U dva klika mišem

ANA MIHALJEVIĆ*

Osnovno o Transkribusu¹

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.

	Back	Save 0 unsaved changes	BrVO_301r_400v - #17 <	200 >	n a	• New -	• •••
E H Ø ▲ @	nasta Masa bitada Stada Stada Stada Stada Stada	амадо апраучерноста и совется справать совется и советс	адиный анитетина, анитетина полого и полого анитетина. Канитетина, анитетика, анитетина санала санала санала и санала изака и санала и санала санала санала санала изака санала санала санала санала санала полого санала санала санала санала и санала санала санала санала полого санала санала санала санала и санала		No transcript	tion for this page R omatic transcription	A+ 2 A- 2
	n dans andre	8 лючальны сирент 9: самоната 19 лочана и прака 19 лочана самоната 19 лочана со прака и прака 19 лочана самоната 19 лочана самоната самоната 19 лочана со прака самоната 19 лочана со прака со прака со прака 19 лочана со прака со прака со прака 19 лочана со прака со прака со прака со прака 19 лочана со прака со прака со прака со прака 19 лочана со прака со прака со прака со прака 19 лочана со прака со пра	כפרישעות האישראינים אוועריאשר אוועריאשר אישראינים אישראינים אישראינים אישראינים אישראינים אישראינים אישראינים א אישראינים אישראינ				Help (3)
ବ ବ ହ t ୧୦ ≺	Addition South the Dickshow Mission Mi	артара со	штания Д. А.Д. 4 «тися корбинкана ча наком янкалийся сласаю ордин на ининого на содисталивается словая в валя изачу чахатолию махипало муст в валя изачу чахатолию махипало на вастика частая применно забеленая как рабинается адника садател пала ча	an and			

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*.

	Text Recognition Layout Fields O	Table O		×
	() ×		Start Recogr	iition
		🌣 Advanced Settings	Credits needed: Available (Personal	-0.25 661 0
 Favorite Models Public Models 13 	Name Search	Train GT Pages	Featured	ID: 51962
Private Models	Universal Lines	24 723	- A A A A A A A A A A A A A A A A A A A	Held
Filter	Mixed Line Orientation			
Q Search	HABE_Baseline new	254		
	Vaybertaytsh.YidTakNL-baseline	228	Universal Lines	
	Danish Newspapers 1750-1850	3 050	Created by Transkribus	6 May 2023
	Text Line Detection in Printed Music Books	217	III Training Set Size	
	Tibetan Pecha	1 530	Pages	24 723

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.

HRVATSKI JEZIK 2024. – broj 4

< Ba	ack Save 0 unsaved changes	BrVO_301r_400v - #17 <	17 200	>	n Progress	· 0 …
	August 2012 Conception of the second	алака ранная ранната чала слада с шылама у алба фалба с сала сала с полите Алько запайства с в собласт Алько запайства с с собласт и с с с с с с с с с с с с с право с с с с с с с с с с с с с с с с с с с		1 2 3 4 5 6 7	Region 2	A+ A- ■
2000 1000 1000 1000 1000 1000 1000 1000	и алабо шчел. Аунин и лице ж.т. ин алаган а	зол аранного индига англина Калина ранного карарарии ин Калина ранного кула уграждийска калина ранного угразованост иналия ула але са са са са са са иналия ула са са са са са са са са иналия ула са са са са са са са са са иналия и са са са са са са са са са са индирата на са са индирата на са		8 9 10 11 12 13 14 15		Hep ③
ري به سنا	CV. HANDRENDER, AND AND COMPLEXED		44	16		0
	אאינטער הינהיה כווואראיקאיניואינאין אינאאראין 27 אאינטער האינטער אינאראין אינאראין אינאראין 28	15 16 17 18 19 19 19 19 19 19 19 19 19 19	-	18 19 20		=

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.

	Text Recognition Layout Fields O	Table O	•		>
	() ×			Start	Recognition 💿
	Language Model			Credits needed: Available (Persona	-1. I 661
 ★ Favorite Models Ø Public Models 224 	Name Search	Words	CER	Language	ed (D: 51170
Private Models (13)	The Text Titan I (Super model)		2.95%	GER, DUT, FRE, FIN, SWE, ENG	
Filter	The German Giant I	15 420 976	8.30%	GER	Same Come
Q Search	The Dutchess I	11 693 499	4.30%	DUT	
Languages	Transkribus Print M1	5 068 310	2.20%	GER, ENG, DUT, FRE, SWE, FIN, POL, ITA, SPA, CZE, SLO, SLO, PO	R. LAT I (Super
🗙 Search	Transkribus French Model 1	1 933 011	7.80%	FRE	
Handwritten or Printed 🛛 👻	Early Portuguese Printing	122 754	2.67%	POR	5 April 2023
Centuries	19th century Danish gothic handwriting v.1.3	557 599	5.33%	DAN	ER, DUT, FRE, FIN

4. slika: Odabir modela za prepoznavanje teksta

Nakon automatskoga očitanja korisnik može u sučelju ispravljati tekst.

