

POKROK V TRANSKRIPCI HISTORICKÝCH RUKOPISŮ

Dušan Katuščák
Lukáš Němec
Vojtěch Říha



**SLEZSKÁ
UNIVERZITA**

FILOZOFICKO-
PŘÍRODOVĚDECKÁ
FAKULTA V OPAVĚ

Úvod

- Historické staré a vzácne tlače, strojopisy a hlavne rukopisy spravidla nie je možné uspokojivo transkribovať pomocou OCR
- Prichádza na pomoc umelá inteligencia
- V snahách sprístupniť historické písomné dedičstvo sa koncentruje pozornosť výskumníkov na *transkripciu* a strojové učenie s použitím *konvolučných neurónových sietí*
- Ide o proces, v ktorom sa nasnímaný *obrázok* mení na *text*.

Ciel' prezentácie

1. Vysvetliť metodiku tvorby modelov transkripcie v platforme Transkribus
2. Informovať o výsledkoch projektu Študentskej grantovej súťaže na Slezskej univerzite v Opave (FPF, Oddelenie knihovníctva)
3. Prezentovať výsledky experimentov študentov

Prečo modely?

- Modely slúžia na transkripciu historických textov
- Využíva sa umelá inteligencia
- Na vytvorenie modelu je potrebné stroj naučiť, čo má robiť
- Učenie prebieha tak, že sa manuálne pripraví tréningový set (Train set) a validačný set (Validation set)
- Strany textu je potrebné prepísať čo najpresnejšie do kvality GT (Ground Truth)
- Spustí sa proces trénovania
- Výsledkom trénovania je MODEL
- Na základe čiastkových modelov je možné pripraviť univerzálne supermodely

Platforma Transkribus

Ako to funguje? Všetko je o učení umelej inteligencie!

Dokument 10 000 strán

65

10

10 000 strán

Ground Truth (GT)

Cvičné strany (25-75 strán/5000-15000 slov - transkribovať: model/ručne

Trénovanie modelu: napr. 75 strán

Rozdelíme na: TRAIN SET a VALIDATION SET

TRAIN SET: napr. 65 strán

VALIDATION SET: napr. 10 strán

Spustíme učenie: napr. 250 cyklov (Výsledok je: CER/WER)

Model použijeme na automatickú transkripciu 925 strán

Tvorba modelu automatického rozpoznávání bohemikálního rukopisného textu v platformě Transkribus

Tvorba modelu automatického rozpoznávání bohemikálního rukopisného textu v platformě Transkribus

Podmínky:

- časové okno 18. – 20. století
- žánrová diferenciac
- různé rukopisné styly a autoři
- regionální sounáležitost

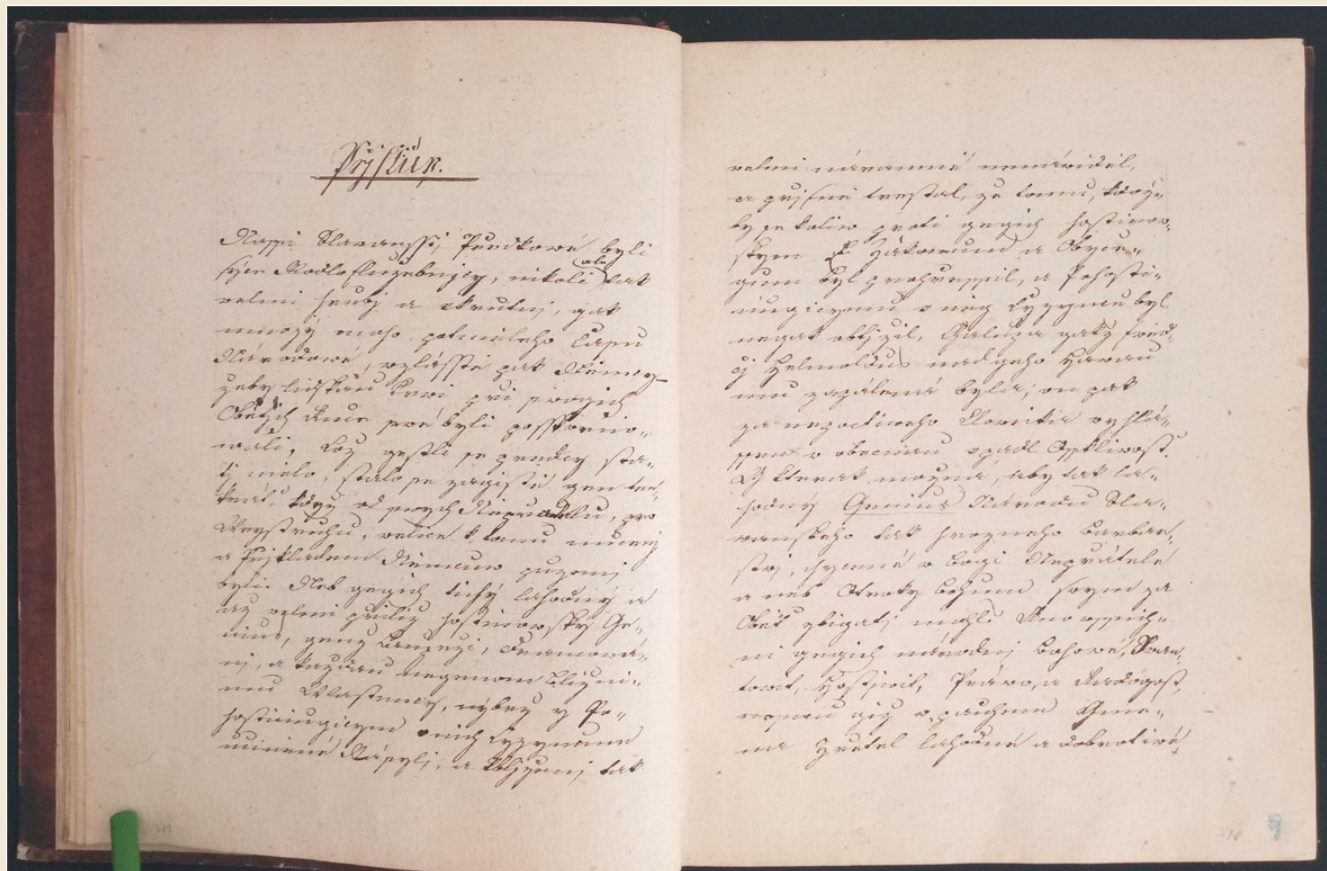
Tvorba modelu automatického rozpoznávání bohemikálního rukopisného textu v platformě Transkribus

Josef Heřman Agapit Gallaš: 1756 – 1840; Hranice,
Přerov
(přírodní vědy, mytologie, národopis)

Otakar Jaroš: 1912 – 1943; Hranice
(vojenství)

František Polášek: 1757 – 1824; Příbor
(náboženské texty)

Mýtické povídky o bozích a bohyních moravských Slovanů



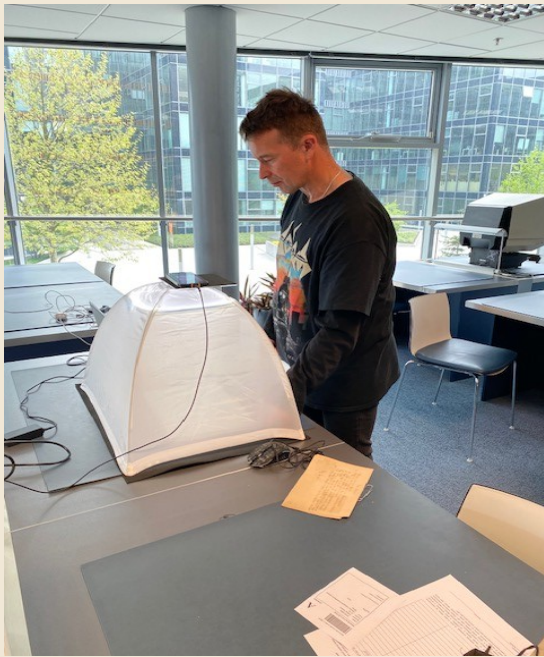
Výzvy:

- Špatná čitelnost původního textu.
- Množství pravopisných chyb.
- Blednutí atramentu.
- Různé písarské styly.

Východisko:

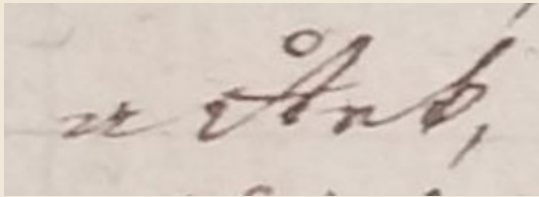
- Dílčí model s chybovostí 8,31 %
- Agregovaný model s chybovostí 6,55 %, který odstranil řadu nejednoznačností.

