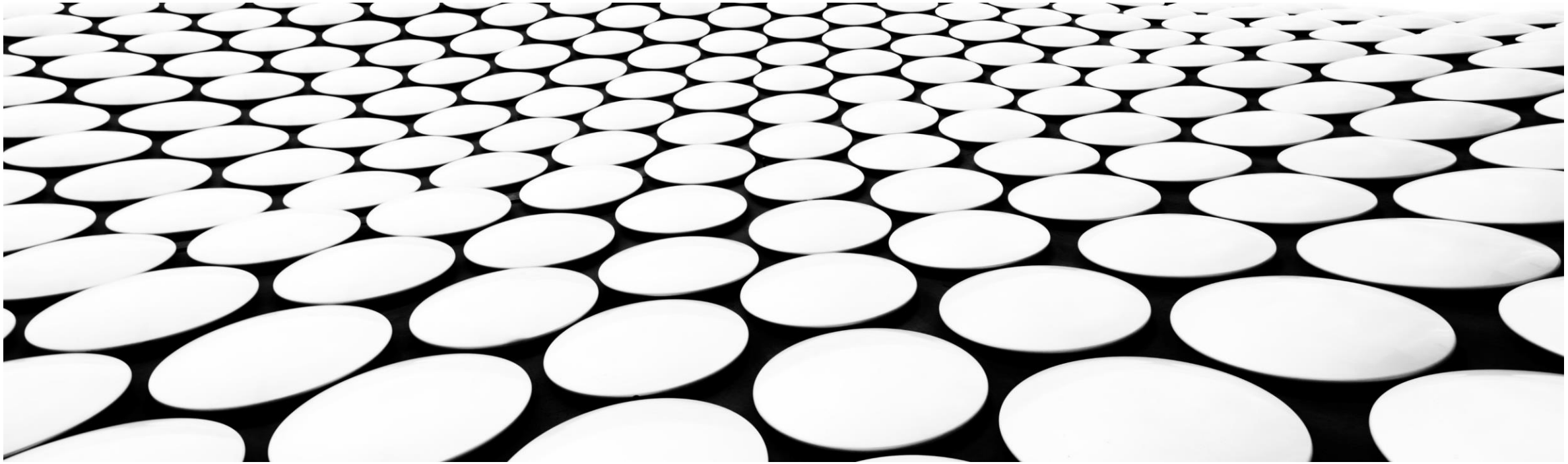
Dušan Katuščák
SVK BB a SU Opava

Transkripcia historických dokumentov v kontexte Digital humanities



Témy prezentácie

1. Digital humanities
2. Projekt SKRIPTOR APVV (ŠVK a UMB Banská Bystrica)
3. Ceny transkripcie
4. Výsledky transkripcie (modely, sprístupnenie)

Digital humanities - Rozpoznávanie textu - Transkripcia

- *Digital humanities* - spoločné *pomenovanie* a prierezovú *metodológiu* pre všetky aplikácie informačných a komunikačných technológií (IKT) v spoločenských a humanitných vedách, odboroch a disciplínach a im zodpovedajúcej praxi.

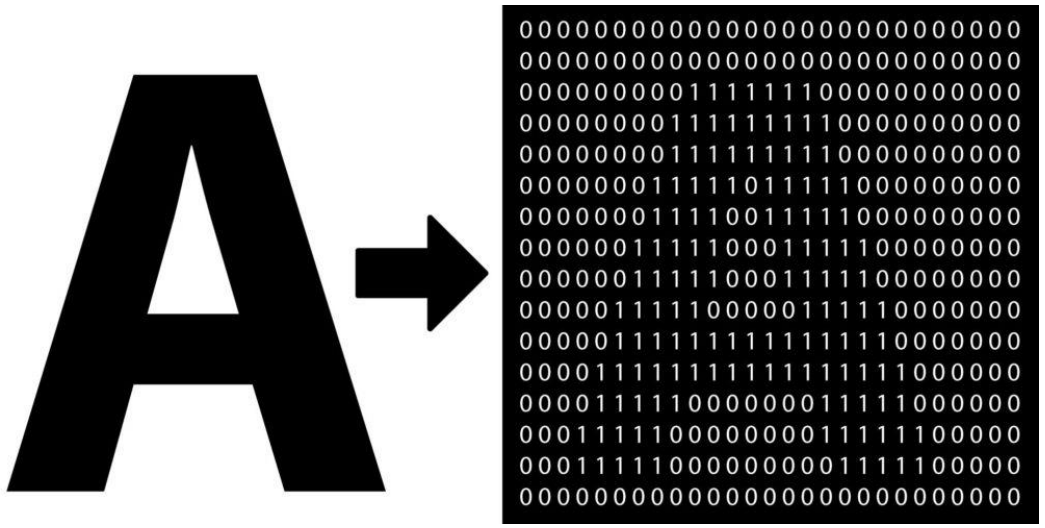
PROJEKT READ (HORIZON 2020); **READ-COOP; TRANSKRIBUS; PROJEKT SKRIPTOR**

Atribúty metodológie *digital humanities*.

a) kooperácia bádateľov;
b) scientizácia v spoločenských a humanitných odboroch;
c) interdisciplinarita;
d) tímovosť a spoluautorstvo (medziinštitučná, medzištátna, univerzity, knižnice, archívy, galérie, múzeá);
e) zapojenie informatikov do výskumu, vzdelávania a sprístupňovania poznatkov;
f) umelá inteligencia

