ANA MIHALJEVIĆ*

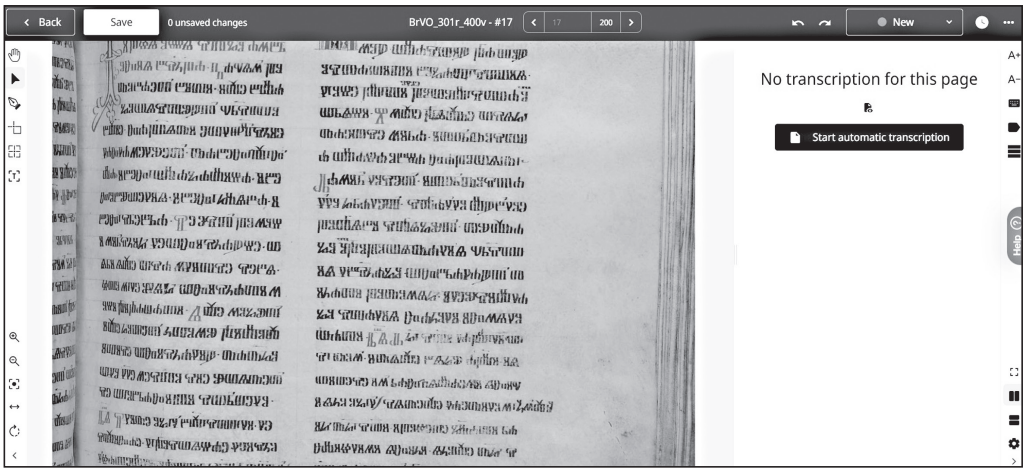
Osnovno o Transkribusu[1]

U posljednjih je nekoliko godina postignut izniman tehnološki napredak u automatskome očitavanju rukopisnih tekstova (HTR – *Handwritten text recognition*). Niz je poslova koji su donedavno zahtijevali često i višemjesečni rad sad zamjenjiv aplikacijama koje u vrlo kratkome vremenu očitaju rukom pisani tekst. Transkribus je jedna od platforma koje omogućuju takvo očitavanje rukopisa. Nastao je u sklopu dvaju projekata: *tranScriptorium* (2013. – 2015.) i *READ* (*Recognition and Enrichment of Archival Documents*, 2016. – 2019.) na Sveučilištu u Innsbrucku. Od 2019. platformu vodi READ-COOP. Transkribus nudi niz modela za prepoznavanje različitih vrsta rukopisa i za različite jezike, ali korisnik može i sam trenirati model koji će biti razvijen na temelju njegove građe. Također, aplikacijom se može služiti za uređivanje teksta, a na tekstu može istodobno raditi više korisnika. Aplikacija je dostupna na mrežnim stranicama https://www.transkribus.org/. Korisnik se prvo mora registrirati i prijaviti u aplikaciju na stranicama https://account.readcoop.eu/.

Nakon ulaska u aplikaciju korisnik treba učitati dokumente na kojima želi raditi pritiskom na opciju *upload files*. Prihvatljivi su formati JPEG/JPG do 10 MB i PDF do 200 MB s najviše 3000 stranica. Nakon učitavanja pojedine slike ili više njih korisnik treba otvoriti sliku s kojom želi raditi.