Mýtické povídky o bozích a bohyních moravských Slovanů

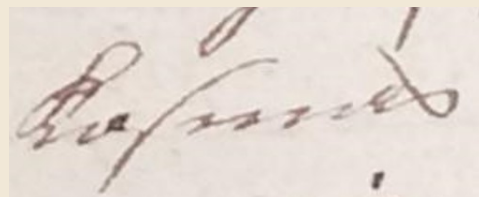


Zachraňujeme historické kulturní dědictví

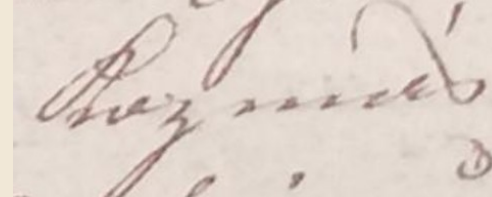
Mýtické povídky o bozích a bohyních moravských Slovanů



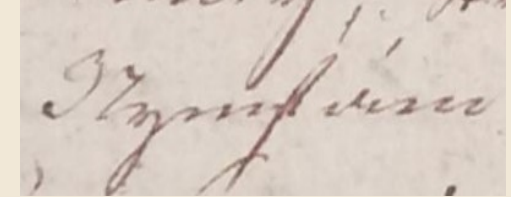
a Řek,



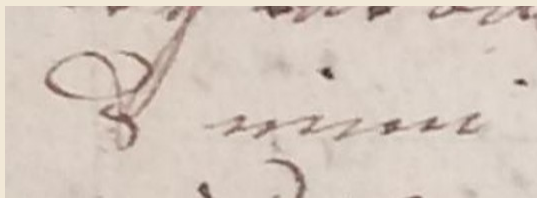
Kosmás



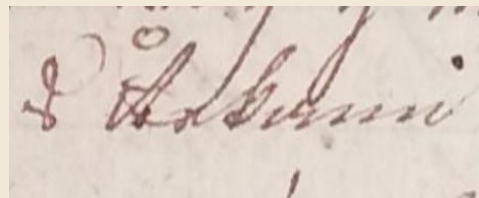
Kozmás



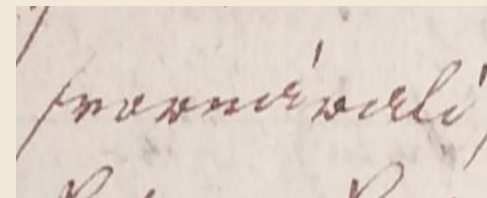
Nymfám



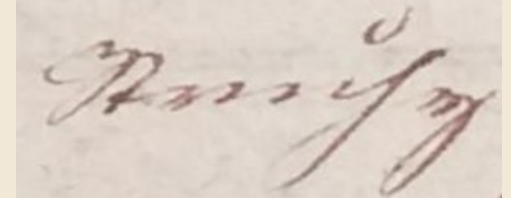
s nimi



s Řekami



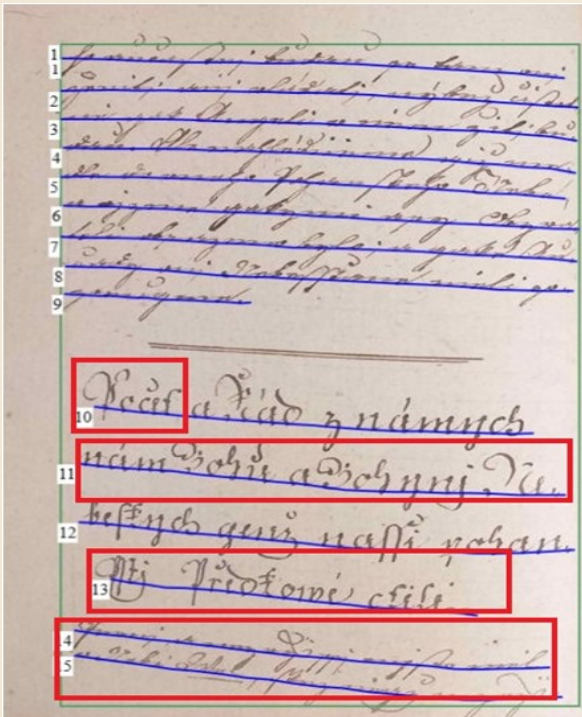
srownáwali ,



Strúhy

Mýtické povídky o bozích a bohyních moravských Slovanů

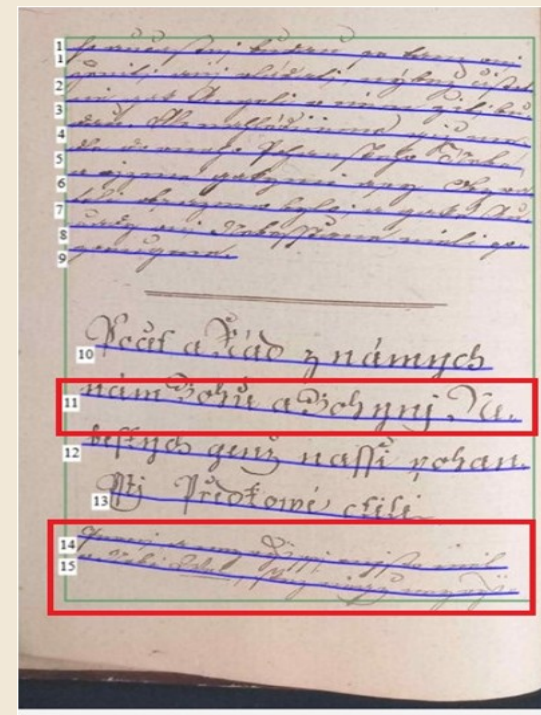
Region 1



1 ho aučafntj búdaú fe tam onj
 2 ženitj anj vládatj, nýbrž číftot-
 3 ně gak Angeli w něm žitj bú-
 4 daú. Ale nahlédněme giž me-
 5 dle do onoho Pohánfkeho Nebe
 6 a wizme gakými afy Obywa-
 7 teli obfazeno bylo; a gaké Au-
 8 řady onj Nebefťtané měli, po-
 9 zorúgmne.
 10 Počt a Řád známých
 11 nám Bohú a Bohynj, Te-
 12 beřkych genž nařfi pohan-
 13 ř Pedewé ctel.
 14 Prnj a neygwřfj mjřtommel
 15 Nb bi Wel, frnz něgž eřwřř-

Model Mystic Absolut s chybovostí 8,31 %

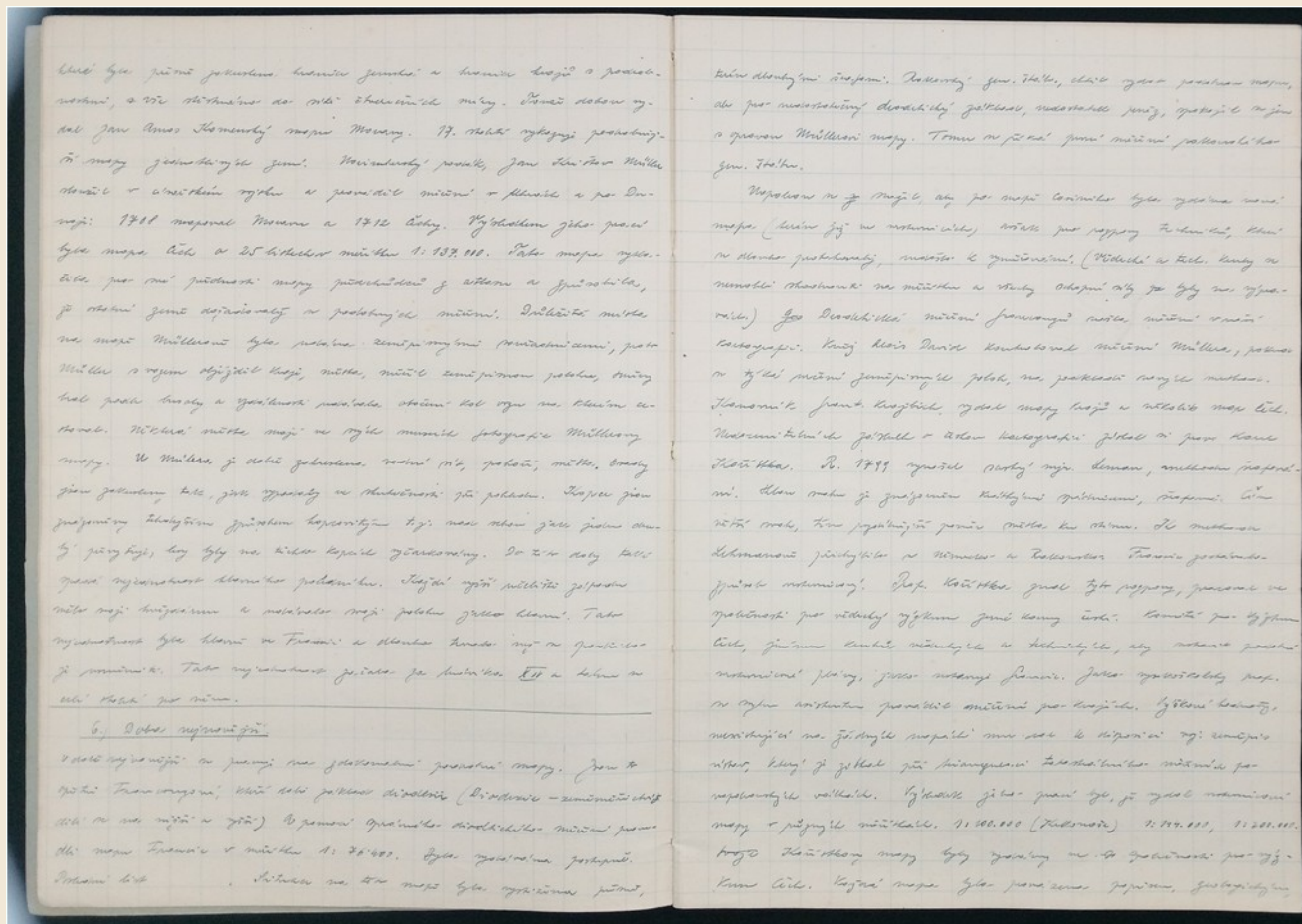
Region 1



1 ho aučafntj búdaú fe tam anj
 2 ženitj anj vládatj, nýbrž číftot
 3 ně gak Angeli w něm žitj bú-
 4 daú. Ale nahlédněme aiž me-
 5 dle do onoho Pohánfkeho Nebe
 6 a wizme gakými afy Obywa-
 7 teli obfáženo bylo; a gaké Au-
 8 řady onj Nebefťtané měli, po-
 9 zorúgmne.
 10 Počet a Řád známých
 11 nám Bohú a Bohynj, Ae-
 12 beřkych genž nařfi pohan-
 13 řtj Předkowé clili.
 14 Prwnj a negwřřřfj mjřto měl
 15 w Nebi Wel řrz něgž neygwřř-

