JOSIP MIHALJEVIĆ*

Može li računalo pročitati tekst na hrvatskome jeziku?

Digitalizacija je proces u kojemu se analogni signal ili informacija koja se nalazi u određenome mediju pretvara u digitalni signal koji predstavlja analogni signal na nekome uređaju. S prikladnom se tehnologijom može digitalizirati gotovo bilo koji fizički objekt. Najčešće se digitaliziraju tekstni sadržaji poput dokumenata i knjiga, koji mogu sadržavati slike i fotografije.

Često se nalazimo u situaciji da trebamo prepisati dugačak tekst iz knjige, s papira ili sa slike. Tu nam računalo može pomoći s pomoću programa za optičko prepoznavanje znakova (engl. *OCR – optical character recognition*), koji mogu unutar slike prepoznati tekst te ga izvesti u tekstni dokument u kojemu se dalje može obraditi te kopirati za korištenje u drugim programima. Na internetu postoji mnogo besplatnih programa za optičko prepoznavanje znakova, od kojih neki prepoznaju i hrvatska slova. Ipak, prije prepoznavanja teksta potrebno je objekt koji sadržava tekst skenirati ili fotografirati. Taj proces stvaranja digitalne slike te njezine obrade zove se digitalizacija.

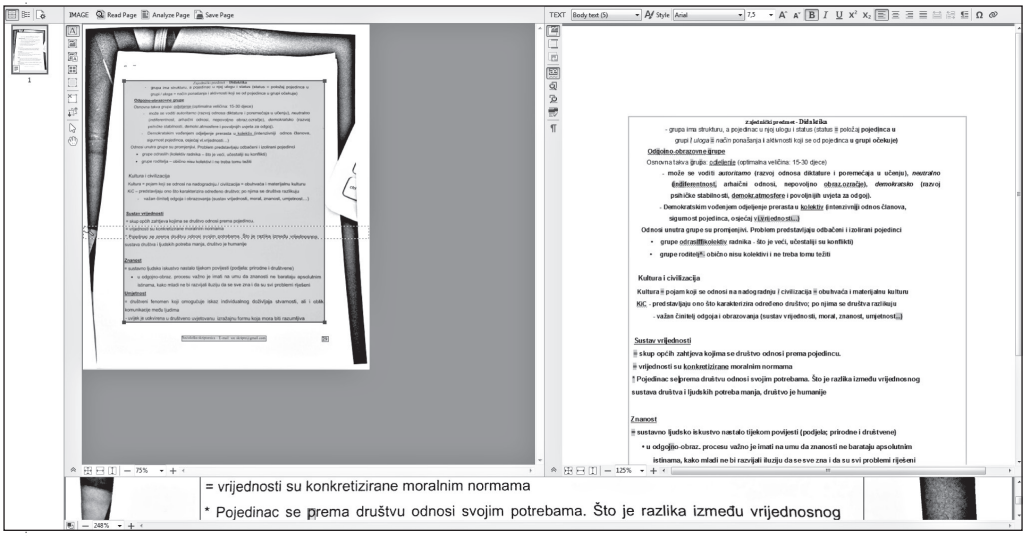
Tekst se može digitalizirati s pomoću stolnih skenera. Nakon skeniranja slike prelazi se na obradu i provjeru kvalitete digitaliziranoga testa. Obrada se digitaliziranoga teksta provodi s pomoću programa za optičko prepoznavanje znakova.

Prepoznavanje teksta unutar slike provodi se s pomoću programa za optičko prepoznavanje znakova, koji omogućuje dodatnu izmjenu prepoznatoga teksta te njegovo spremanje u određeni tekstni dokument (npr. Wordov .docx).