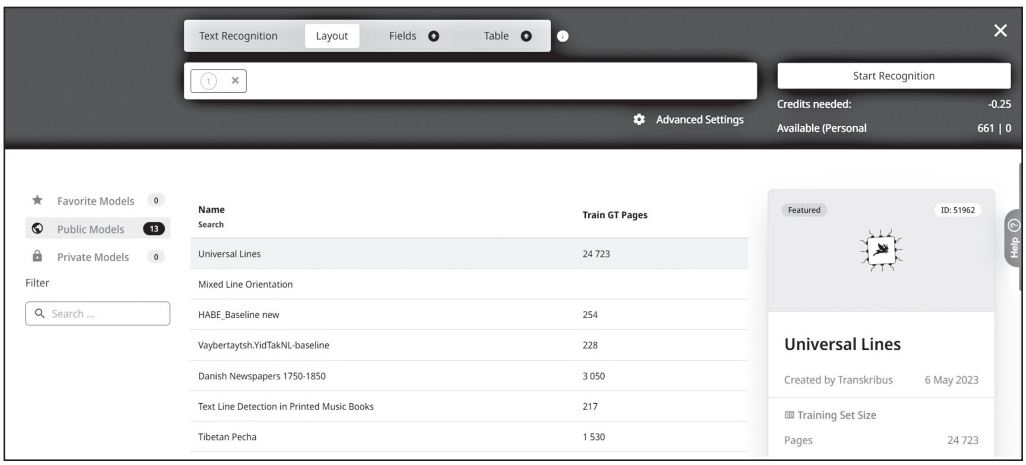
* Ana Mihaljević znanstvena je suradnica na Odjelu za Rječnik crkvenoslavenskoga jezika hrvatske redakcije u Staroslavenskome institutu.

[1] Ovaj je rad nastao u okviru projekta *Razvoj modela digitalne infrastrukture Staroslavenskoga instituta – DigiSTIN*, koji financira Europska unija – NextGenerationEU. Za iznesene stavove i mišljenja odgovorna je samo autorica te ti stavovi ne odražavaju nužno službena stajališta Europske unije ili Europske komisije. Ni Europska unija ni Europska komisija ne mogu se smatrati odgovornima za njih.



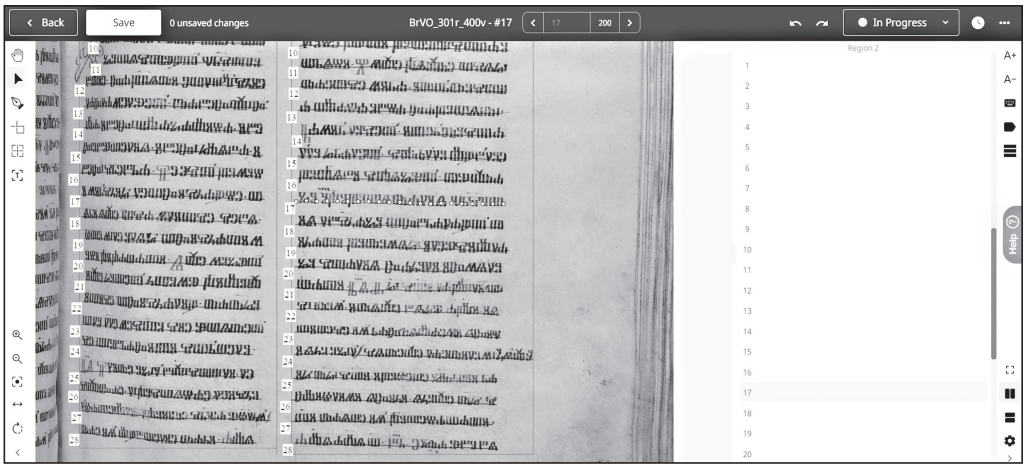
1. slika: Slika izvora otvorenoga u aplikaciji Transkribus

Nakon učitavanja dokumenta korisnik treba prvo automatski prepoznati linije u dokumentu odabirom opcije *recognition* u gornjemu lijevom dijelu izbornika (ikona sa slovom T). Sljedeći je korak odabir opcije *layout* te jednoga od modela za prepoznavanje izgleda teksta. U većini slučajeva primjenjiv je osnovni model *Universal lines*.