Agregovaný model s chybovostí 6,55 %

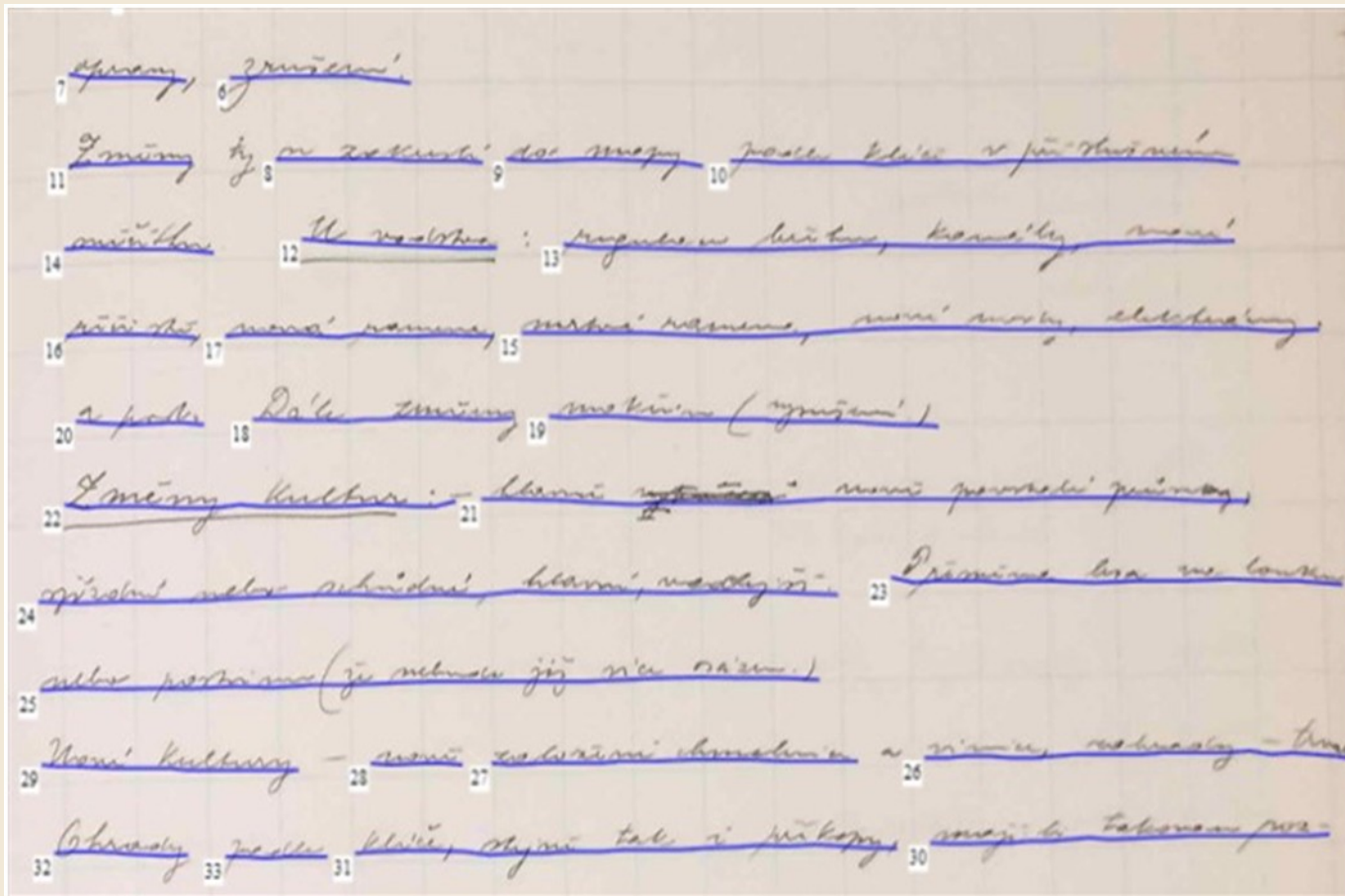
Otakar Jaroš: Nauka o terénu (školní sešit)



Výzvy:

- Drobné písmo.
- Špatně čitelné.
- Degradace papíru.
- Dílčí vybledlost psaného textu.
- ČTVEREČKOVANÝ PAPIR.

Otakar Jaroš: Nauka o terénu (školní sešit)



Výzvy:

- Drobné písmo.
- Špatně čitelné.
- Degradace papíru.
- Dílčí vybledlost psaného textu.
- Chybná segmentace textu
- ČTVEREČKOVANÝ PAPÍR.

Otakar Jaroš: Nauka o terénu (školní sešit)

Details

Name:

Creator:

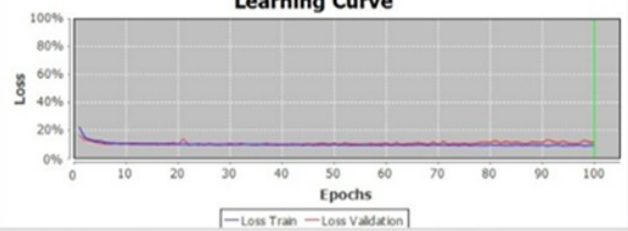
Description:

Parameters:

Document Type: Show advanced parameters...

Nr. of Words: Nr. of Lines:

Learning Curve



Loss on Train Set Loss on Validation Set

Layout Analysis Configuration

Selected model:

Baseline detection settings

Minimal baseline length Param-Value:

Baseline accuracy threshold Param-Value:

Use trained separators Param-Value:

Max-dist for merging baselines Param-Value:

Image scaling Param-Value:

Region detection settings

Method

Nr of Text-regions Param-Value:

Východisko:

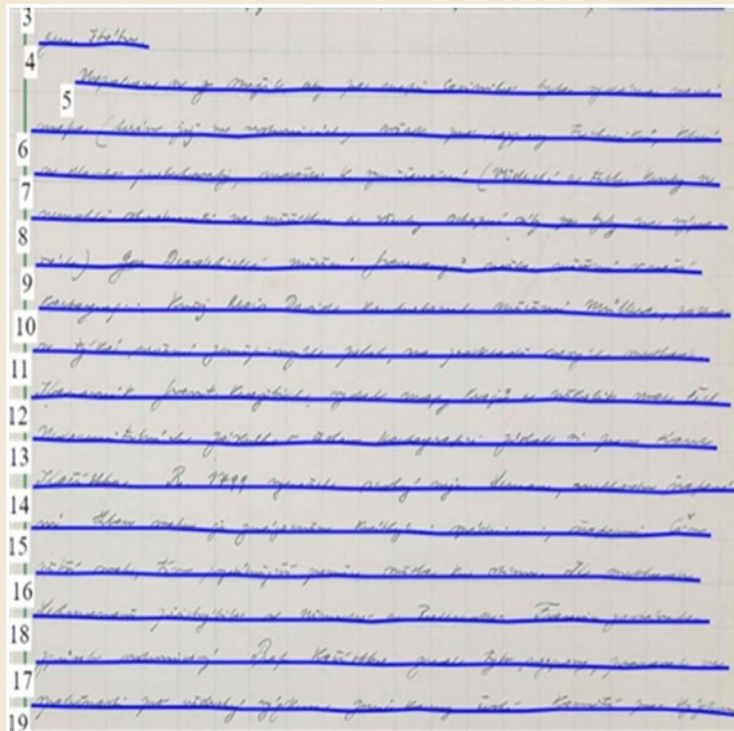
- Experiment s vytvořením modelu na segmentaci textových linek s uvedenými parametry.

Otakar Jaroš: Nauka o terénu (školní sešit)



Odborné vědecké soustředění v Jazernici...víkendový pracovní oddechový pobyt účastníků projektu.

Otakar Jaroš: Nauka o terénu (školní sešit)



4 gen. štábu.	::
5 Napoleon se je snažil, aby po mapě lasiného byla vydána nová	::
6 mapa (terén již ve vrstevnicích, avšak pro rozpory techniků, které	::
7 se dlouho protahovaly, nedošlo k vyměřování (Vědecké a tech. kruhy;	::
8 nemohli shodnouti na měřítku a všechny schopné síly pe byly na	::
9 výpra-	::
10 vách.) Jes Deodetická měření francouzů našla měření v naší	::
11 kartografii. Kněž Alois David kontroloval měření Müllera, pokud	::
12 se týká určení zeměpisných poloh, na podkladě nových method.	::
13 Kanovník frant. krajbich vydal mapy krajů a několik map čech.	::
14 Nedocenitelných zásluh o českou kartografii získal si prov koue	::
15 Kořístka. K. 1799 vynošel saský mjr. Lemon, metodu šrafová	::
16 ní. Klon svahu je znázorněn krátkými spádnicemi, šrafamá. Čín	::
17 větší svah, tím rozdílnější poměr světla ku stínu. K methode	::
18 způsob vrstevnicový. Pof. Kořístka znal tyto rozpory, pracoval ve	::

- Ukázka experimentálního ověření naší hypotézy s potvrzením správnosti úsudku o nutnosti použití výchozího „pomocného“ modelu pro segmentaci textových linek v případě automatického rozpoznávání rukopisu na „čtverečkovaném“ podkladě.

