetnosti često se doživljavaju kao javno vlasništvo, no ne smijemo zaboraviti da svako od njih ima svojeg autora, bez čijeg dopuštenja se ne smije koristiti (reproducirati, distribuirati, uvoziti, prevoditi, itd.). Prema Građanskom zakoniku Ruske Federacije: „U slučaju kada knjižnica daje na korištenje kopije djela koja su zakonito ušla u civilni promet za privremeno besplatno korištenje, takva je upotreba dopuštena bez pristanka autora ili drugog nositelja autorskog prava i bez plaćanja naknade. Istodobno, digitalne kopije radova koje knjižnice daju za privremeno korištenje, uključujući i redoslijed zajedničkog korištenja knjižničnih resursa, mogu se osigurati samo u prostorima knjižnice pod uvjetom da je isključena mogućnost izrade kopija tih djela u digitalnom obliku“ (Članak 1274. Slobodno korištenje djela u informativne, znanstvene, obrazovne ili kulturne svrhe). Time je zabranjena besplatna produkcija elektroničkih digitalnih kopija djela, skeniranje i digitalizacija. Isključivo autor ima pravo stavljanja predmeta autorskih prava na Internet.

U Rusiji, kao i u većini europskih zemalja, autorska prava traju za vrijeme autorova života i 70 godina nakon njegove smrti. Novi je zakon 2008. godine dopustio knjižnicama digitalizaciju knjiga koje imaju znanstvenu i obrazovnu vrijednost, a 2014. digitalizaciju svih knjiga koje nisu bile ponovno izdane na teritoriju zemlje u zadnjih 10 godina, no digitalno izdanje može biti izrađeno samo u jednom primjerku, dostupno samo unutar knjižnice. Time je postignut samo jedan cilj digitalizacije, odnosno očuvanje knjiga u digitalnom obliku, no široka dostupnost knjiga i dalje je onemogućena (Mehancev 2019). Knjige u digitalnom obliku unutar digitalnih knjižnica mogu imati tri vrste pristupa: pristup omogućen samo knjižničarima, samo posjetiteljima knjižnice ili univerzalan pristup.

Usprkos tome na Internetu je moguće pronaći različite knjige u digitalnom obliku za objavu koji njihovi autori nisu dali dopuštenje i tako se naravno krši autorsko pravo. Neke digitalne knjižnice pokušavaju zaobići autorska prava na doista maštovite načine. Jedan od primjera je i nekadašnja piratska digitalna knjižnica Librusec (rus. Либрусек). Njezin je vlasnik Ilja Larin (rus: Илья Ларин), Rus koji živi u Ekvadoru, te je za svoju stranicu tvrdio kako djeluje prema zakonima Ekvadora, te se na taj način ne krše autorska prava. U međuvremenu je Librusec

postala legalna internetska trgovina. Do sada se zakon nije previše miješao u internetsku sferu, no iako se to možda i promijeni u budućnosti, online piratstva uvijek će biti. Većini korisnika Interneta to naravno odgovara, jer žele neograničen pristup informacijama. (Atiz)

10. Zaključak

Želja znanstvenika da stvore stroj koji bi oponašao ljudske funkcije oduvijek je postojala. Iako je tehnologija optičkog prepoznavanja znakova silno napredovala u posljednjih par desetljeća, program koji bi čitao tekst jednako dobro kao što to čini čovjek i dalje ne postoji. Osim rukom pisanih tekstova, najveće probleme OCR programima zadaju stari tekstovi.

Prije same digitalizacije i optičkog prepoznavanja znakova potrebno je napraviti dobar plan kako bi se odredile ključne stavke digitalizacije: svrha digitalizacije, predviđeni budžet, stručnjaci koji će voditi cijeli projekt i sam odabir gradiva za digitalizaciju. Kako bi rezultat digitalizacije bio zadovoljavajući, potrebno je posebnu pažnju posvetiti svakoj od faza digitalizacije: odabiru gradiva za digitalizaciju, samoj digitalizaciji gradiva, obradi i kontroli kvalitete, zaštiti gradiva u elektroničkoj okolini, pohrani i prijenosu digitalnog gradiva, pregledu i korištenju digitalnog gradiva i održavanju digitalnog gradiva. Kod svake faze valja imati na umu o kakvoj je vrsti gradiva riječ, te o cilju digitalizacije.

