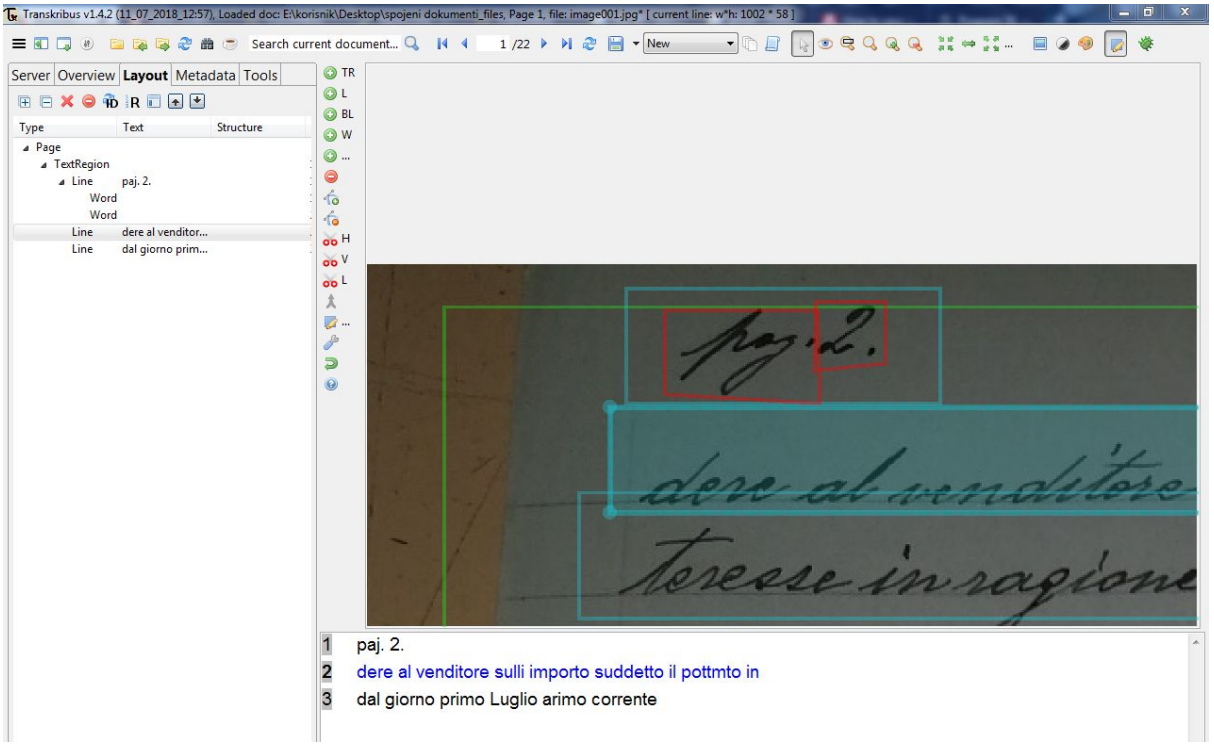
Doktorand Josip Mihaljević prisustvovao je 18. listopada 2018. radionici *Transkribus: optičko prepoznavanje znakova rukopisnih tekstova*. U prvome dijelu radionice, koji se održao u Rektoratu sveučilišta u Zagrebu, Damir Boras, Vlatka Lemić i Hrvoje Stančić govorili su o problemima digitalizacije te optičkoga prepoznavanja znakova i međunarodnim projektima povezanima s digitalizacijom i upravljanjem digitalnim dokumentima poput ICARUS-a.