Ti programi mogu prepoznati slova unutar slike na temelju podataka o izgledu znakova te njihovih međusobnih veza u oblikovanju riječi. Tim se programima koriste knjižnice, muzeji i arhivi te druge kulturne ustanove kako bi za skenirane sadržaje omogućili prepoznavanje znakova. Uglavnom točno prepoznaju otisnute brojeve i osnovna latinična slova, ali imaju problema s prepoznavanjem rukopisa zato što su rukopisna slova jedinstvena oblika te katkad nečitka. Većini su programa za prepoznavanje teksta problem i dijakritički znakovi jer se ne nalaze u bazi programa. Ipak se na internetu mogu naći programi koji podržavaju hrvatski jezik. Jedan je od boljih programa za prepoznavanje hrvatskih znakova Abby FineReader,

* Josip Mihaljević profesor je informatike u Školi za medicinske sestre Vrapče.

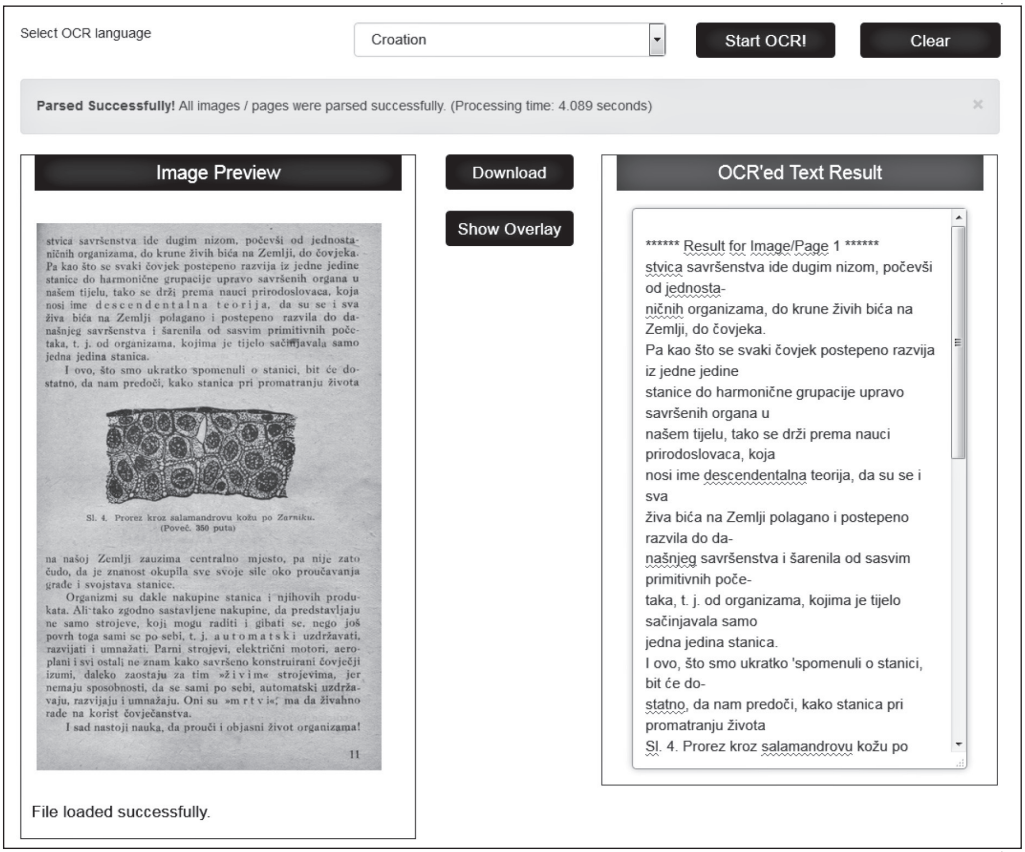
koji je ujedno i program za skeniranje. On omogućuje za svaku skeniranu sliku izravnu obradu slikanoga teksta te označuje moguće pogreške u prepoznavanju, koje se mogu ručno ispraviti. Nažalost, to je komercijalni program koji se katkad dobiva besplatno s kupljenim skenerom.