Prije samog optičkog prepoznavanja znakova potrebno je odabrati program koji će se koristiti, bio on komercijalan ili besplatan. O toj odluci ovisit će predviđeni budžet i vremenski rok. Nakon toga potrebno je provesti prethodnu obradu slike kako bi kvaliteta skeniranih stranica bila što bolja i napraviti segmentaciju u kojoj će se stranica podijeliti na svoje komponente: tekst, slike i tablice. Kako bi sljedeća faza prepoznavanja znakova bila uspješna, važno je da prve dvije fazu budu pravilno provedene. Glavne metode prepoznavanja znakova su prepoznavanje na temelju predložaka i prepoznavanje na temelju svojstava oblika. Kod prve metode svaki zasebni znak se uspoređuje s gotovim predlošcima, dok se kod druge metode znak dijeli na svoje komponente kako bi bio prepoznat prema svojim značajkama. U zadnjoj se fazi znakovi ponovno sastavljaju kako bi se dobio cjelovit tekst, te se taj tekst provjerava kako bi sve otkrivene greške bile ispravljene.

Do grešaka dolazi ako je izvornik u lošem stanju s puno mrlja i poderotina, ako sve faze digitalizacije nisu pravilno provedene ili ako OCR program nije najtočniji. Neke od najčešćih grešaka su odbijanje znakova, zamjena znakova, zamjena velikog i malog slova, spajanje

dviju riječi, dijeljenje jedne riječi na više dijelova i pogrešno postavljeni interpunkcijski znakovi.

Na točnost OCR-a može se utjecati tako da se ispravno provede postupak prethodne obrade i pravilno namjesti rezolucija, te trenira program. Današnji OCR programi imaju veoma visoku točnost, no program koji bi imao stopostotnu točnost ne postoji. Točnost uvelike ovisi o vrsti i kvaliteti izvornika, pa će tako točnost biti niža, ako je izvornik star i u lošem stanju. Kod OCR postupka starih knjiga probleme stvaraju izblijedjeli tekstovi kod kojih je kontrast između teksta i pozadine jako malen te stari fontovi koji se danas više ne koriste. Osim njih, dolazi i do leksičkih i grafemskih problema, jer se pravopis i vokabular korišten u starim knjigama danas više ne koristi. Do tih problema dolazi i kod digitalizacije starih ruskih knjiga, jer se u mnogim koristi usko funkcionalan staroslavenski jezik koji se uvelike razlikuje od današnjeg suvremenog ruskog jezika. Osim što se u njemu pojavljuju slova koja se danas više ne koriste, u staroslavenskom se pojavljuje i dvojina, vokativ i glagolska vremena čijih oblika više nema u suvremenom ruskom.

Za potrebe rada provedeno je istraživanje u kojem su se uspoređivala dva programa za optičko prepoznavanje znakova, Abbyy FineReader i Transkribus, na primjeru *Ruske gramatike* Vladimira Vasiljeviča Lomonosova. Iako je Abbyy FineReader komercijalan program, rezultati dobiveni njime nisu bili zadovoljavajući, te je točnost dobivenog teksta bila samo 56,48%. Puno boljim se pokazao Transkribus, besplatan program koji je još uvijek u fazi razvoja. Kod njega je bilo potrebno uložiti više vremena i truda prilikom segmentacije i transliteracije, no točnost dobivenog teksta bila je čak 97,60%. Jednom istreniran model moguće je kasnije koristiti i na drugim tekstovima s jednakim fontom i pisanim istim jezikom.