Agregovaný model Finale 2.0

Celkové parametry modelu

Details

Name: Language:

Creator:

Description:

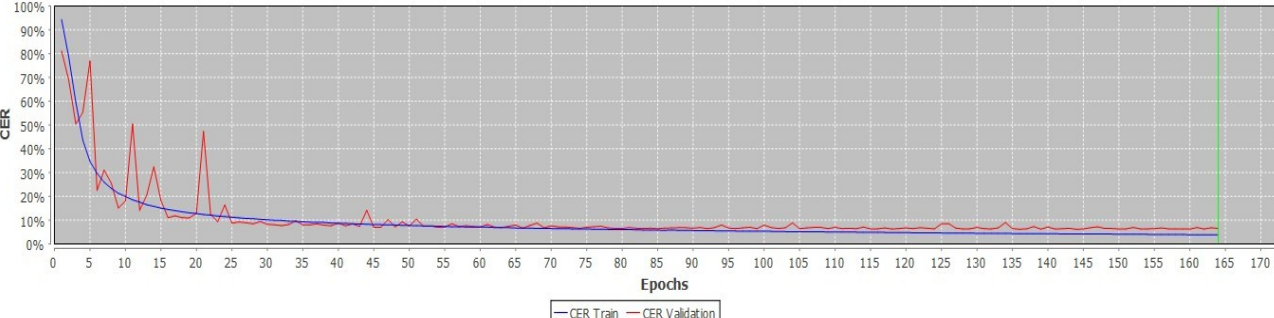
Parameters:

Max epochs	250
Early stopping	20
Epochs trained	164
Learning rate	0.0003
Batch size	24

Document Type:

Nr. of Words: Nr. of Lines:

Learning Curve



CER on Train Set: CER on Validation Set:

Doba trénování: 11 hod. 26 min.

Počet epoch: 164 (bylo nastaveno 250)

Počet stran v kvalitě GT: 514

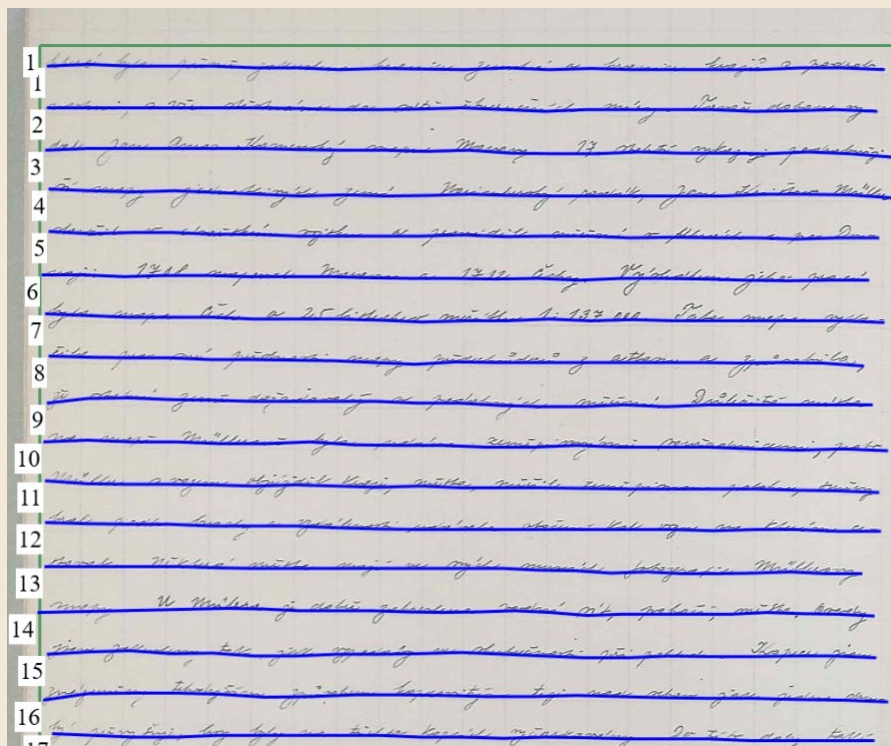
Počet slov trénovací sady/ ověřovací sady:
87.945 / 10.808

Počet řádků trénovací sady / ověřovací sady:
14.530 / 1.785

Celková chybovost modelu: 6,56 %

Otakar Jaroš: Nauka o terénu (školní sešit)

Ukázka použití modelu Finale 2.0



Region 1

1 které byla přesně zakreslena hranice zemská a hranice krajů s podrob-
 2 nostmi, s vše stětnáno do sítě čtverečních míry. Tonže dobou vy-
 3 dal Jan Amos Komenský mapu Moravy. 17. stobetí vykazuje podrobněj.
 4 sí mapy jednotlivých zemí. Norimberský rodák, Jan Křištov Müller
 5 sloužil v čísařském vojsku a prováděl měření v Ulhrách a po Dn-
 6 naji 1708 mapoval Moravu a 1712 Čechy. Výsledkem jeho prací
 7 byla mapa lech o 25 listech v měřítku 1: 137.000. Tato mapa vytla-
 8 čila pro své přednosti mapy předchůdců z atlasu a způsobila
 9 že ostatní země dožadovaly se podobných měření. Důležitá místa
 10 na mapě Müllerově byla udána zeměpisnými souřadnicemi, proto
 11 Müller s vozem objížděl kraje, města, měřil zeměpisnou polohu, směry
 12 bral podle busoly a vzdálenosti udávala otočení kol vozu na kterém ce-
 13 stoval Některá města mají ve svých muscích Jstografie Müllerovy
 14 mapy U můlera je dobře zakreslena vodní síť, pohoří, města. osady
 15 jsou zakresleny tak, jak vypodaly ve skutečnosti při pohledu. Kopce jsou

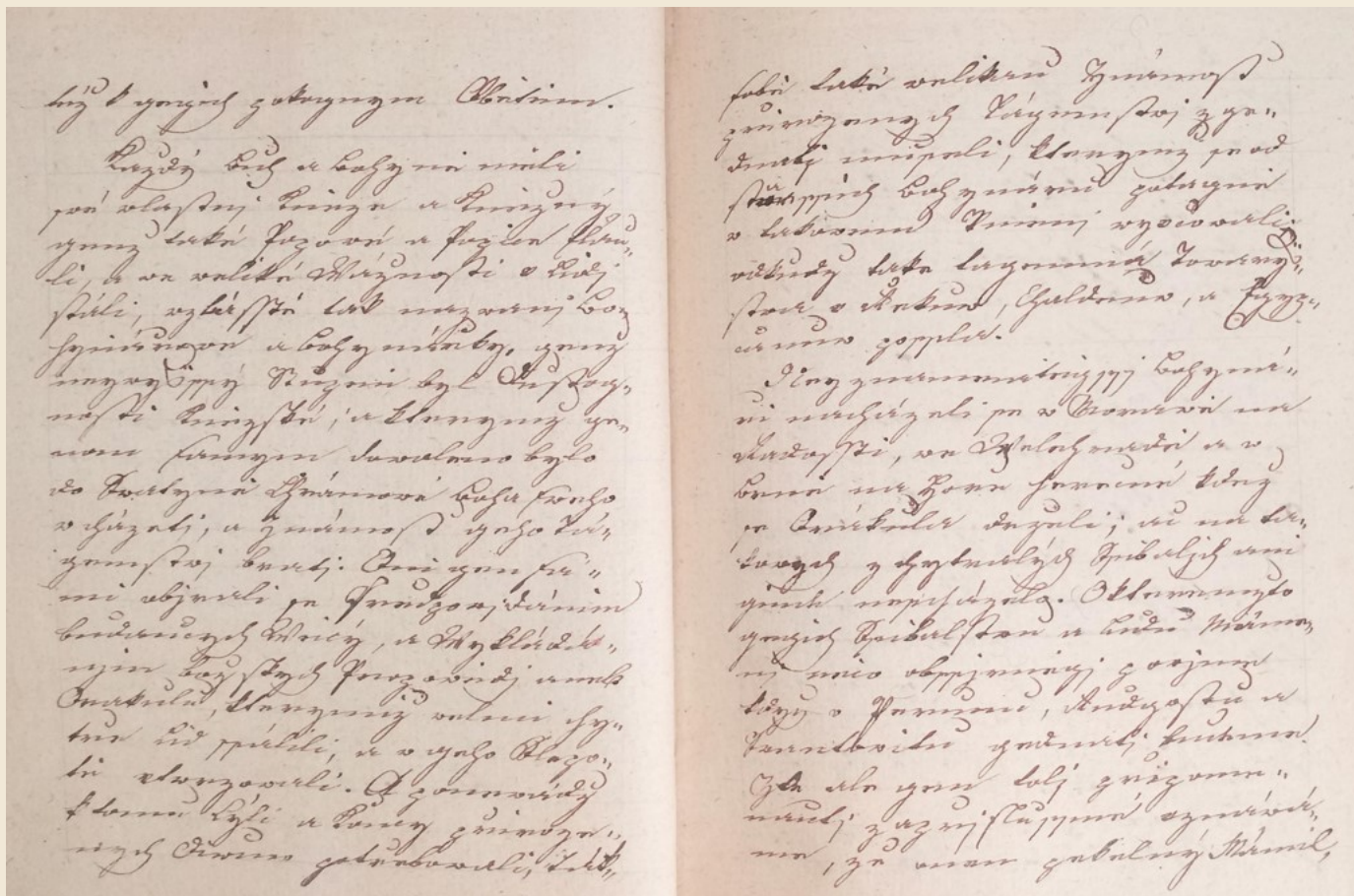
Josef Heřman Agapit Gallaš : Mystické povídky o bozích a bohyních