1. slika: ispravljanje prepoznatoga teksta unutar slike s pomoću programa Abby FineReader

Postoji besplatna mrežna inačica toga programa (https://finereaderonline.com/en-us), koja služi isključivo za prepoznavanje znakova unutar slika te njihov izvoz u različitim formatima poput Worda, PowerPointa i PDF-a. Za uporabu toga programa potrebno se registrirati na stranici. Program vrlo precizno prepoznaje tiskane znakove s rijetkim pogreškama, koje se lako mogu uočiti u Wordu ako se uključi pravopisni provjernik za hrvatski jezik. Moguće je uključiti prepoznavanje teksta za više jezika, pa čak i više pisama, što može biti korisno kad prepoznajemo tekst unutar slike koji je osim latinice napisan na ćirilici, kineskome, arapskome ili kojemu drugom pismu. Također postoje opcije za prepoznavanje umjetnih jezika poput esperanta te programskih jezika poput Jave, Pascala i jezika C++. Besplatna mrežna inačica toga programa nudi prepoznavanje znakova za najviše deset stranica mjesečno po korisničkome računu. To nije problem za one koji povremeno rade s manjim dokumentima, ali nije dobro ako treba prepoznati tekst skenirane knjige. Od besplatnih programa za skeniranje koji prepoznaju hrvatske znakove preporučio bih NAPS2 (https://sourceforge.net/projects/naps2/), koji dobro funkcionira kad treba prepoznati ravno skeniranu stranicu, ali ne nudi mogućnost spremanja dokumenta u Wordu, nego isključivo u PDF-u.

Postoje i mnogi drugi mrežni programi za prepoznavanje hrvatskih znakova poput Free Online OCR-a (https://www.onlineocr.net/) i i2OCR (http://www.i2ocr.com/), ali oni uglavnom dobro prepoznaju hrvatske znakove samo ako je riječ o ravno skeniranoj i čitljivoj slici. Program OCR.space (https://ocr.space/) precizniji je od Free Online OCR-a i i2OCR jer može s velikom točnošću prepoznati tekst sa staroga papira te sliku s fotoaparata. Taj program može se neograničeno i besplatno upotrebljavati bez registracije. Program podržava hrvatski jezik, vrlo je precizan pri prepoznavanju znakova skeniranih starijih tekstova ili slika s mobitela. Omogućuje i usporedbu teksta i slike kako bi se tekst mogao urediti prije nego što se izveze u .txt formatu. Nedostatak mu je što ne omogućuje spremanje teksta u Wordu, ali .txt datoteke mogu se otvoriti na svim računalima te lako kopirati u Word. Još je jedan problem toga programa što ne omogućuje učitavanje datoteka koje imaju više od 5 MB. Većina dokumenata ipak ne prelazi to ograničenje, osim visokvalitetno skeniranih knjiga koje imaju mnogo stranica.



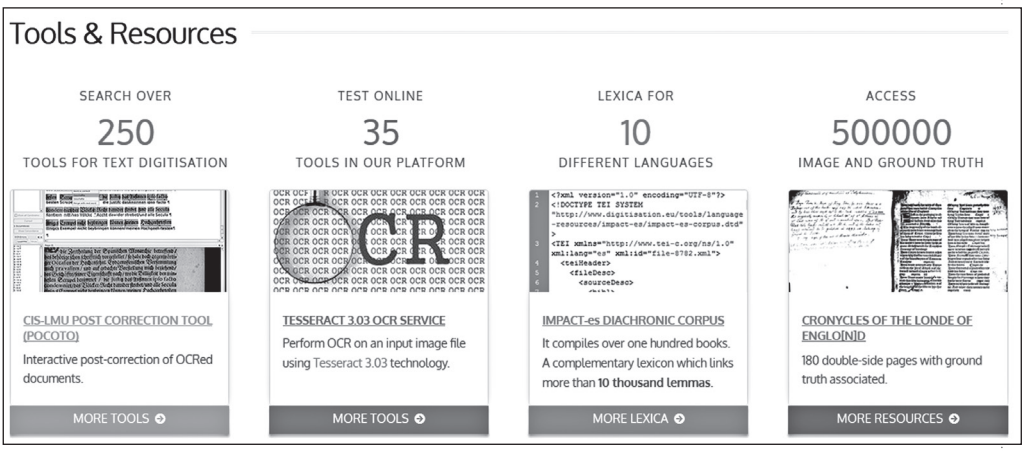
2. slika: programom OCR.space prepoznat tekst unutar slike može se izravno ispraviti te kopirati ili spremiti u tekstni dokument