Moguće je zaključiti kako za OCR starih ruskih knjiga ponekad nisu dovoljni obični OCR programi, već je potrebno koristiti specijalizirane programe poput Transkribusa, koji se može koristiti i za rukopise, ili je potrebno dodatno treniranje unutar običnog OCR programa. No u tim je slučajevima potrebno uložiti i više vremena i truda kako bi se dobio dobar rezultat, za razliku od običnih OCR-a programa u kojima program obavi većinu posla, ali je nakon toga potrebno uložiti više vremena na korigiranje.

Za sam kraj valja dodati kako je trenutačno putem Interneta omogućen pristup velikom broju ruskih knjiga u digitalnom obliku, ali pri tome valja pripaziti i na autorska prava, koja se često krše prilikom objavljivanja digitalnih knjiga.

Literatura

Abbyy. ABBYY FineReader 14 for Windows. URL: https://www.abbyy.com/en-au/finereader/ (13.03.2019.).

Abbyy technology portal. OCR Character Classifier. URL: https://abbyy.technology/en:features:ocr:classifier (13.03.2019.).

Abdulwahhab Hamad, K., Kaya, M. A Detailed Analysis of Optical Character Recognition Technology. // International Journal of Applied Mathematics, Electronics and Computers (2016) str. 244-248. URL: http://dergipark.gov.tr/download/article-file/236939 (13.03.2019.).

Ačimović, A. Upotreba ćirilice na istočno- i južnoslavenskom prostoru (od postanka do suvremenih azbuka, sociopolitički pogled). Zagreb: diplomski rad, 2018.

Andrianov 2009. Андрианов, А. И. Сравнение OCR-систем на основе точности анализа изображения. // Бизнес-информатика 4, 10 (2009). str. 44-45. URL: https://cyberleninka.ru/article/v/sravnenie-ocr-sistem-na-osnove-tochnosti-analiza-izobrazheniya (13.03.2019.).

Atiz. Проблема авторского права: Авторское право и сканирование библиотечного фонда. URL: http://atiz.ru/avtorskoe-pravo (13.03.2019.).

Azad, S., Jain, K. CAPTCHA: Attacks and Weaknesses against OCR Technology. // Global Journal of Computer Science and Technology 13, 3 (2013) str. 14-17.

Bondarenko 2012. Бондаренко, С. Интернет-портал «Руниверс»: в помощь исследователю. 30. kolovoza 2012. URL: http://urokiistorii.ru/article/51450

Britannica. Paul Gottlieb Nipkow. URL: https://www.britannica.com/biography/Paul-Gottlieb-Nipkow (13.03.2019.).

Cornell University Library/Research Department. How scanners work. URL: http://preservationtutorial.library.cornell.edu/technical/technicalB-02.html (08.04.2019.).

Čelić, Ž. Latinski metajezik – matrix slavenskih gramatika. Utjecaj latinskoga na hrvatski i istočnoslavenske jezike, prikazan jezičnim nazivljem, opisom glasova i oblika u hrvatskome i istočnoslavenskim jezicima. Zagreb: doktorska disertacija, 2008.

Dalbir, Singh, S. K. Review of Online & Offline Character Recognition. // International Journal Of Engineering And Computer Science 4, 5 (2015), str. 11729-11732

Damjanović, S. Staroslavenski jezik. Zagreb: Hrvatska sveučilišna naklada, 2003.

Eikvil, L. OCR Optical Character Recognition. 1993. URL: https://www.nr.no/~eikvil/OCR.pdf (13.03.2019.).

Evtjuhin. Евтюхин В. Б. "Российская Грамматика" М. В. Ломоносова. URL: http://www.ruthenia.ru/apr/textes/lomonos/add02.htm (13.03.2019.).

Gbooks. О проекте. URL: http://gbooks.archeologia.ru/ (08.04.2019.)