Ukázka použití modelu Finale 2.0

	Region 1
1	a neb Pop pohodlně seděj a wffe což se
2	w Chrámě bliže nj djlo widěj a flyffe-
3	tj mohlj neb Uffi gegj byly tak mjftr-
4	ně Sformowané že y flabo proně-
5	ffená Sliwa zřetedlně od něg flyfa-
6	ná bytj mohla. Proftředkem onoho
7	Túlipanta, a neb Zwonečka Kwjtk
8	genž se mezy geho Rtami wynachá-
9	zel, mohl sedjcy wněm bohynář ne
10	gen wffechno widěj nýbrž y také
11	aúftná Orakúla wydawatj čímž
12	powěrečnj Pohané, neznage Popúw
13	fwych keykljřkych Proftředkúw a
14	Sfalby w Slepotě fwogj welmi upew-
15	nen byl. Zdaliž wfichni Slawani
16	

Agregovaný model Finale 2.0

Ukázky rukopisných vzorků:



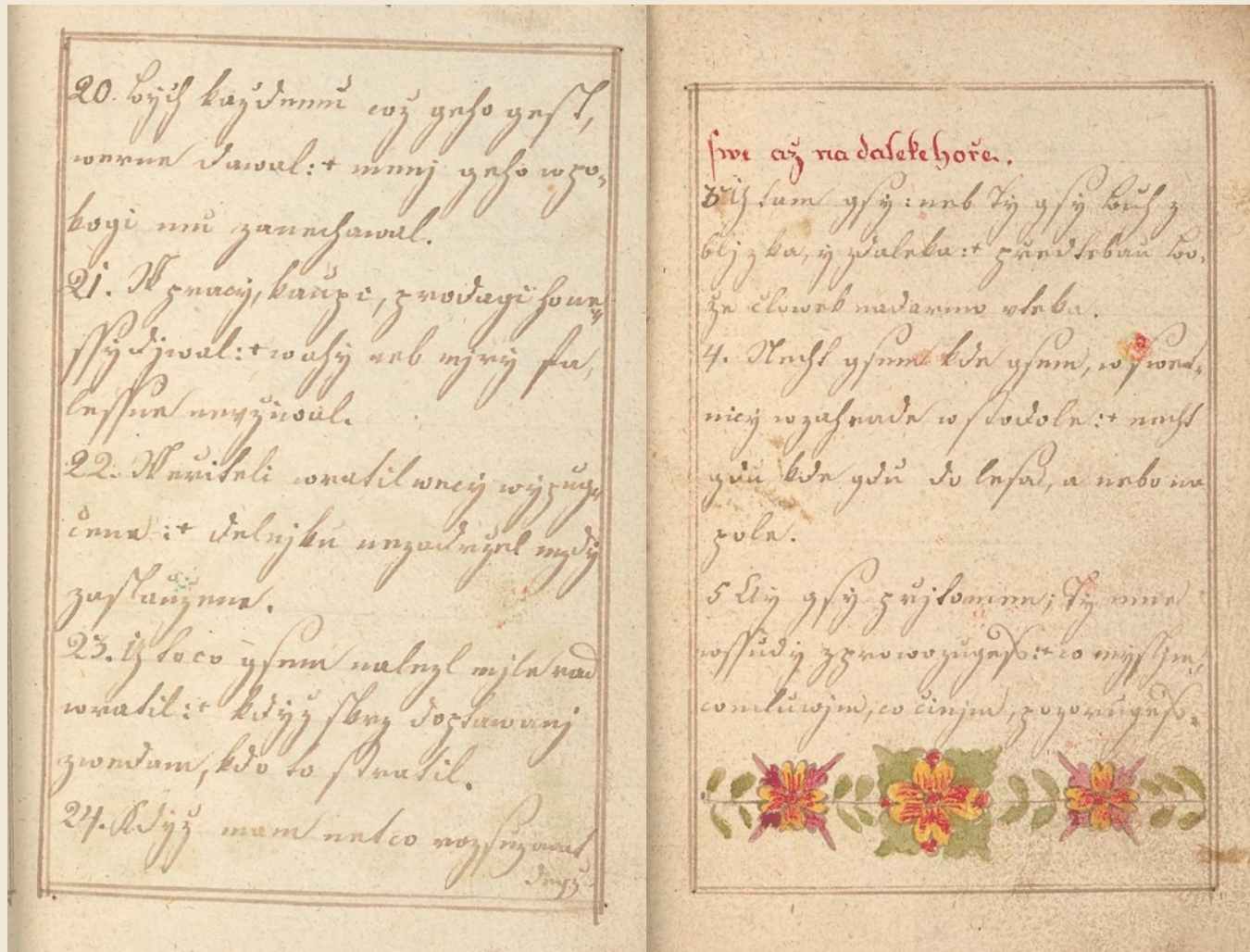
Josef Heřman Agapit Gallaš
Mythické povídky o bozích a bohyních
moravských Slovanů

Uloženo v:

MZA Brno, G 11, sign. 838, čeština, papír,
rukopisná kniha, originál, vázáno v
tvrdých, polokožených deskách, šířka 195
mm, výška 250 mm, stopy po pův. pag.,
starší fol. 125; stará sign.: Schr. 224, pův.
287, červ.

Agregovaný model Finale 2.0

Ukázky rukopisných vzorků:



František Polášek
Pravé poznání boha...

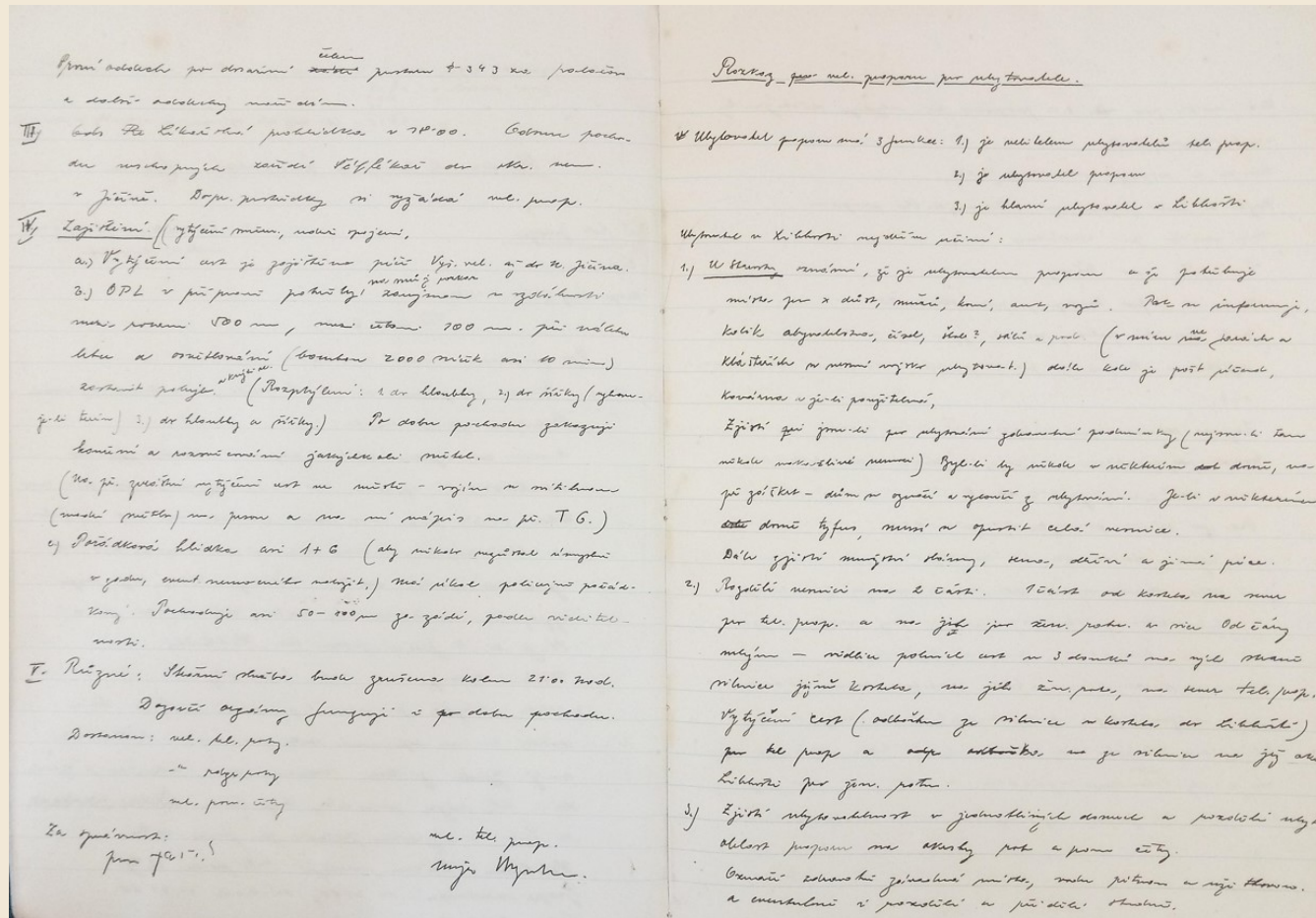
Uloženo v:

Sbírková knihovna Vlastivědného muzea v
Olomouci, Sign. K-24073, dostupný z:

https://new.manuscriptorium.com/hub/catalog/default/detail/single/manuscriptorium%7CVMO-VMOK_24073_0U6ABL2-cs?lang=cs

Agregovaný model Finale 2.0

Ukázky rukopisných vzorků:



Otakar Jaroš
Nauka o terénu, školní sešit, linkovaný
papír

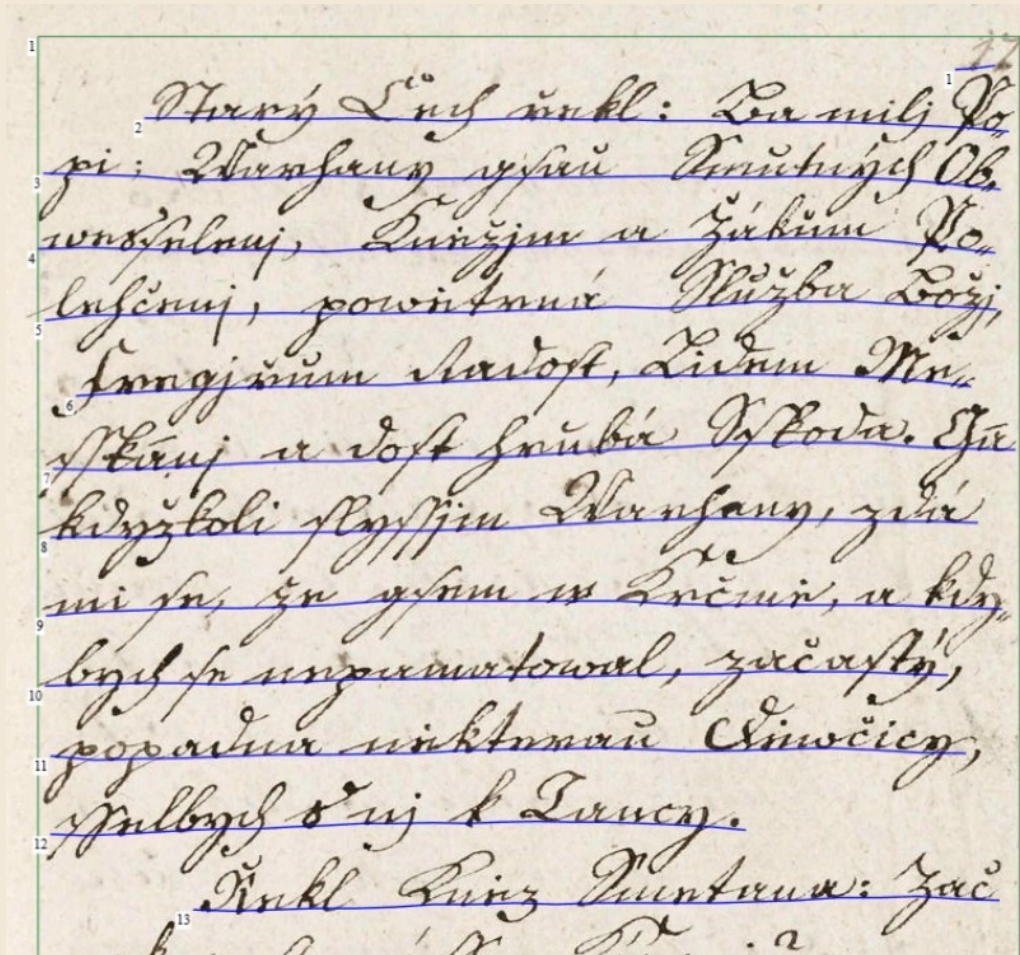
Uloženo v:

Historická expozice 71. mech. praporu
„Sibiřského“ v Hranicích. Zapůjčeno z
pozůstalostní sbírky rodiny.

Digitalizát vytvořen s laskavým svolením
kurátora muzea nrtm. Radima Cába.

Agregovaný model Finale 2.0

Ukázka použití modelu na textu, který nebyl trénován



Region 1

- 1 17
- 2 Stary Čech řekl: Ba milj Po-
- 3 pi: Wárhany gfaú Gmútných Ob-
- 4 webfelegj, Kněžjm a Žákúm Po-
- 5 lehčenj, powětrú Slúžba Božj,
- 6 Přepjřúm Radost, Lidem Me-
- 7 řkánj a doft hrúbá Sfkoda. Gu-
- 8 kdyžloli flyřřjm Wárheny, zdá
- 9 mi fe že gfem w Krčině, a kdy-
- 10 bych fe nepamatowal, začařký
- 11 popadna některou děwčicy
- 12 řelbych s nj k Čaúcy.
- 13 Řekl Kněz búletanx : Zač
- 14 pak geft nářfe Kázágj?
- 15 Stary Čech řekl: Rázánj wá-

Zdroj:

Rozličné písně starožité

Uloženo v:

Moravská zemská
knihovna v Brně
RKP-0048.022; Rkp 59

AUTOMATICKÁ
TRANSKRIPCE
POMOCÍ PLATFORMY
TRANSKRIBUS

VOJTĚCH ŘÍHA



Pozehnana brádrů oklesně,
Křesťů a nerozdilna Trojice
Božka, Duch Otce, Duch Syn
Duch Duch Dvati Ami v wž

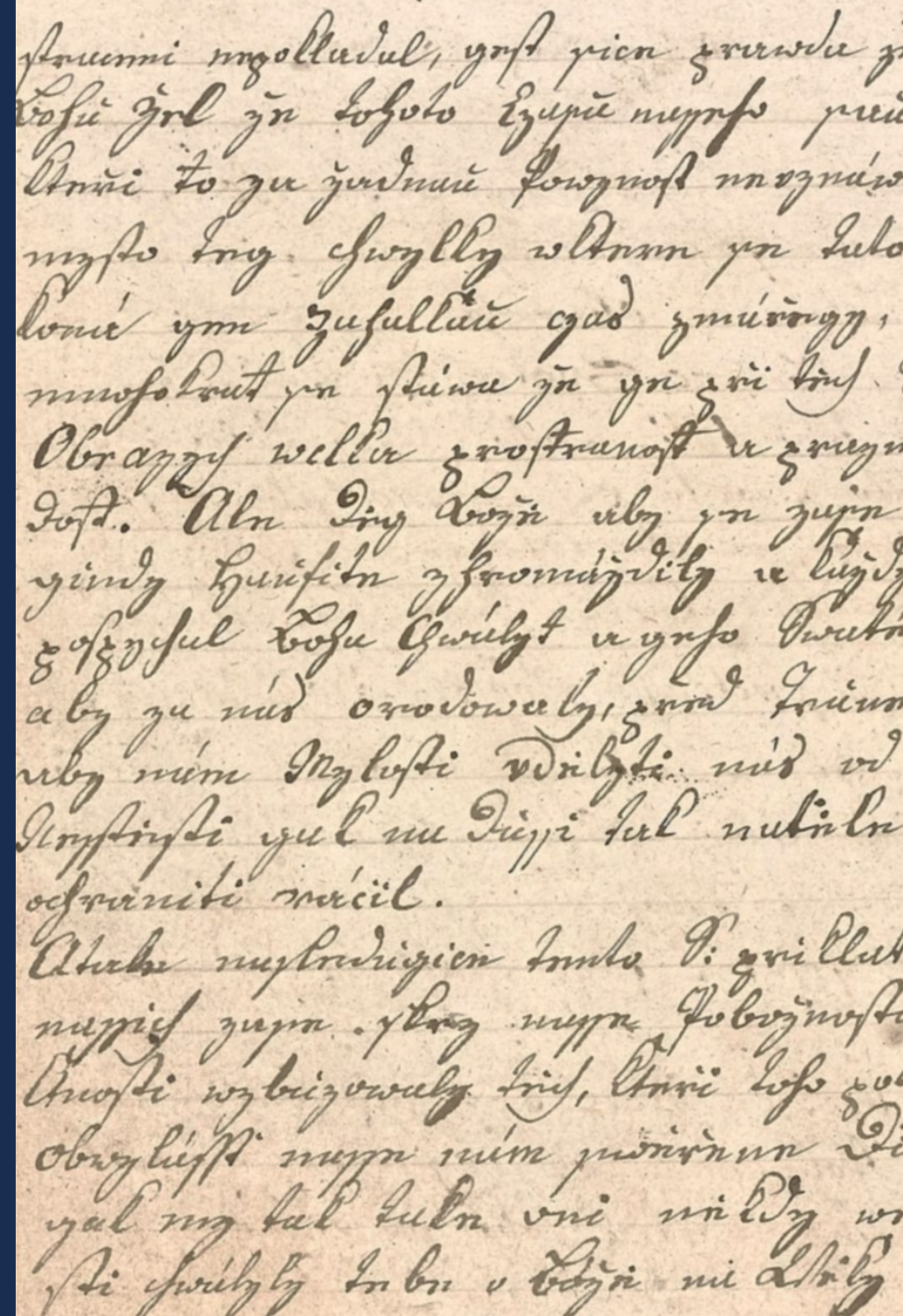
sti, a nešodnosti se nekóřil, a neponizoval.
Ach Bože Troj jediný! tobě patří čest,
sláva a dílo činění. Ale ty Bože! v sa,
me pochvále oblybeni nemáš, Ty od nás
žádáš dětinství větění, ty od nás žádáš
lásku, podanost, a důvěrnost. *fošle Bože*
my ti dnes lyto Avěti měni a me mēdaū

Paní Bohu služba wděčná
A křesťům Radost věčná.
Křesťomů modlitbě, aby ona
tjm spisegi a bezpečnegi do nebe k Bo-
hů se man ozi maha dno s úřad a uřidati
zdravugi tebe mradosny den
narozeny tveho zamiteg knem
oneg světe gšny panenko potěše-
ny a radosti srdce meho tebe

MODEL AGREG-8

Model Agreg-8

- Společné rysy dokumentů
 - Století 18.-19. století
 - Tematika modliteb a katolických písní
 - Podobnost písma – cílem je postupně rozšiřovat data pro dosažení větší rozmanitosti znaků
 - Licence Creative Commons (BY-NC-SA)



Handwritten text in a cursive script, likely a prayer or religious text, written on aged paper. The text is dense and fills the right side of the image. It appears to be a historical document, possibly a prayer book or a collection of religious songs, as mentioned in the adjacent text. The script is a form of cursive, characteristic of the 18th or 19th century. The paper shows signs of age, with some discoloration and wear.