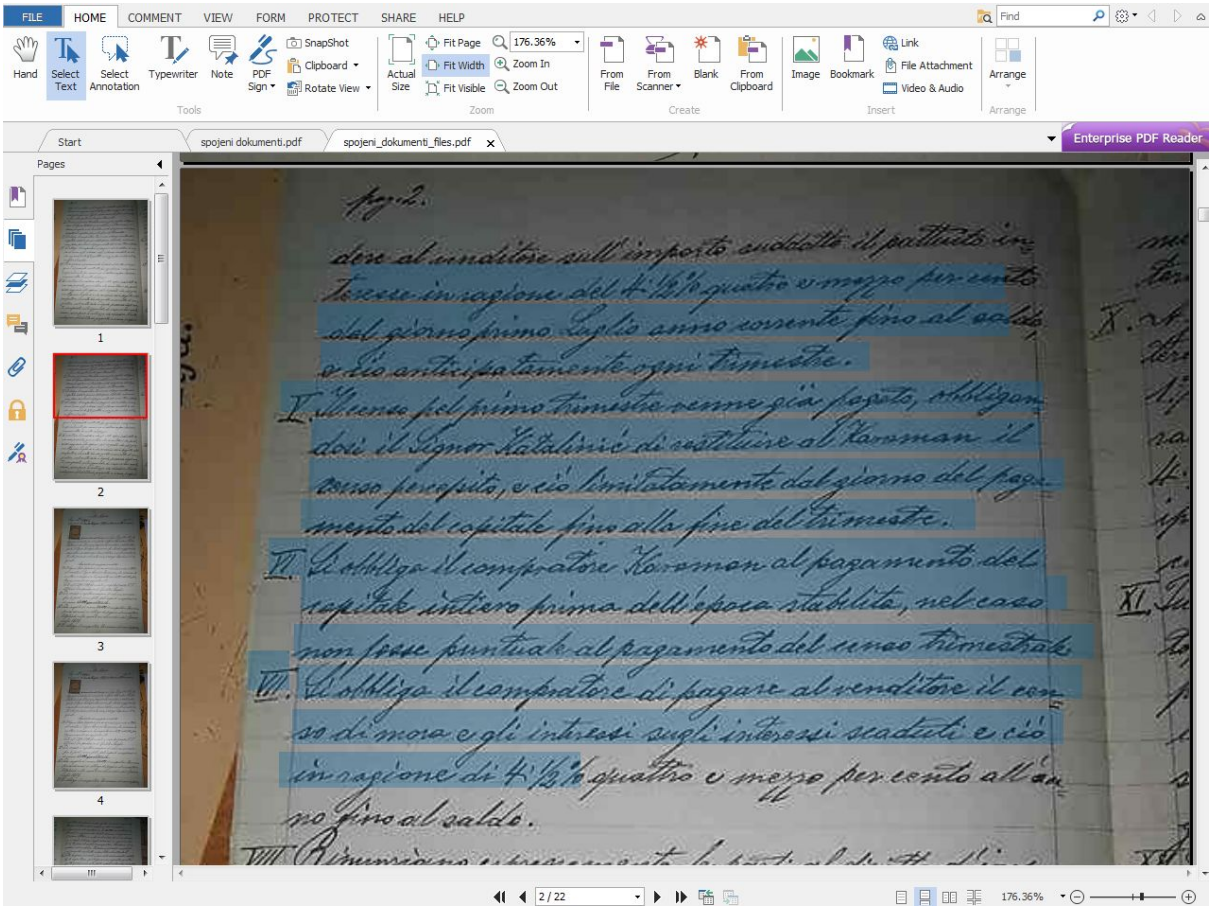
Unutar projekta READ, koji je financiran iz fondova Europske unije, razvio se program Transkribus koji je izrađen za optičko prepoznavanje rukopisnih znakova. Program je nastao kako bi se pronašao način za automatsko optičko prepoznavanje znakova povijesnih dokumenata. Program pokušava automatski prepoznati slova rukopisa na temelju unesenoga skupa podatka kojima se koristi za trening te dodatno omogućuje pretraživanje sadržaja digitaliziranih rukopisa. Podržava i stare rukopise poput hebrejskoga, ćirilice i arapskoga. Program se može besplatno preuzeti te trenutačno ima oko 15 000 korisnika. U programu se pohranjuju slike i prijepisi koji se koriste za trening prepoznavanja znakova. Program može točniji prepoznati znakove ako ima više primjera određenoga rukopisa koji prepoznaje. Uglavnom se program trenira da prepoznaje rukopise određenih osoba. Moguće je stvoriti i modul koji sadržava rukopise više osoba, ali sigurnije je imati odvojene module za prepoznavanje znakova svakoga autora. Preciznost je programa za svaki znak oko 90 posto ako program sadržava dovoljno velik uzorak stranica s obrađenim rukopisom. Pogreške u prepoznavanju znakova mogu se ručno ispraviti. Pretraživati se može s pomoću ključnih riječi. Kao primjer uporabe toga programa spominje se obrada 75 000 zapisa Jeremyja Benthama u Londonu. Digitalizirano je 95 000 stranica za koje je napravljen ručni prijepis koji se koristio za treniranje programa da prepozna ostale znakove skeniranih Benthamovih rukopisa. Za označavanje