Gliha, I., Autorsko pravo, zbirka propisa, Zagreb, 2000, str. 1; Henneberg, I., Autorsko pravo, Zagreb, 2001, str. 1-5; Henneberg, I., Nazivlje u autorskom pravu, Zbornik Hrvatskog društva za autorsko pravo (HDAP), vol. 1, 2000, str. 5-6.

Google Drive Help. Convert PDF and photo files to text. URL: https://support.google.com/drive/answer/176692?co=GENIE.Platform%3DDesktop&hl=en (13.03.2019.).

Gupta, M. R., Jacobson, N. P., Garcia, E. K. OCR binarization and image pre-processing for searching historical documents // Pattern Recognition. 40, 2(2007), str. 389-397. URL: https://www.sciencedirect.com/science/article/pii/S0031320306002202 (13.03.2019.).

Hauser, A.W. OCR Postcorrection of Historical Texts. 4. listopada 2007. URL: http://www.cip.ifi.lmu.de/~hauser/papers/histOCRNachkorrektur.pdf (13.03.2019.).

Holley, R. How Good Can It Get?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. // D-Lib Magazine 15, 3/4(2009). URL: http://www.dlib.org/dlib/march09/holley/03holley.html (13.03.2019.).

Ilindra, A. Top 6 Best OCR Software to Extract Text from Images. URL: https://www.geekdashboard.com/best-ocr-software-to-extract-text-from-images/ (13.03.2019.).

INION RAN. ИНИОН РАН. История библиотеки. URL: http://inion.ru/library/istoriia-biblioteki/ (13.03.2019.).

Iris. Readiris 17, the PDF and OCR solution for Windows. URL: http://www.irislink.com/EN-US/c1729/Readiris-17--the-PDF-and-OCR-solution-for-Windows-.aspx (13.03.2019.).

Islam, N, Islam, Z., Noor, N. A Survey on Optical Character Recognition System // Journal of Information & Communication Technology 10, 2(2016). URL: https://arxiv.org/ftp/arxiv/papers/1710/1710.05703.pdf (13.03.2019.).

Ivanov 1990. Иванов, В. В. Историческая грамматика русского языка. Москва: Посвещение, 1990. URL: http://mling.ru/iazik/russe/gramm_hist.pdf (13.03.2019.).

Katić, T., Klarin, S., Obhođaš, A., Bukovac, D., Seiter-Šverko, D. Smjernice za odabir građe za digitalizaciju. 2007.

Lomonosov 1788. Ломоносов, М. В. Российская грамматика. Санкт-Петербург: Императорская Академия Наук, 1788. URL: http://elib.gnpbu.ru/text/lomonosov_rossiyskaya-grammatika_1788/go,0;fs,1/ (13.03.2019.).

Matthews, K. The 3 Best Free OCR Tools to Convert Your Files Back Into Editable Documents. 26. listopada 2017. URL: https://www.makeuseof.com/tag/3-free-ocr-tools-convert-files-editable-documents/ (13.03.2019.).

Mehancev. Механцев, Ж. Проблемы оцифровки литературы и норм авторского права в цифровую эпоху. URL: https://nauchkor.ru/pubs/problemy-otsifrovki-literatury-i-norm-avtorskogo-prava-v-tsifrovuyu-epohu-595122695f1be749c9c635c9 (13.03.2019.).

Mithe, M., Indalkar, S., Divekar, N. Optical Character Recognition // International Journal of Recent Technology and Engineering. 2, 1(2013). URL: https://pdfs.semanticscholar.org/6a4b/4f04d5ce3c3592832eb40c23cc8fc5a9131e.pdf (13.03.2019.).

Nacional'naja Elektronnaja Biblioteka. Национальная электронная библиотека. О проекте. URL: https://нэб.рф/about/ (13.03.2019.).

Nield, D. Best OCR software of 2019. URL: https://www.techradar.com/news/best-ocr-software (13.03.2019.).