OCR (Optical character recognition)

- Rozpoznávací model je naprogramovaný ručne
- ABBY FR
- Presnosť transkripcie 87%-92%
- Až 100% - možno trénovať
- Staré tlače - problém



HTR+ = Handwritten Text Recognition

- **Umelé neurónové siete (ANN)**
- **Stroj - Rozpoznávací model sa sám učí na príkladoch!**
- **Množstvo dát, trénovanie, učenie stroja**
- **CER (Character Error Rates)** je miera chybovosti znakov (porovnáva pre danú stranu celkový počet znakov (n) vrátane medzier s minimálnym počtom vložení (i), nahradenia (s) a vymazania (d) znakov, ktoré sú potrebné. získať výsledok *Ground Truth*. Ide teda o chyby v porovnaní s presným textom.
- **Vzorec na výpočet CER** je nasledujúci: CER = [(i + s + d) / n]*100. Každá malá chyba v prepise je štatisticky plnohodnotná chyba. To znamená, že každá chýbajúca čiarka, „u“ namiesto „v“, dodatočná medzera alebo dokonca veľké písmeno namiesto malého písmena sú zahrnuté v CER ako chyba.
- **PyLaia** – ako **HTR+** s možnosťou nastavovať parametre

Platforma Transkribus

- *Transkribus expert klient* – Transkribus | AI powered Handwritten Text Recognition (readcoop.eu)
- kľúčový inovatívny nástroj pre transkripciu historických rukopisných a tlačených dokumentov
  - komplexná platforma na digitalizáciu,
  - rozpoznávanie textu podporované umelou inteligenciou
  - prepis a vyhľadávanie historických dokumentov –
  - z akéhokoľvek miesta, kedykoľvek a v akomkoľvek jazyku

Alternatívy Transkribus

- Pred výskumníkmi v budúcnosti stojí úloha vypracovať kritériá hodnotenia funkcionality a kvality nástrojov, aplikácií a platforiem transkripcie. (Metaanalýza)

Existuje celý rad iných nástrojov transkripcie:

- *OCR4all*
- *eScript*
- *Rescribe*
- *Pero.cz*
- *ABBYY Cloud OCR SDK* k nemu vyše 10 alternatív
- *Online OCR*
- *Kofax*
- *Omnipage,*
- *Geekersoft OCR Word Recognition*
- *i2OCR ai*

Ekosystém Platformy Transkribus

- **READ-COOP SCE** (Societas Cooperativa Europeae – SCE), od 1. júla 2019 About us - READ-COOP (readcoop.eu).
  - Cieľ udržať a ďalej rozvíjať platformu *Transkribus*
  - *94 000 používateľov*
  - *40 mil obrazov*
  - *20 mil rozpoznaných strán*
- ***Transkribus Lite*** – jednoduchá transkripcia v prehliadači Transkribus
- ***Read&Search*** – *portál na sprístupnenie zbierok a dokumentov* Read&search - READ-COOP (readcoop.eu)
- ***ScanTent a DocScan –*** *podpora snímania* ScanTent - Professional scanning with your Smartphone (readcoop.eu)


23/02/2022

Ceny a efektívnosť

Manuálna transkripcia – 10-15 eur/strana

Automatická transkripcia – Transkribus ca 0,12 € - 0,14 €/strana s DPH

Transkribus? [Transkribus Credits & Pricing - READ-COOP (readcoop.eu)]

Cena snímania:

ScanTent a DocScan – 1000 strán ? 50 € (0,05 €/strana)

Profesionálne skenovanie 742.. -60 obr. (670 obr) = ca 1000 € s DPH (bez postproc.). Teda ca 1,50 €/strana

Hospodáriť s kreditmi!

Ceny

READ co·op

Transkribus ScanTent read&search The COOP Resources Account 0 EN

500 CREDITS
FREE

Sign up for free

- Get 500 Credits free on signup
- Only applicable to one "collection"
- Can be used with any "engine"

How many pages can I process with this?

Engine	Handwritten Pages	Printed Pages
PyLaia	500	3000
HTR+	400	2500

Pages can be mixed and matched

3 000 CREDITS — 720€ ONE-TIME

Add to cart

- Can be bought **as often as you like**
- Shareable with **any** "collection"
- Can be used with any "engine"

How many pages can I process with this?

Engine	Handwritten Pages	Printed Pages
PyLaia	3 000	18 000
HTR+	2 400	15 000

Pages can be mixed and matched

3 000 CREDITS — 648€ ONE-TIME

Become a Member

- Can be bought **as often as you like**
- Shareable with **any** "collection"
- Can be used with any "engine"

How many pages can I process with this?

Engine	Handwritten Pages	Printed Pages
PyLaia	3 000	18 000
HTR+	2 400	15 000

Pages can be mixed and matched

All prices shown on this page include Austrian **VAT of 20%.**

Suma / kredity	Stroj PyLaia Počet strán rukopis/cena €	Stroj PyLaia Počet strán tlač/cena €	Stroj HTR+ Počet strán rukopis/cena €	Stroj HTR+ Počet strán rukopis/cena €
648 €/3000	3 000/0,216€	18 000/0,036€	2 400/0,27€	15 000/0,043€
1944 €/10000	10 000/0,194€	60 000/0,0324€	8 000/0,24€	50 000/0,038€