određenih znakova na stranicama koristio se TEI XML. Bentham je s vremenom oslijepio, pa je njegov rukopis postao manje čitljiv te ga je bilo teže prepoznati. Zbog toga su napravljena dva odvojena modula za trening njegovih rukopisa. Napravljena je stranica na kojoj se može mrežno pretraživati rukopise. Upute za rad s programom napisane su na wiki stranicama. Spominje se i tehnički proces rada u programu te problemi poput izblijedjelih i umrljanih znakova u rukopisu. Često se pojavljuju i tekstovi napisani u dva stupca te tekstovi u tablicama koji se teže mogu automatski obraditi. Kao

algoritam prepoznavanja koristi se Neural Network, koji pamti sekvence podataka za prepoznavanje znakova. *Keyword spotting* tehnologija je koja omogućuje automatski pronalazak određenih riječi u slikama rukopisa. Koristi se u programu za pronalaženje određenih rukopisa koji sadržavaju tu riječ. Inačica program koji je dostupan na službenoj stranici zove se TranskribusX. Predstavljen je i program za digitalizaciju na mobitelu DocScan, koji u kombinaciji s programom TentScan izravno prebacuje slike na Transkribus.

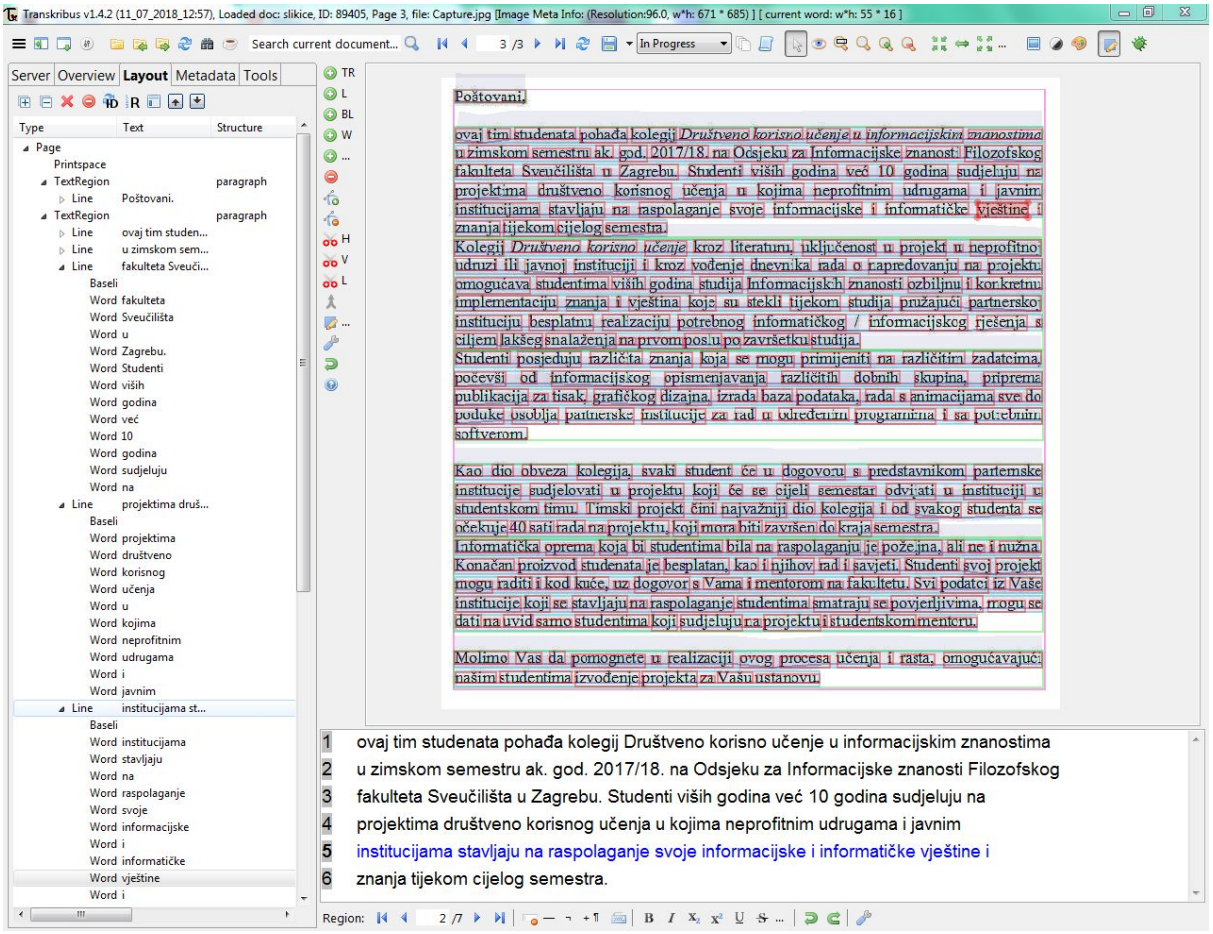
U drugom dijelu radionice radilo se u programu Transkribus. Za korištenje programa prvo je potrebna registracija s pomoću Googleova računa. Svaki polaznik trebao je učitati svoje slike rukopisa te označiti na njima linije teksta za koje se trebao napraviti prijepis. Kasnije se utvrdilo da postoji opcija za automatsko označavanje linija u tekstu te da se mogu upotrijebi postojeći istrenirani moduli za prepoznavanje teksta kako bi se automatski napravilo optičko prepoznavanje teksta. Nažalost, nije bilo mogućnosti prepoznavanja za hrvatskih rukopisnih tekstova, pa točnost prepoznavanja nije bila dobra.