Nikitin-Perenski. Никитин-Перенский, А. О новых поступлениях в электронную библиотеку ImWerden. URL: http://sites.utoronto.ca/tsq/29/nikitin29.shtml (13.03.2019.).

Northeast Document Conservation Center. Handbook for digital projects: A Management Tool for Preservation and Access. 2010. URL: https://www.nedcc.org/assets/media/documents/dman.pdf?fbclid=IwAR2Pr1ms0VVD7y5PkdhF2SfUKpwGldhu1TvRmuLlw0v3-t92GAYo5OSsVm8 (13.03.2019.).

Nuance. OmniPage Ultimate: Easy, reliable and robust. URL: https://www.nuance.com/print-capture-and-pdf-solutions/optical-character-recognition/omnipage/omnipage-ultimate.html (13.03.2019.).

Pirker, J., Wurzinger, G. Optical Character Recognition of Old Fonts - A Case Study. URL: http://ipsitransactions.org/journals/papers/tar/2016jan/p3.pdf (13.03.2019.).

Popović, M. Osnove staroslavenskog za studente ruskog jezika. Zagreb: Sveučilišna naklada Liber, 1983.

Radošević, D. Postupci i problemi optičkog prepoznavanja znakova. //Journal of Information and Organisational Sciences. 20, 2 (1996). URL: http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=117350 (13.03.2019.).

READ. How To Train A Handwritten Text Recognition Model In Transkribus. URL: https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf (13.03.2019.).

Shahi, M., Ahlawat, A. K., Pandey, B. N. Literature Survey on Offline Recognition of Handwritten Hindi Curve Script Using ANN Approach. // International Journal of Scientific and Research Publications. 2, 5 (2012). URL: https://pdfs.semanticscholar.org/ca62/8fe973547f252cb175b0d2975893e2d8c33b.pdf (13.03.2019.).

Sharma, R. Extract Text From Images With These Best OCR Software. 2017. URL: https://beebom.com/best-ocr-software/ (13.03.2019.).

Singla, S. K., Yadav, R. K. Optical Character Recognition Based Speech Synthesis System Using LabVIEW. // Journal of Applied Research and Technology. 12, 5(2014). str. 919-926. URL: https://www.sciencedirect.com/science/article/pii/S166564231470598X (13.03.2019.).

Smith, A. Why Digitize?. Council on Library and Information Resources, Washington, D.C. veljača 1999. URL: https://www.clir.org/pubs/reports/pub80-smith/pub80-2/ (13.03.2019.).

Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A., Fink, F. OCR of Historical Printings of Latin Texts: Problems, Prospects, Progress. 2014. URL: http://springmann.net/papers/2014-DATeCH-Springmann.pdf (13.03.2019.).

Stančić, H. Digitalizacija. Zagreb: Zavod za informacijske studije, 2009.

Tanner, S., Muñoz T., Ros P. H. Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. // D-Lib Magazine 15, 7/8(2009). URL: http://www.dlib.org/dlib/july09/munoz/07munoz.html (13.03.2019.).

Verma, A., Arora, S., Verma, P. OCR- Optical Character Recognition // International journal of advance research in science and engineering. 5, 9 (2016). str 181-191. URL: https://pdfs.semanticscholar.org/041f/3b52994a1653e27797e349f42bebac35d246.pdf (13.03.2019.).

Vynckier, I. How OCR works URL: http://www.how-ocr-works.com/index.html (13.03.2019.).

Woodford, C. Optical character recognition (OCR) URL: https://www.explainthatstuff.com/how-ocr-works.html

Popis slika

Popis tablica

Problem prepoznavanja znakova iz starijih ruskih knjiga tijekom procesa digitalizacije, na primjeru Gramatike M. V. Lomonosova