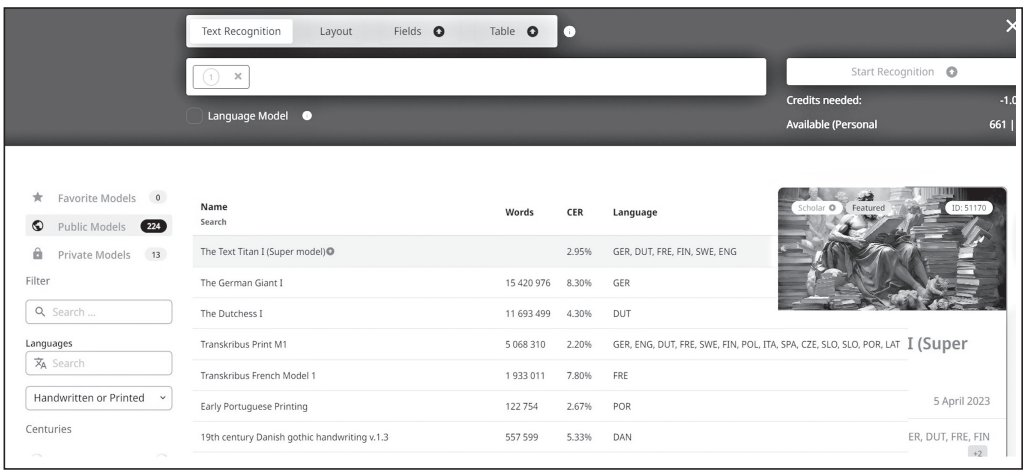
2. slika: Prepoznavanje linija – odabir modela *Universal lines*

Novija inačica Transkribusa ima razvijene modele i za prepoznavanje tablica te drugih složenijih grafičkih oblika. Nakon odabira modela korisnik treba pritisnuti opciju *start recognition*. Sustav automatski prepoznaje stupce i retke, a korisnik ih može prilagođivati i ispravljati ono što je sustav automatski prepoznao s pomoću opcija u gornjemu lijevom kutu.



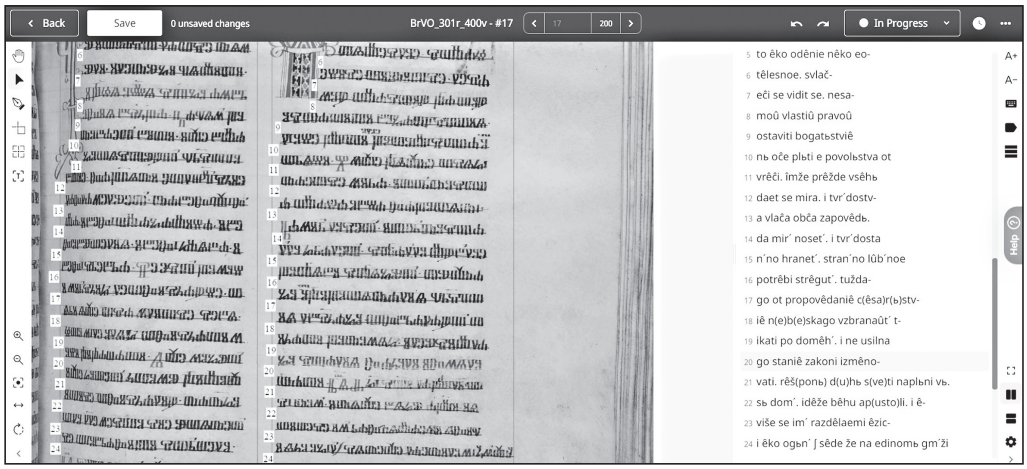
3. slika: Slika teksta s prepoznatim linijama

Korisnik zatim može prepisivati ili transliterirati tekst po redcima ili može iskoristiti neki od dostupnih modela za automatsko očitavanje. Za uporabu postojećih modela korisnik treba ponovno odabrati opciju *start recognition*, ali ovaj put odabrati opciju *text recognition* te neki od postojećih modela. Modeli se mogu pretraživati prema jeziku, tipu pisma i drugim značajkama.