Projekt Skriptor

- Na Slovensku sme začali pracovať s platformou *Transkribus* v roku 2017
- Projekt APVV SKRIPTOR – SVK BB a UMB BB (APVV-19-NEWPROJECT-17816 (2020-2024). *Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov)*
- *Transkribus* - používame stroj umelej inteligencie *HTR+* (Handwritten Text Recognition) a *PyLaia*
- Tieto stroje zatiaľ nemôžu okamžite automaticky transkribovať rôzne historické rukopisy
- Učiť stroj na konkrétny typ písma a rukopisu.
- Hlavným cieľom praktických experimentov v projekte SKRIPTOR je v súčasnosti tvorba *modelov* transkripcie.
- Transkripcia geografických slovacikálnych historických zbierok a dokumentov (slovenčina, tiež čeština, latinčina, maďarčina, poľština ai.)
- Tvorbu modelov transkripcie na základe väčších zbierok, ktoré obsahujú stovky a tisíce strán
- Tvoríme veľmi dobré až excelentné modely transkripcie archívnych dokumentov a starých tlačí

Výskumníci SKRIPTOR

- P. Maliniak - transkripcia rukopisných kázní Izáka Abrahamidesa
- K. Kováčová - transkripcie nemeckej rukopisnej kuchárskej knihy z roku 1667
- B. Snopková – editovanie transkribovaných textov
- I. Poláková – editovanie textov, organizačné záležitosti projektu
- P. Kunec a absolvent štúdia histórie M. Katreniak - transkripcii kanonických vizitácií.
- M. Mikušková a L. Nižníková - transkripcii komplikovanej historickej tlače
- O. Tomeček transkripcie novolatinského rukopisu reambulačného protokolu.
- M. Bôbová - digitalizácie vo výskume dejín knižnej kultúry
- A. Kurhajcová - transkripcioa rukopisu J. M. Hurbana.
- I. Nagy - transkripcie Csákósovho katalógu korešpondencie Kohárvovcov
- D. Katuščák – transkripcia historických rukopisov a tlačí

Kabinet digital humanities?

Zameranie:

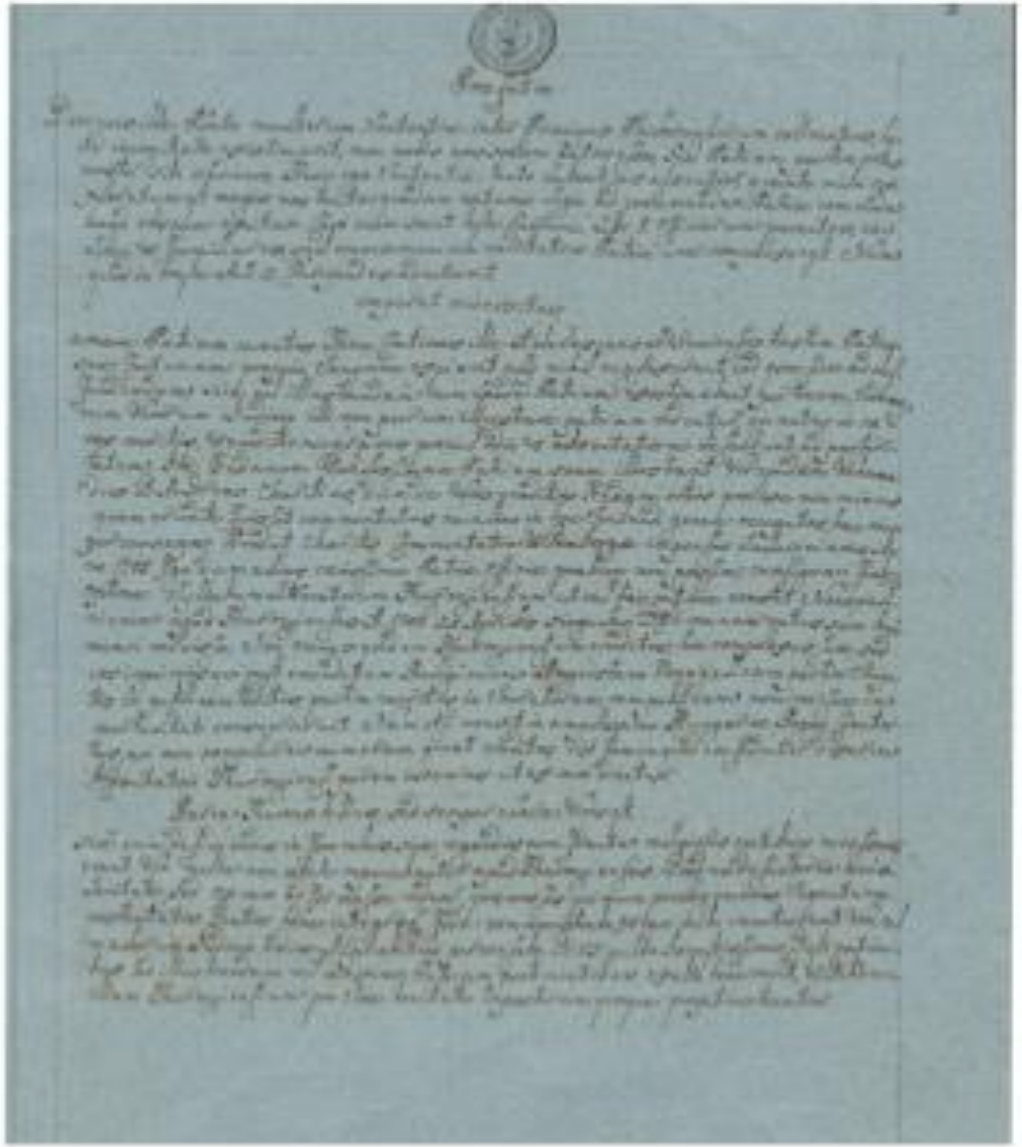
- Výskum, vývoj, vzdelávanie v DH
- Analýzy trendov a nástrojov DH
- Špecifický výskum a vývoj v oblasti transkripcie historických rukopisných a tlačených textov
- Metodika a poradenstvo, expertízy
- Startup a podnikateľský zámer
- Nové projekty výskumu

Atribúty:

- Interdisciplinarita
- Medzinárodná spolupráca (V4?)
- Spolupráca pamäťových a vzdelávacích inštitúcií
- Kooperácia expertov, inštitúcií, krajín !!!