Sažetak

Ovaj se diplomski rad bavi problemima do kojih dolazi prilikom digitalizacije starih ruskih knjiga. Budući da starim knjigama prijeti opasnost od njihovog propadanja i uništavanja, veoma je važno da se na vrijeme krene s postupkom digitalizacije kako bi se one sačuvale barem u svom digitalnom obliku. Problemi do kojih dolazi prilikom OCR-a takvih starih knjiga vezani su uz loše stanje u kojem se knjige nalaze, ali i uz zastarjeli jezik kojim su one često pisane, a koji se danas u tom obliku više ne koristi. U radu se detaljno opisuje postupak digitalizacije i optičkog prepoznavanja znakova. U istraživačkom dijelu uspoređuju se dva različita programa za OCR na primjeru *Ruske gramatike* Mihaila Vasiljeviča Lomonosova, te se daju preporuke za provođenje uspješne digitalizacije starih knjiga.

Ključne riječi: digitalizacija, optičko prepoznavanje znakova, OCR, stare ruske knjige, staroslavenski

The problem of recognizing characters from old Russian books during the digitization process, on the example of Grammar by M. V. Lomonosov

Abstract

This graduate thesis deals with the problems that arise when digitizing old Russian books. Since old books are in jeopardy of their decay and destruction, it is very important to initialize the digitization process as soon as possible in order to preserve them, at least in their digital form. The problems with the OCR of such old books are related to the poor physical state of the books as well as to the use of obsolete language. This thesis describes in detail the procedure of digitization and optical character recognition. The research part compares two different OCR programs on the example of *Russian grammar* by Mikhail Vasilyevich Lomonosov, and gives recommendations for successful digitization of old books.

Key words: digitization, optical character recognition, OCR, old Russian books, Old Church Slavonic

Аннотация

В данной дипломной работе рассматриваются проблемы, возникающие при оцифровке старых русских книг. Оцифровка — преобразование текста, изображений, звука, движущихся изображений (фильмы и видео) или 3d объектов в цифровой формат. Основная причина оцифровки — защита и сохранение книг, находящихся под угрозой деградации, а также обеспечение широкой доступности книг в разных странах мира.

Оцифровка состоит из 7 фаз: подбор материалов для оцифровки, оцифровка материалов, обработка и контроль качества, защита материалов в электронной среде, хранение и передача цифрового материала, просмотр и использование цифрового материала и сопровождение цифрового материала. Перед началом оцифровки необходимо определить предусмотренный бюджет и срок завершения проекта и решить, какие книги должны быть оцифрованы первыми. Книгу можно оцифровать с помощью цифровой камеры или сканера, а полученное изображение затем проходит через программу OCR. Наконец, оцифрованный материал должен быть защищен, сохранен и передан пользователям.

Чтобы сделать процесс оцифровки максимально простым и быстрым, ученые разработали оптическое распознавание символов. Оптическое распознавание символов (OCR) — это технология, с помощью которой рукописи, печатные тексты и документы, записанные в цифровой форме, преобразуются в текстовые документы, которые можно обрабатывать. Существует большое количество коммерческих и бесплатных OCR программ. При выборе программы бюджет и срок завершения проекта являются самыми важными факторами.

Фазы оптического распознавания символов: предыдущая обработка, анализ изображений, т. е. сегментация, распознавание символов и последующая обработка. После сканирования изображения необходимо устранить нежелательный шум, полученный пятнами на изображении, но без потери существенной информации, и нужно получить хороший контраст между текстом и фоном. В течение сегментации страница делится на текст, изображения и таблицы. Текст далее делится на строки, слова и символы. После этого программа OCR распознает символы на основе шаблонов и на основе свойств формы. Распознанные символы должны быть затем вновь собраны в полный текст, а в полученном тексте необходимо проверить наличие ошибок. Эта

проверка может быть выполнена вручную пользователем или с помощью встроенных словарей в рамках программы.

Некоторые из наиболее распространенных ошибок в распознавании символов — невозможность распознать символ, замена двух символов, замена прописных и строчных букв, объединение двух слов, разделение одного слова на несколько частей или неправильно поставленная пунктуация. Точность результата оптического распознавания символов в наибольшей степени зависит от качества оригинала, так что точность старых текстов будет намного ниже, чем у более новых материалов. Точность оптического распознавания символов можно улучшить обучением, но это возможно только у некоторых программ OCR.