K Back Save 0 unsaved changes	BrV0_301r_400v - #17 < 17 200 >	r 🖌 🕒 In Progress 🗸 🕓 🚥	
A C SHUMMATIN YAURTS IDAM	Ana Dunaling granthand	s to êko odênie nêko eo-	
RUDRUNAL BROCHTERK SKE	Har Burger Burger Burger	6 têlesnoe. svlač-	
This a super states and a super states	Res with auth to many the town	7 eĉi se vidit se, nesa-	1
Calls mark mark mark the the wark my		8 moû vlastiû pravoû	1
а Стана в стан	William Contraction of the state of the stat	9 ostaviti bogatustviê	1
EB h ((10) saunasannanna guisanna		10 nb oĉe plati e povolastva ot	
C W Philip Baddin Antes Sunnahiragoo	abbarra A dute pastille) and a	11 vrêĉi, îmže prêžde vsêhь	
12	12 And State State State State State	12 daet se mira, i tvr'dostv-	
13 A.L. month and the first of the second	13 Unide Sch Strength Burght State	13 a vlaća obća zapovédь.	
	AMPARAL STAR STAR STAR START PRACTICE	14 da mir' noset', i tvr'dosta	2
15 15 15 15 15 15 15 15 15 15 15 15 15 1	CRACHOLIN RANGER MILLING MILLING PAR	15 n'no hranet', stran'no lûbinoe	1
REWEIT III E CH. 474. 16 Photos 16	peanares semararen ansembra	16 potrébi strégut, tužda-	Τ.
Main Kallar Koundan and Marken Mar	unicent ARVANDADADAMANERIE EX	17 go of propovédanié c(ása)r(k)stv-	
A DER CERTINEYA ASTU CIMANA	A What by a minane to be mining	s) iš p(a)b(a)skago vzbrapa)tí t.	
M KULANAR HOLD ALA WE WIN BUL		is leaf no domôh' i no urilna	
O BE SEA CITIES AND MARIAMENT	19 V. G. Churchard and A. Carrier Milling of Annual States of Annual State	is na stania zakoni izmano	
mentisent generula incursions	20 m. Luniawaw Dath-AVS BURAWAYS	zu voti, rôč(popu) d(u)tu stvo/ti populuti v	: [
SUBSESTIMPAUSE ATVIN, Dr. 6 D. / 9	1 minuta 16 A. d. an ar ins walfantykann	21 vau. res(pons) d(u)ns s(ve)u napisni vs.	
	22	22 s6 dom : Ideze benu apjustojii. Le-	ı.
	UII8IID AN Laborradhabababa Albusy	23 vise se im razdelaemi ezic-	1
< 24 24 A CUMA AND AND AND AND AND AND AND AND AND AN	24 24	24 Leko oguni jisede žeina edinomu gmiži	·

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.

A+ A-	Configuration	×
-	Tags in collection	Structural tags ~
-	Decide which tags you would like to menu of the image area.	o have available in the context
-	Tag	Visibility
	paragraph	
I	heading	
	caption	
	header	
	footer	
=		

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

A+	Configuration	×
A-	Tags in collection	Textual Tags ~
•	Decide which tags you would like to of the text area.	have available in the tag overlay
-	Tag	Visibility
	abbrev	
	blackening	
	add	
	sic	
	speech	
-		

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

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

Export options	
Standard Export	~
Select formats to be included in your export	
Images	
Docx files	
Transkribus PDF	
Text Files (TXT)	
Page XML	
Export structural elements to Mets O	

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*.

À Tran	skribus	Desk	(‡) Models	🗖 Sites	Jobs
Dashboard	Gallery	Q Search text across all co	ollections	~ 0	ilobal Text Sear
Welcome	to your AI Training Dashboard!			+ 1	rain New Mode

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).

Text Recognition	Model Training D	ata Validation Data	Model Setup	Start	;
	Remove	Title Pages			Next >
	×	Homilijar 🥒 30-33,42	,49-53,56,58-59,61-67		I documents selected - 13065 words
					1
10% of your Training Data w	ill be used as Validation Set o	r choose to manually select your Vali	dation Set		
2 pages Validation 5	et 10% of train data	10% of selected pages Manual Selection >			

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.

Model Name*		Model	Preview
Model Name			
Description*			
Description			H
	0	Ny.	TT
mage URL	optional	Model Name	
Image URL		by Ana Mihaljević	16.10.2024
anguage*			
Y Search			
anguage is required and at least one has to be selected			
Centuries			21 Recommended
Centuries Base Model Select a pre-existing model	to use as the base for your	rown model.	21 Recommended
Centuries Base Model Select a pre-existing model	to use as the base for your	rown model.	21 Recommended
Centuries D Base Model Select a pre-existing model	to use as the base for your	own model.	21 Recommended

11. slika: Opis novoga modela

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

Private Model	٨
Test	¥
by	23 September
amihaljevi@gmail.com	2024
沟 Languages	CRO
🖽 Training Set Size	
Pages	68
% CER (Accuracy)	5.89%
🖬 Centuries	1-21
📭 Trained on	handwritten

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. Transkribus također redovito org/. organizira besplatne mrežne seminare o uporabi aplikacije.