Transkripcia fraktúry (Schwabacher)

Date	OCR method	Train file		Valid file		Accuracy CER		ID
		pages	lines	pages	lines	Train	Valid	model
20210824	OCR base 29418	7	8092	1	888	0,20%	0,91%	36160
20210905	OCR base 29418	9	11231	4	1179	0,18%	1,07%	36358
20210912	OCR base 29418	17	20805	5	2252	0,39%	0,44%	36550
20210913	OCR base 36550	7	2462	3	276	0,03%	1,78	36607



Rukopis Martina Laučeka v zbierke Collectanea (od krasopisu k voľnejšiemu rukopisu)

Model 'ANDREJ KMET'

Name:

ANDREJ KMET

Language:

slo

Description:

Andrej Kmeť (November 19, 1841, Szénásfalu, Austrian Empire (today Bzenica, Slovakia) - February 16, 1908, Turócszentmárton (today Martin, Slovakia)) was a Slovak botanist, ethnographer, archaeologist, and geologist. [1] He identified several new species of plants and created a herbarium with 72,000 specimens. He was one of the first researchers who carried on modern archaeological excavations in Central Europe. In 1892, he founded the Slovak Learned Society (Slovak: Slovenská učená spoločnosť), which later became nucleus of the Slovak Academy of Sciences. He was also known for his bitter criticism of alcoholism. Andrej Kmeť was interred in the National Cemetery in Martin. The collection contains personal letters written by Andrej Kmeť to various addressees. Letters of Andrej Kmeť from the Archives of the Slovak National Museum in Martin were used for transcription. The experiment is part of applied research - project APVV-19-NEWPROJECT-17816 (2020-2024): Innovative disclosure of written heritage of Slovakia through the automatic transcription of historical manuscripts. (Matej Bel University and the State Scientific Library in Banská Bystrica). ScanTent and DocScan with Samsung 6 were used for digitization. I used 185 pages Trains SET (28672) words and 26 pages Validation Set (4703 lines) of GT to create the ANDREJ KMEŤ model. Slovak language with diacritical marks, which

Parameters:

Nr. of Epochs 200

Document Type: Handwritten

Show advanced parameters...

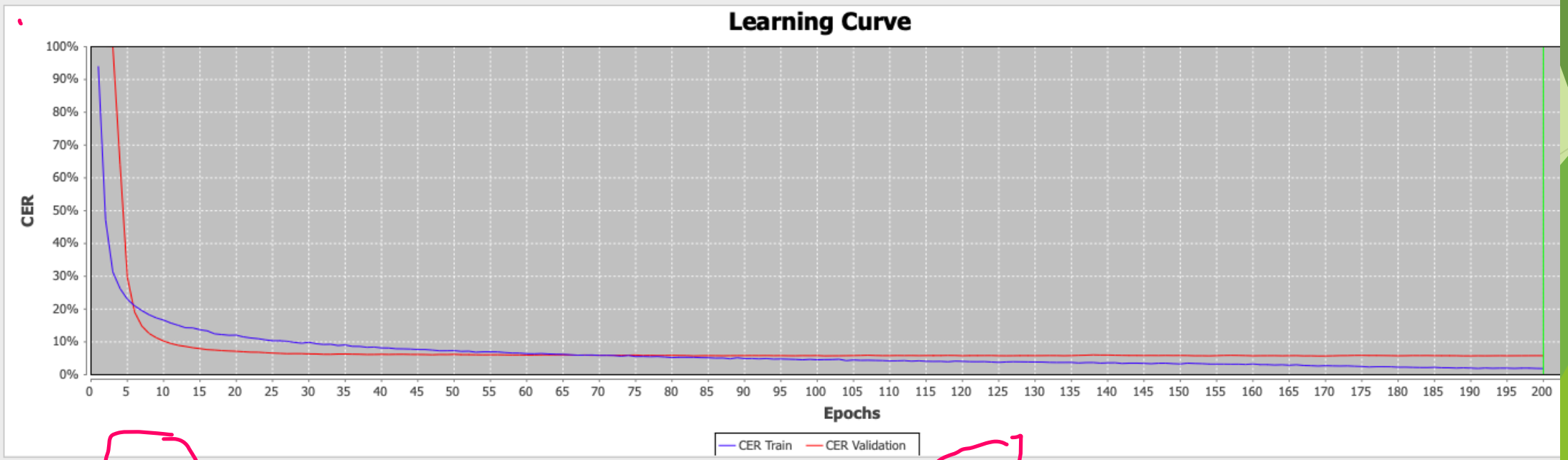
Nr. of Words:

28672

Nr. of Lines:

4703

Save | Show Train Set | Show Validation Set | Show Characters



CER on Train Set: 1.87%

CER on Validation Set: 5.79%

MODELY TRANSKRIPCIE RUKOPISNEJ KOREŠPONDENCIE ANDREJA KMEŤA

MODELS OF TRANSCRIPTION OF ANDREJ KMEŤ'S HANDWRITTEN DOCUMENTS

DATE	Method	Model	Training set		Validation set		CER accuracy set		Number of cycles	CER/WER	
		ID	pages	lines	pages	lines	training	validation	(epochs)	characters	words
20190125	CITlabHT+	**10135**	125	22549	26	3497	1.15%	3.37%	200	5.97%	21.60%
20190201	CITlabHT+	**10410**	152	29905	46	4499	1.27%	2.97%	200	6.19%	22.13%
20190205	CITlabHT+	**10548**	166	29411	46	4573	1.37%	1.84%	200	5.91%	21.87%
20201012	CITlabHT+	**26809**	111	18071	98	2921	0.44%	7.25%	500	6.08%	21.87%
20210410	CITlabHT+	**31888**	119	19291	13	3126	1.15%	5.16%	200	3.77%	12.27%
20210821	CITlabHT+	**36009**	185	28672	26%	4703	1,8%	5.79%	200	2.48%	7.73%

Search for...

Documents | Fulltext (Solr) | Tags | **KWS (CITlab)**

Search in: Current document | Partial Matches | Case-sensitivity | Expert Syntax

Confidence Threshold: 0.05

Queries

Keyword 1

Search

Search Results

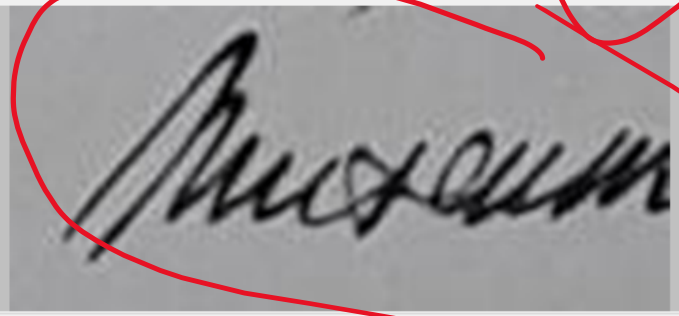
Created	Status	Queries	Duration
08.10.22 16:46:49	Completed	"Medvecký" (4)	2.04 sec.
08.10.22 16:35:38	Completed	"Holuby" (9)	2.87 sec.
07.10.22 20:52:21	Completed	"Prenčov" (28)	2.56 sec.
07.10.22 20:50:03	Completed	"MUSEUM" (27)	8.44 sec.
07.10.22 20:48:42	Completed	"Martin" (21)	2.58 sec.
07.10.22 20:47:24	Completed	"Stiavnica" (5)	3.61 sec.
07.10.22 20:33:47	Completed	"SITNO" (10), "KUCHTA" (5), "NIE...	9.14 sec.

Keyword Spotting Results

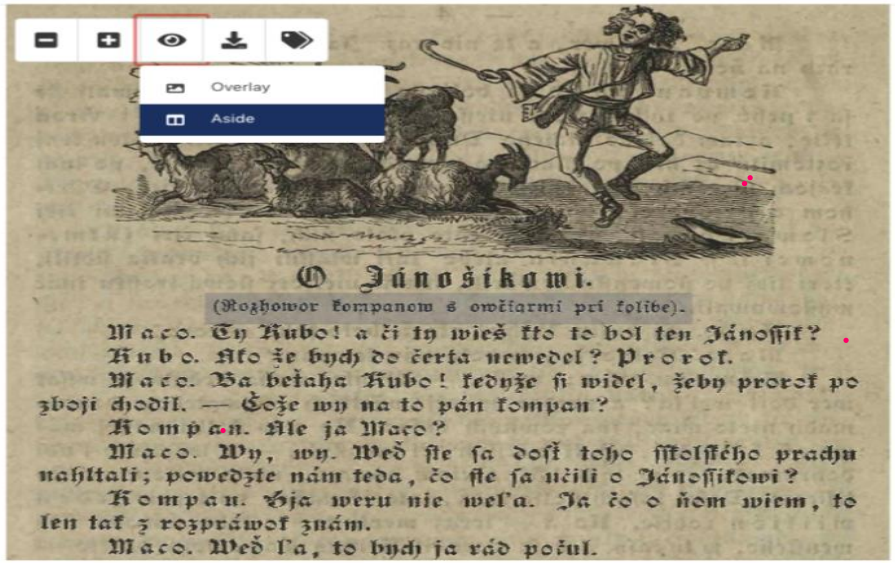
"MUSEUM" (27 hits)

Confidence	Page Nr.	Line transcription	Prev...
0.9561	81	p. Dr. Šamkovi pre Museum, a ten že	
0.9375	71	staré obrazy pre Museum z bočného oltá	
0.9293	100	Vecam od Vás pre Museum teším sa	
0.9176	69	tozár nemá rád Museum ani teraz	
0.8739	100	Museum, ale bol kunšt dostať ju! Lebo	
0.8715	95	pre Museum nemohli predať staré roč.	
0.8585	27	škoda pre náš národ a pre naše Museum, že	
0.8406	113	či nám zhabú museum, alebo od	
0.8403	82	alebo Museum vína.. Prečo nepošlú di	
0.7498	15	naň a či ho zadržíme pre Museum, poťažme	
0.6647	48	museum.	
0.6646	105	Museum kúpil som za 20 zl.	
0.3914	131	niekto napíše, že musem chodiť po drače	
0.3616	105	Dodek daroval Museumu	
0.2918	13	Teraz kúpil som pre meseum ostrovida	