U tablici su navedene prednosti i nedostatci programa za prepoznavanje hrvatskoga teksta.

	Abby FineReader 12 -14	ABBYY FineReader Online	NAPS2	OCR.space
cijena	1190 kn – 1250 kn (+ PDV) program se katkad može dobiti besplatno s kupljenim skenerom	besplatan	besplatan	besplatan
ograničenje korištenja	nema ograničenja u radu s opcijama nakon što se program kupi ili dobije s licencijom	po korisničkome računu može se obraditi najviše 10 stranica mjesečno	nema ograničenja	ne mogu se obraditi dokumenti koji imaju više od 5 MB
spremanje prepoznatoga teksta	može se spremiti kao Word, OpenOffice, Excel, PDF, elektronička knjiga (ePub) i čista tekstna datoteka (.txt)	opcije za spremanje iste su kao i kod komercijalnoga programa	može se spremiti jedino unutar slike kao PDF dokument	može se spremiti kao čista tekstna datoteka (.txt) iz koje se poslije može lako kopirati tekst u druge dokumente

Treba samo uzeti u obzir da su AbbyFinereder i NAPS2 programi za skeniranje koji sadržavaju program za prepoznavanje znakova, a OCR.space i ABBYY FineReader Online mrežni su programi koji jedino omogućuju prepoznavanje teksta unutar slike te njegov izvoz u tekstnoj datoteci. Svi programi mogu precizno prepoznati hrvatske znakove, jedino ih NAPS2 nešto slabije prepoznaje, ali je i on dovoljno dobar za većinu namjena.

Iz ove kratke usporedbe možemo zaključiti da je najbolji program Abby FineReader, ali on nije besplatan. Njegova je mrežna inačica besplatna, ali ima ograničenje na samo deset stranica. OCR.space je zbog toga najbolji besplatni mrežni program za prepoznavanje hrvatskoga teksta. Najbolje je njime se koristiti u kombinaciji s programom FineReader Online. OCR.space može skenirati dokumente manje od 5 MB. Njih možemo obraditi s pomoću programa ABBYY FineReader Online. Programom NAPS2 možemo se koristiti kad izravnim skeniranjem iz dokumenta ili knjige treba dobiti pretraživ dokument u PDF-u. Još neki alati za prepoznavanje i obradu teksta unutar slika mogu se naći na stranici IMPACT Centre of Competence (https://www.digitisation.eu/), na kojoj se registriranim korisnicima nude mrežne poveznice i usluge alata za prepoznavanje teksta.



3. slika: poveznice na različite alate i resursa za prepoznavanje teksta na stranicama organizacije IMPACT Center of Competence

Programi za prepoznavanje teksta još nisu dovoljno dobri da prepoznaju rukopise zbog njihove različitosti i često nečitljivosti, ali i ta se tehnologija polako razvija. Google i mnoge druge tvrtke sa svojom reCAPTCHA metodom za potvrdu autentifikacije traže od korisnika da prepišu rukopis sa slike kako bi pristupili sadržaju. To se ne radi samo kako bi se spriječio pristup štetnim programima koji ne mogu prepisati nasumično odabran tekst nego i da se ono što korisnici prepišu ubaci u bazu za prepoznavanje teksta kako bi se u budućnosti poboljšala tehnologija za prepoznavanje znakova. Možemo se nadati da će računalo uskoro moći prepoznati i rukopis.