При оцифровке старых книг проблема заключается в недостаточном контрасте между текстом и фоном из-за пожелтевших листов бумаги и выцветавшего текста и шрифтов, которые сегодня больше не используются. Кроме того, старые вариации правописания и устаревший словарный запас затрудняют контроль при оптической проверке текста. Большое количество старых русских книг написано на старославянском языке или на более старой версии русского языка и очень сильно отличаются от текстов, написанных на современном русском языке. При оптическом распознавании символов старых русских книг многие программы обнаружат проблемы со всеми словами, словоформами и правописанием, которые отличаются от современного стандартизированного русского языка.

Таким образом, у многих OCR программы возникнут проблемы с распознаванием букв, которых больше нет в стандартном русском языке: „ѱ“, „ѯ“, „ѡ“, „ѵ“, „ѧ“, „ѳ“, „і“ и „ї“ и старых глагольных времен, падежей и склонений.
В рамках этой дипломной работы было проведено исследование, сравнивающее две разные программы для оптического распознавания символов: Abbyy FineReader и Transkribus. Анализ проводился по *Российской грамматике* Михаила Васильевича Ломоносова, написанной в 1755 году. Abbyy FineReader — одна из наиболее широко используемых коммерческих OCR программ, разработанная российской компанией Abbyy.

Ее использование было довольно простым. После анализа страницы, программа автоматически распознала все символы. У программы возникли проблемы с

распознаванием таблиц и изогнутых скобок при анализе, и в итоге удалось заметить большое количество неправильно идентифицированных символов.

Transkribus является бесплатной платформой для автоматического распознавания, транслитерации и поиска исторических документов. Он является частью проекта READ (Распознавание и обогащение архивных документов), финансируемого Европейским Союзом. В Transkribus потребовалось больше времени для сегментации страницы, потому что программа не делила текст точно на строки и поэтому требовалась ручная сегментация. После этого было необходимо ввести расшифрованный текст первых 30 страниц для обучения модели, что не было проблемой, так как книга уже была в цифровом формате. В противном случае этот процесс потребовал бы намного больше времени. Через некоторое время команда Transkribus разработала модель распознавания символов, которая работала очень хорошо.

Результаты обеих программ были проверены с помощью аналитических инструментов ISRI, которые показали, что точность результата программы Abbyy FineReader была только 56.48%, а программы Transkribus 97.60 %. Можно сделать вывод, что для оцифровки старых книг требуются специализированные программы, поскольку в классических программах OCR слишком много ошибок распознавания символов, и исправление этих ошибок займет слишком много времени.

В России была признана необходимость оцифровки книг, поэтому многие российские библиотеки инициировали проекты оцифровки и русские книги сейчас доступны на различных веб-сайтах. Но нужно обратить внимание и на возможное нарушение авторских прав, которое иногда возникает при публикации книг в цифровом виде.

Ключевые слова: оцифровка, оптическое распознавание символов, OCR, старые книги, старославянский язык

Životopis

Ivana Cencelj rođena je 30. travnja 1993. godine u Zagrebu. Godine 2011. završila je Gimnaziju Tituša Brezovačkog. Od 2012. godine studira ruski jezik i književnost i informacijske i komunikacijske znanosti na Filozofskom fakultetu Sveučilišta u Zagrebu. Godine 2016. završava preddiplomski studij na informacijskim i komunikacijskim znanostima te upisuje diplomski, smjer arhivistike. Godine 2017. završava preddiplomski studij na Odsjeku za istočnoslavenske jezike i književnosti te upisuje diplomski, prevoditeljski smjer. Iste godine sudjeluje na 19. Svjetskom festivalu mladih i studenata u Krasnojarsku i Sočiju, te joj je velika želja ponovno posjetiti Rusiju.