Preview

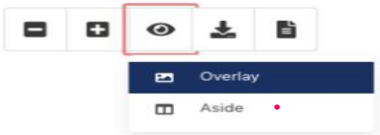


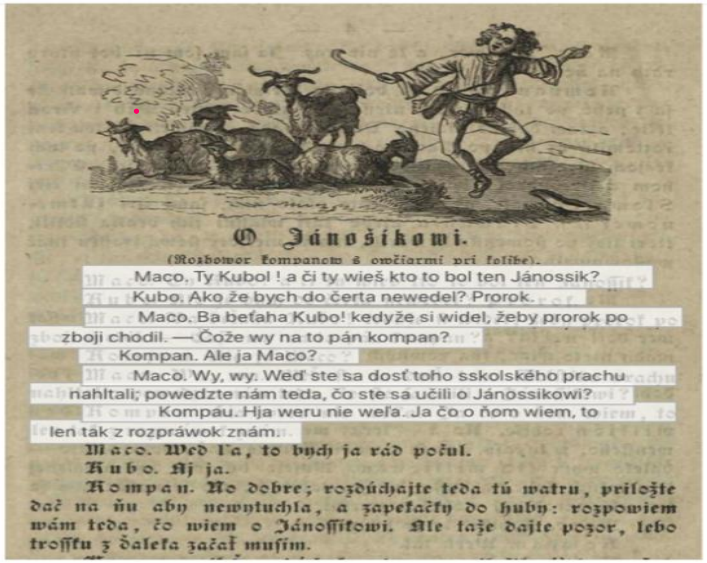
Close

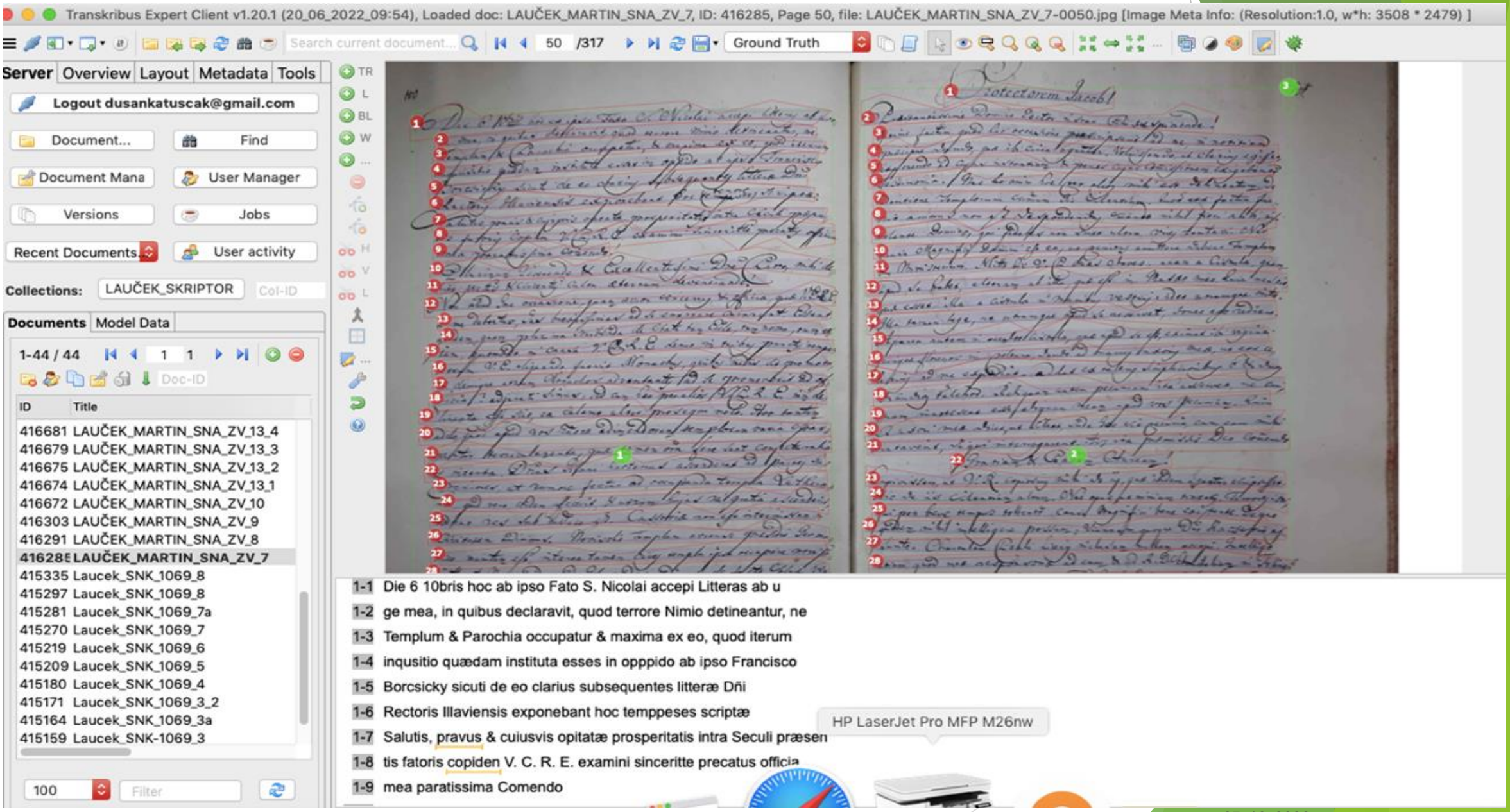


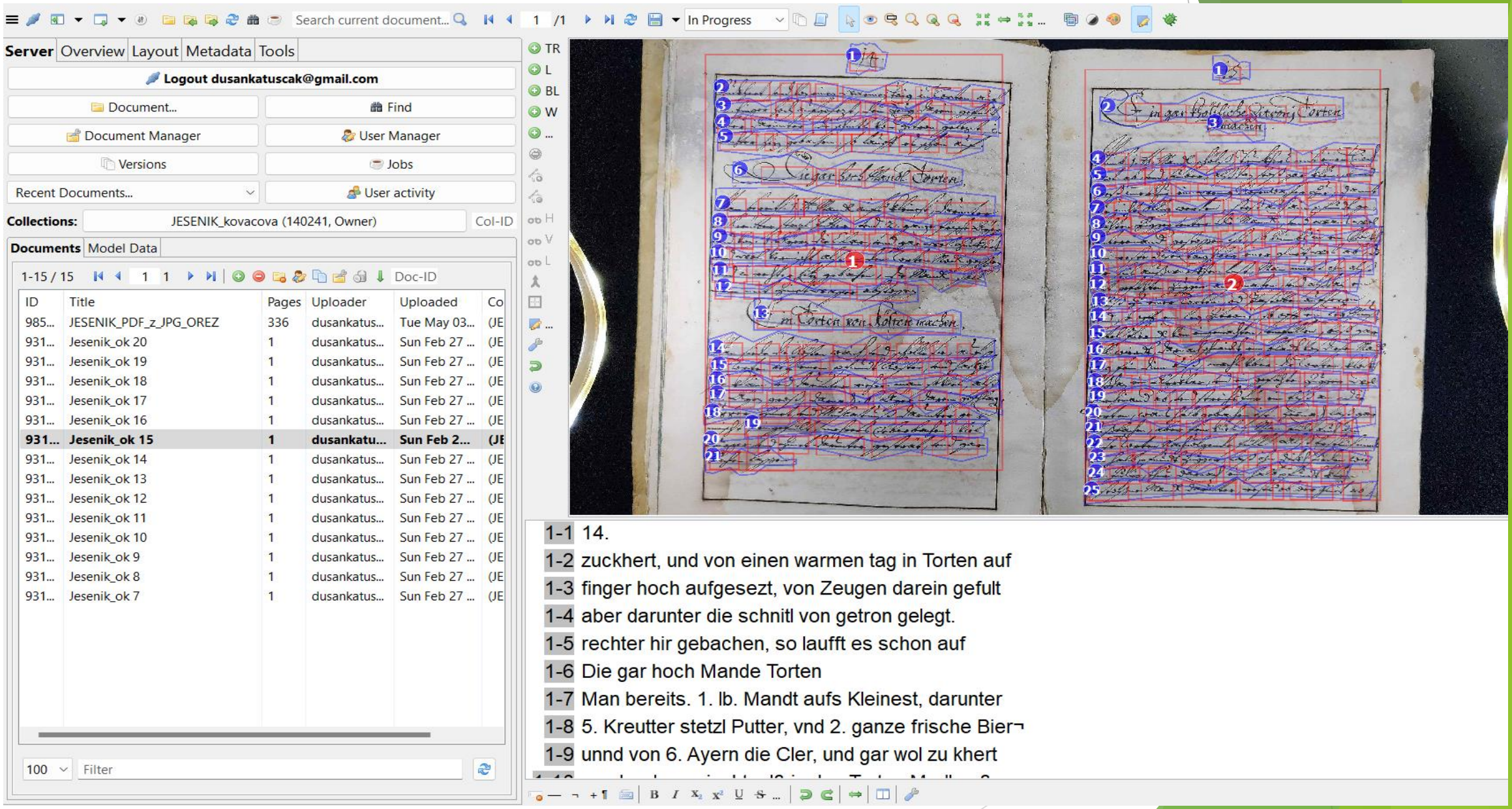
O Jánošíkowi.
(Rozhowor kompanow s owčiarmi pri kolibe).
Maco. Ty Kubo! a či ty wieš kto to bol ten Jánossik?
Kubo. Ako že bych do čerta newedel? Prorok.
Maco. Ba beťaha Kubo! kedyže si widel, žeby prorok po zboji chodil. — Čože wy na to pán kompan?
Kompan. Ale ja Maco?
Maco. Wy, wy. Weď ste sa dosť toho sskolského prachu nahltali; powedzte nám teda, čo ste sa učili o Jánossikowi?
Kompau. Hja weru nie weľa. Ja čo o ňom wiem, to len tak z rozpráwok znám.
Maco. Weď ľa, to bych ja rád počul.
Kubo. Aj ja.
Kompan. No dobre; rozdúchajte teda tú watru, priložte dač na ňu aby newytuchla, a zapekačky do huby: rozpowiem wám teda, čo wiem o Jánossikowi. Ale taže dajte pozor, lebo trossku z ďaleka začať musím.
Kompan, (keď rozdúchal watru, a priločil ráždia na ňu).
No začnite teda.