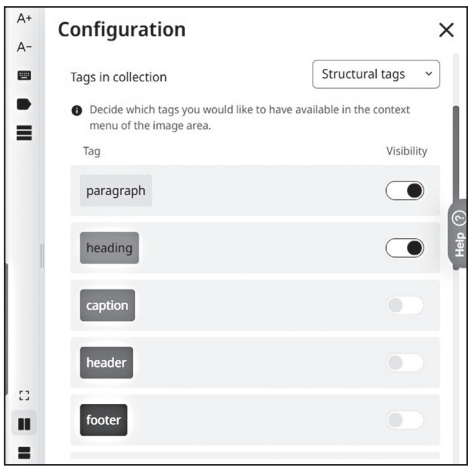
4. slika: Odabir modela za prepoznavanje teksta

Nakon automatskoga očitanja korisnik može u sučelju ispravljati tekst.

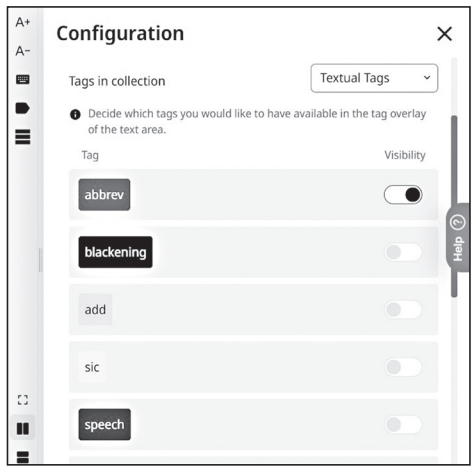


5. slika: Izgled sučelja nakon primjene modela za prepoznavanje teksta

Korisniku je dostupan i sustav oznaka (*tags*) kojima se mogu označavati različiti dijelovi teksta te različite pojave u tekstu.