Da bi se napravio vlastiti modul za optičko prepoznavanje znakova određenog rukopisa, korisnik mora za najmanje 50 skeniranih odkumenat imati precizno označene riječi s prijepisom. Nakon toga može se kontaktirati proizvođače programa da stvore modul za prepoznavanje znakova na temelju postojećega uzorka. Obrađene rukopise može se izvesti u različitim formatima poput datoteka Word, .txt, .xml te PDF-a. U PDF dokumentu na slikama rukopisa moguće je kopirati tekst.



Program također ima u sebi integriran program Abby Finereader za optičko prepoznavanje tiskanih znakova. Taj program podržava podržava hrvatski jezik te ima visoku točnost. Program nije besplatan. Postoji, međutim, mrežna inačica koja ima ograničenja u vezi s korištenjem. Unutar Transkribusa može se koristiti besplatno bez ograničenja.



Polaznici su isprobali i Android aplikaciju *DocScan* za optičko prepoznavanje znakova, koja zahtijeva korištenje uređaja *ScanTent. ScanTent* funkcionira kao mali šator u koji se položi dokument ili knjiga za skeniranje. Na vrhu šatora nalazi se podloga s rupom za postavljanje mobilnoga uređaja. Program *DocScan* automatski radi digitalizaciju za svaku novo prepoznatu stranicu. Može primijetiti pokrete ruke te tek uključuje skeniranje kad korisnik mirno položi papir. Skenirani dokumenti mogu se učitati za obradu na stranicama Transkribusa.