read &search Browse Search About Language
Home • Browse • Jánošík
< 3 32 Go >






Transkribus Expert Client v1.20.1 (20_06_2022_09:54), Loaded doc: LAUČEK_MARTIN_SNA_ZV_7, ID: 416285, Page 50, file: LAUČEK_MARTIN_SNA_ZV_7-0050.jpg [Image Meta Info: (Resolution:1.0, w*h: 3508 * 2479)]
Search current document...
50 /317
Ground Truth
Server
Overview
Layout
Metadata
Tools
Logout dusankatuscak@gmail.com
Document...
Find
Document Mana
User Manager
Versions
Jobs
Recent Documents
User activity
Collections:
LAUČEK_SKRIPTOR
Col-ID
Documents
Model Data
1-44 / 44
Doc-ID
ID
Title
416681 LAUČEK_MARTIN_SNA_ZV_13_4
416679 LAUČEK_MARTIN_SNA_ZV_13_3
416675 LAUČEK_MARTIN_SNA_ZV_13_2
416674 LAUČEK_MARTIN_SNA_ZV_13_1
416672 LAUČEK_MARTIN_SNA_ZV_10
416303 LAUČEK_MARTIN_SNA_ZV_9
416291 LAUČEK_MARTIN_SNA_ZV_8
416285 LAUČEK_MARTIN_SNA_ZV_7
415335 Laucek_SNK_1069_8
415297 Laucek_SNK_1069_8
415281 Laucek_SNK_1069_7a
415270 Laucek_SNK_1069_7
415219 Laucek_SNK_1069_6
415209 Laucek_SNK_1069_5
415180 Laucek_SNK_1069_4
415171 Laucek_SNK_1069_3_2
415164 Laucek_SNK_1069_3a
415159 Laucek_SNK-1069_3
100
Filter
1-1 Die 6 10bris hoc ab ipso Fato S. Nicolai accepi Litteras ab u
1-2 ge mea, in quibus declaravit, quod terrore Nimio detineantur, ne
1-3 Templum & Parochia occupatur & maxima ex eo, quod iterum
1-4 inqusitio quædam instituta esses in opppido ab ipso Francisco
1-5 Borcsicky sicuti de eo clarius subsequentes litteræ Dñi
1-6 Rectoris Illaviensis exponebant hoc temppeses scriptæ
1-7 Salutis, pravus & cuiusvis opitatæ prosperitatis intra Seculi præsen
1-8 tis fatoris copiden V. C. R. E. examini sinceritte precatus officia
1-9 mea paratissima Comendo
HP LaserJet Pro MFP M26nw


Search current document...
1 /1
In Progress
Server
Overview
Layout
Metadata
Tools
Logout dusankatuscak@gmail.com
Document...
Find
Document Manager
User Manager
Versions
Jobs
Recent Documents...
User activity
Collections:
JESENIK_kovacova (140241, Owner)
Col-ID
Documents
Model Data
1-15 / 15
Doc-ID
ID
Title
Pages
Uploader
Uploaded
985... JESENIK_PDF_z_JPG_OREZ 336 dusankatus... Tue May 03...
931... Jesenik_ok 20 1 dusankatus... Sun Feb 27 ...
931... Jesenik_ok 19 1 dusankatus... Sun Feb 27 ...
931... Jesenik_ok 18 1 dusankatus... Sun Feb 27 ...
931... Jesenik_ok 17 1 dusankatus... Sun Feb 27 ...
931... Jesenik_ok 16 1 dusankatus... Sun Feb 27 ...
931... Jesenik_ok 15 1 dusankatu... Sun Feb 2...
931... Jesenik_ok 14 1 dusankatus... Sun Feb 27 ...
931... Jesenik_ok 13 1 dusankatus... Sun Feb 27 ...
931... Jesenik_ok 12 1 dusankatus... Sun Feb 27 ...
931... Jesenik_ok 11 1 dusankatus... Sun Feb 27 ...
931... Jesenik_ok 10 1 dusankatus... Sun Feb 27 ...
931... Jesenik_ok 9 1 dusankatus... Sun Feb 27 ...
931... Jesenik_ok 8 1 dusankatus... Sun Feb 27 ...
931... Jesenik_ok 7 1 dusankatus... Sun Feb 27 ...
100
Filter
TR
L
BL
W
1-1 14.
1-2 zuckhert, und von einen warmen tag in Torten auf
1-3 finger hoch aufgesezt, von Zeugen darein gefult
1-4 aber darunter die schnitl von getron gelegt.
1-5 rechter hir gebachen, so laufft es schon auf
1-6 Die gar hoch Mande Torten
1-7 Man bereits. 1. lb. Mandt aufs Kleinest, darunter
1-8 5. Kreutter stetzl Putter, vnd 2. ganze frische Bier¬
1-9 unnd von 6. Ayern die Cler, und gar wol zu khert

Rozpracované dokumenty a zbierky

- Andrej Kmeť - ca 3000 s. - rukopis - prevažne slovenčina
- Martin Lauček - min 20 000 s. - rukopis - prevažne latinčina
- kuchárska kniha z roku 1667 - Archív Jeseník – 800 s., rukopis, nemecký kurent
- *Mestečko Tisovec*. Kurentálny protokol - Rukopis, slovenčina, ca 500 s. 1780
- *Školská zápisnica*, rukopis, slovenčina, 1836, ca 60 s. 1836 – rukopis, slovenčina, maďarčina
- Časopis Lužica 1909 – tlač, vlastný model transkripcia – presnosť 99,2% (bežné OCR – 87%-92%)