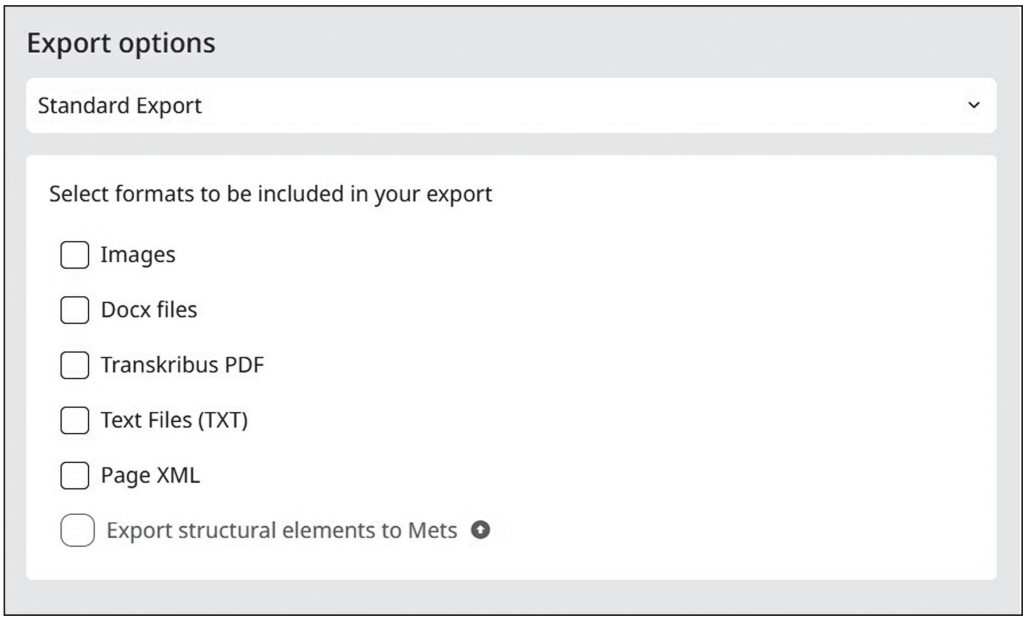


6. slika: Dio oznaka kojima korisnik može označiti tekst – strukturne oznake



7. slika: Dio oznaka kojima korisnik može označiti tekst – tekstne oznake

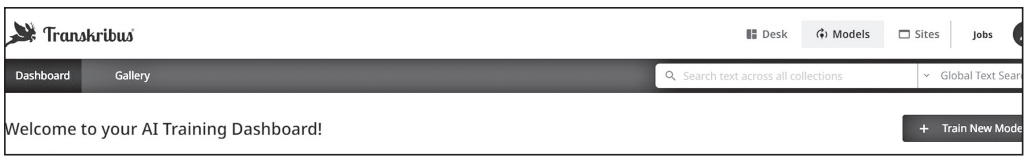
Na kraju se prepisani tekst može izvesti u različitim formatima (*export*).



8. slika: Mogućnosti izvoza teksta

Korisnik može i sam trenirati model na temelju teksta koji je prepisao. Trebao bi unijeti najmanje dvadeset stranica prepisanoga teksta uz sliku. Od skupa za treniranje sustav izdvaja određeni postotak stranica na kojima će se model testirati i s pomoću kojega će se odrediti postotak pogrešaka. Novi se model može trenirati ili samo na osnovi prepisanoga teksta ili na osnovi prepisanoga teksta i već postojećega modela.

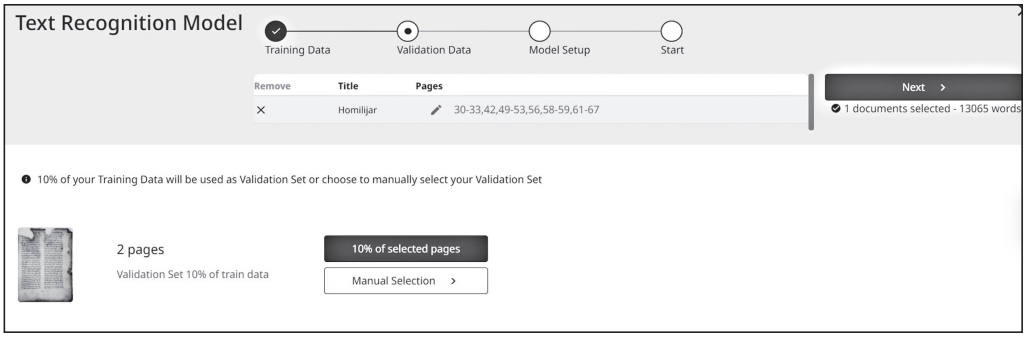
Kako bi trenirao novi model, korisnik mora pritisnuti na opciju *models* i odabrati opciju *train new model*.



9. slika: Treniranje novoga modela – odabir opcije *train new model*

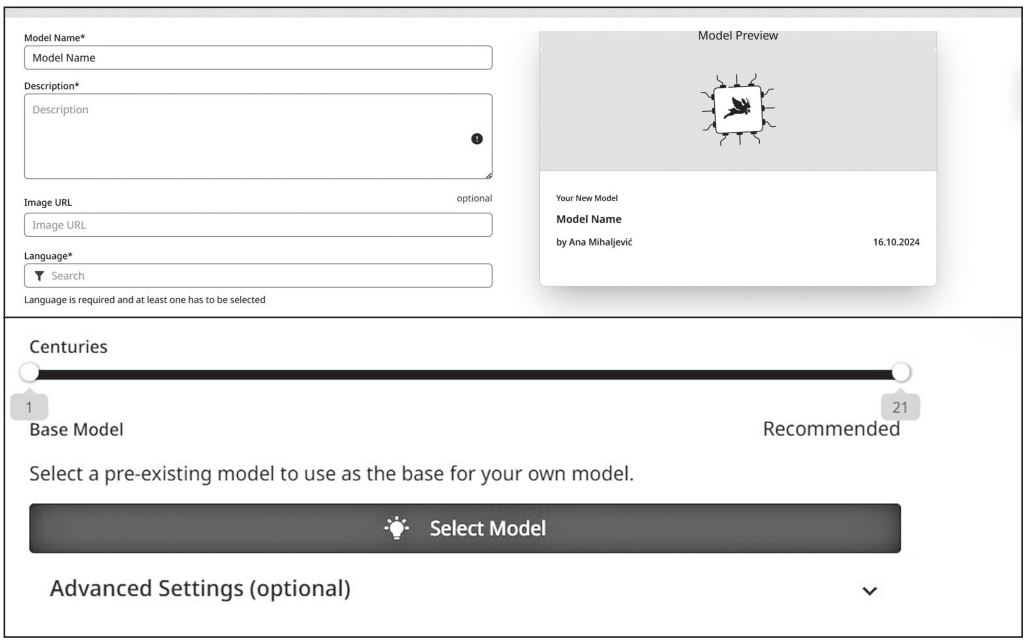
Zatim mora odabrati zbirku slika i prijepisa na kojemu želi trenirati model i niz stranica koje su prepisane te postotak stranica koje želi izdvojiti kao *validation set*, tj. stranice na kojima će se provoditi provjera. Za treniranje modela najbolje je odabrati najjasnije i najčitljivije dijelove teksta, prednost je ako su izabrane stranice iz različitih dijelova

rukopisa jer se pretpostavlja da će one biti leksički raznovrsnije. Što je više teksta u skupu za treniranje, rezultati će obično biti bolji (ako nije riječ o jako teško čitljivim stranicama).



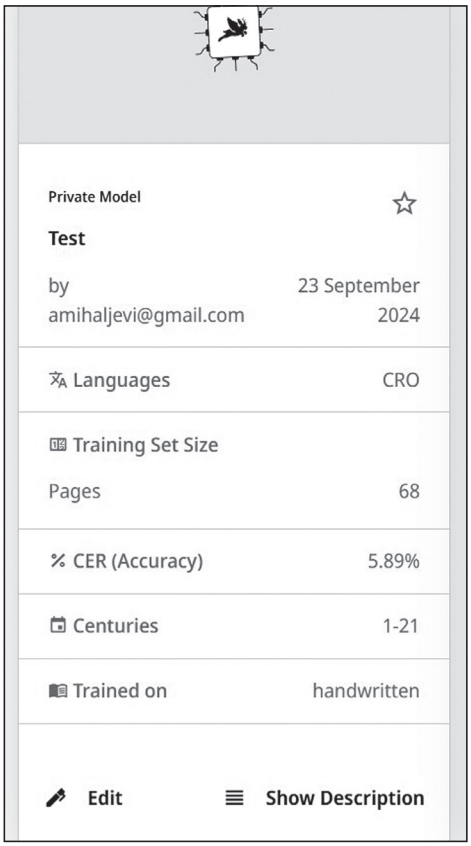
10. slika: Treniranje novoga modela – odabir seta za treniranje i seta za provjeru

Novi je model potrebno nazvati i opisati te definirati za koji je jezik i za izvore iz kojega stoljeća u prvome redu namijenjen.



11. slika: Opis novoga modela

Po završetku treniranja modela sustav donosi informacije o uspješnosti modela na stranicama za provjeru.



12. slika: Rezultati uspješnosti novoga modela

Nova inačica Transkribusa pruža mogućnost i objave gotovoga rukopisa unutar Transkribus Sites.

Transkribus je iznimno koristan alat za sve koji se bave rukopisnim tekstovima. Bitno olakšava i ubrzava rad s takvom građom. Omogućava brže procesuiranje većega broja dokumenata, a novije inačice omogućuju i lakše predstavljanje građe. Detaljnije upute za uporabu Transkribusa dostupne su u videima koje Transkribus objavljuje na stranici na YouTubeu https://www.youtube.com/@transkribus te na stranici za pomoć https://help.transkribus.org/. Transkribus također redovito organizira besplatne mrežne seminare o uporabi aplikacije.