2 . Cesta Svatocellenská

- Datace: 1733-1766
- Instituce: Muzeum Jindřichohradecka
- Signatura: RK 037
- Rozměry : 18 cm x 11,5 cm
- Problematika podobnosti i/y
- Používání obrovských iniciál (segm.)

Krista a Jani Swieta : kteraz
žadneho neopaustiti a zadnim
nepohradat w zlydni o Jano
Dobrotiwa cesima ktilosti
a wotproa nam utweho mileh
dina w ssech hrstehuro odpust
ni, abichom pobožna u ktil
twe Swate Bozeti sobie p
pominali, potom wiecneho
Blahostarwenstwi dosahli hrz
stiedrost Jana nasseho Gezi
Krista czechoz ysi ti Jano kreci
porodila ktemiz v Dohem otcem
y d duchem Swatim ziro yst ak
alucye w trogici dokonali Su

על פניו
הוא גדול
הוא קטן

tinuoyi wylkiffel

We Gmenū

Trögice,

tebestuzze Lita 20

3. Radostná cesta

- Datace: 1829-1884
- Instituce: Moravské zemské muzeum
- Signatura: ST 2272
- Rozměry: 19,5 cm x 16,5 cm
- Různost při psaní diakritiky

la opuštěná, pro bolest do Město
padla, skoro mrtvá syžála
O nejdobrotivější Jámě Mě
Přijete brzo Orvati opuštěni, neopá
mne, ^{in muppi} mne, m m m jarmuteyed, a prot
stwic, ne nej mne padnaoti, m
galau: málo myslu, neb za
odvrat ^{in muppi} od mne nemoc, Inor,
nebezpečné wod rozlyti, studly
Ojic, Laurky, Grombyli, a Lrip
Ano rseck byli nestěti,
O nejdobrotivější, Opuštěná!
Boji! Přijete brzo S: opuštěnj, ne
Přij mne ^{in muppi} romem časem jivo
Orozlastně ale neopáštaj m
Přij loner jivo la mēso, neopa
mne, ^{in muppi} Pj nā smrtelné hostely
Laudis, od celého Orvati opuště
plny bolesti, ležeti buda, neop
mne ^{in muppi} Džj stráčený, gazy m
nejšťetější Jmeno wyee ne

Büch Büch Büch

Pariti utunula Lüdi, nüle

Beni assemohaucy

Putowati zhuari

4 . Modlitby, písně a litanie

- Datace: okolo roku 1826
- Instituce: Moravské zemské muzeum
- Signatura: ST 2193
- Rozměry: 22 cm x 18 cm
- Různé druhy písma v jednom dokumentu (unclear tag)

twym, we wyznani prawe
stawu Trogyce wieczne po
a w mocy Wlechnosti, Se
se klaneti, prosyme tebe, a
wyry stalosti, proti wssim
wenswym opatreni byly, st
Bezisse Xrysta pana naszego
s teba u ziw gest, a bral u
gednoti Duchu swateho r
Zuhy na wely weli kmo

3. Otinnym u 33 Truwat Mar
Pliwu Otij z Pym z Dufu
z: Galozj bylu na gwelku nji
wzlicy wj na Libly Wnlic

Sidzj na wile Kuznort l' P.
Janina: drcim w Dofu:
Pliwu Otij st: zel Otinnym. D
zo Trilrut z Truwat Merga:

5 . Modlitební knížka

- Datace: 1700-1750
- Instituce: Muzeum Jindřichohradecka
- Signatura: RK 087
- Rozměry: 22 cm x 16 cm
- Průsvitnost textu protilehlých stran

Altari posvěcený tvaú štoc
Božstau pěstovaný, a k gste
naděgi Spasený našeho přit
men byl, deq snilost abychom
na tebe, kterého zde pod Spůs
baú Chleba vidjme, w Nebes
wěčně, patřili, a tam se s teba
wěčně, radowali, Amen.

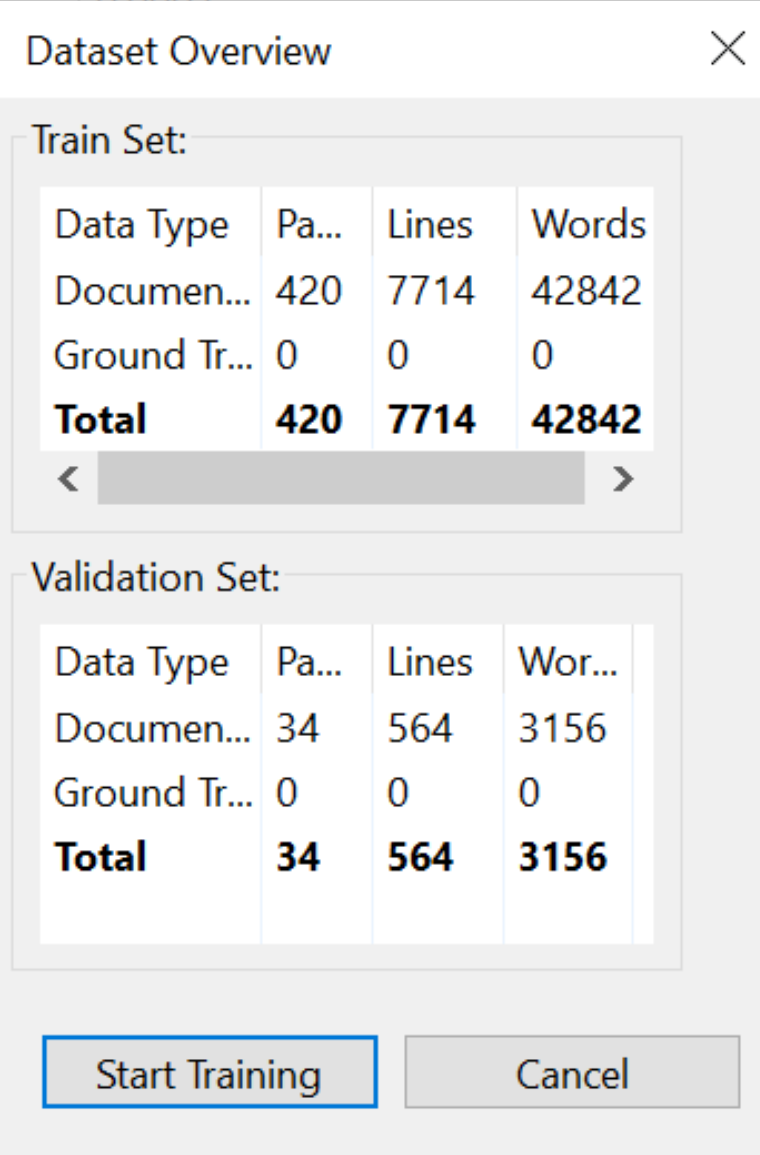
Kněz Kalich pozdvihügi
Z Jan Krysta Pana nřmřtř křen
teče, a gste

Modlitba
Dřesvatá Krav, která
gste w Nřnřsvětęgšřch Křan
Spasitele meho k obmřtř wřřech
mřch hřřchřw hognę, tekla, ob
meg wřřectnř Nřcřřřotř hřřchř
mřch, a mnę očřřřřenř, k wěčnř
mř životř zachowęg. Dřpanę

Geřřřř Kryste! genř gřřř chřřř

Model Agreg-8

- Počet trénovacích cyklů: 250
- Délka trénování: 21h 37m
- Celkový počet slov: 45998
 - *Trénovací sada*: 42842
 - *Validační sada*: 3156
- Počet stran GT: 454
- Výška řádku: 140 px



Dataset Overview

Train Set:

Data Type	Pa...	Lines	Words
Documen...	420	7714	42842
Ground Tr...	0	0	0
Total	420	7714	42842

Validation Set:

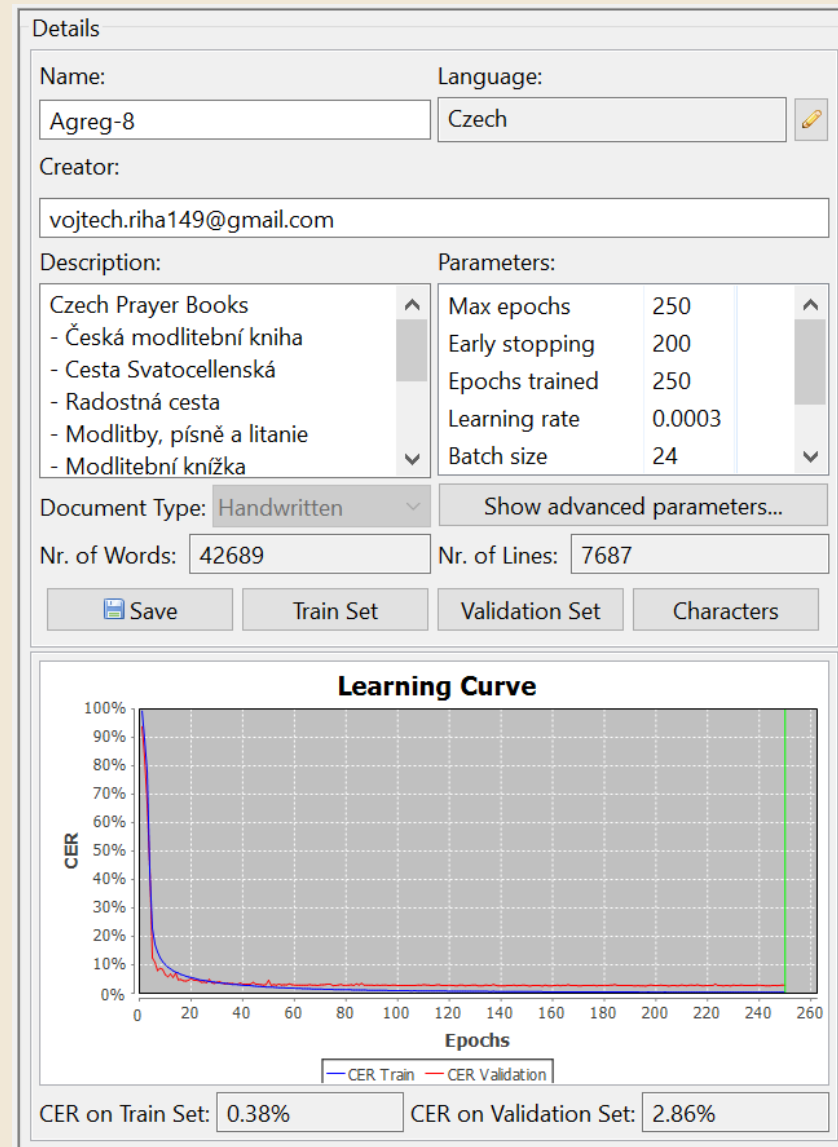
Data Type	Pa...	Lines	Wor...
Documen...	34	564	3156
Ground Tr...	0	0	0
Total	34	564	3156

Start Training Cancel

Model Agreg-8

■ Výsledek

- CER na trénovací sadě: **0,38%**
- CER na validační sadě: **2,86%**
- nejlepší epocha CER: **2,60%**
- nejlepší epocha WER: **16,57%**



Model Agreg-8

Model 'Agreg-8'

Name: Agreg-8 Language: Czech

Creator: vojtech.riha149@gmail.com

Description: Czech Prayer Books
- Česká modlitební kniha
- Cesta Svatocelestenská
- Radostná cesta
- Modlitby, písně a litanie
- Modlitební knížka
Doplnění verze 7 (20 stran 5. dokumentu)
Korekce

Parameters:

Max epochs	250
Early stopping	200
Epochs trained	250
Learning rate	0.0003
Batch size	24
Normalized height	140
Omitted Tags	unclear

Show advanced parameters...

Document Type: Handwritten

Nr. of Words: 42689 Nr. of Lines: 7687

Save Train Set Validation Set Characters

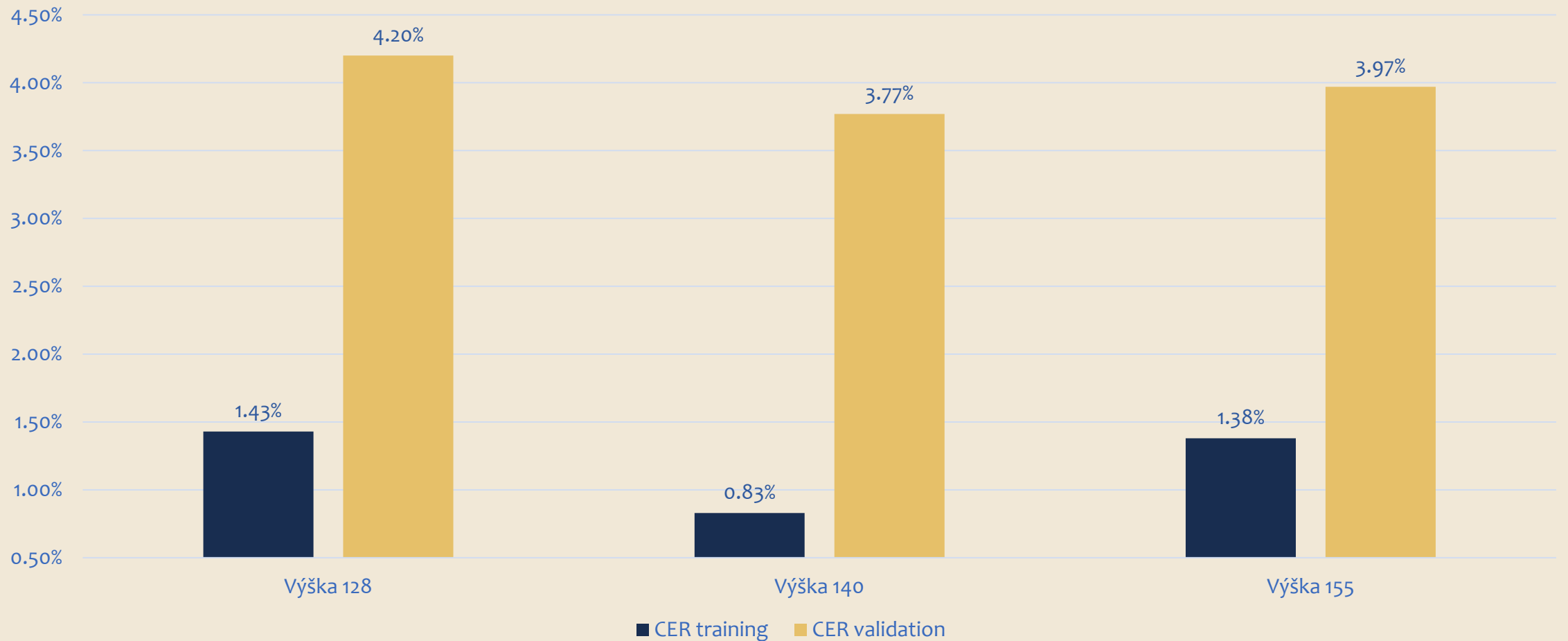
Learning Curve

Epochs	CER Train	CER Validation
0	100%	100%
5	~10%	~15%
10	~5%	~8%
20	~3%	~5%
50	~2%	~3%
100	~1.5%	~2.5%
250	0.38%	2.86%

CER on Train Set: 0.38% CER on Validation Set: 2.86%

Korekce výšky řádku u modelu ČMK-70

Úspěšnost dílčích modelů

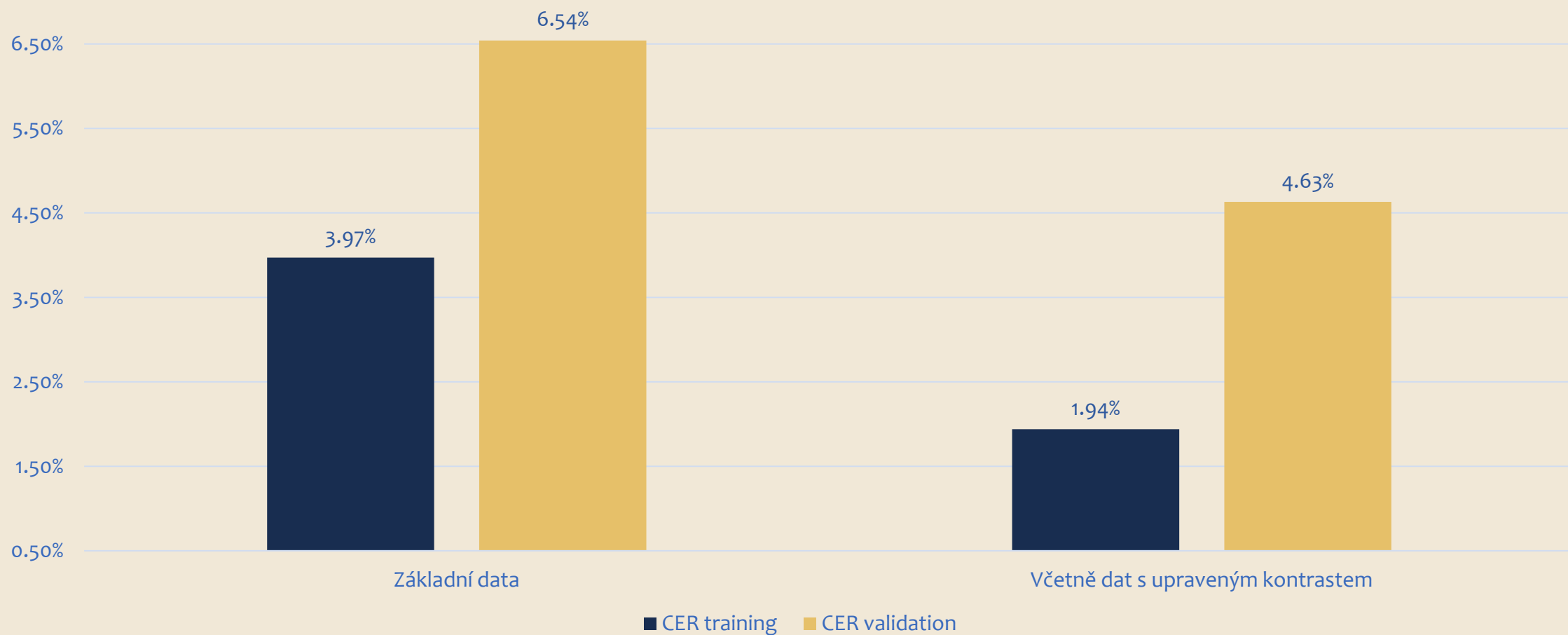


Augmentace obrazových dat

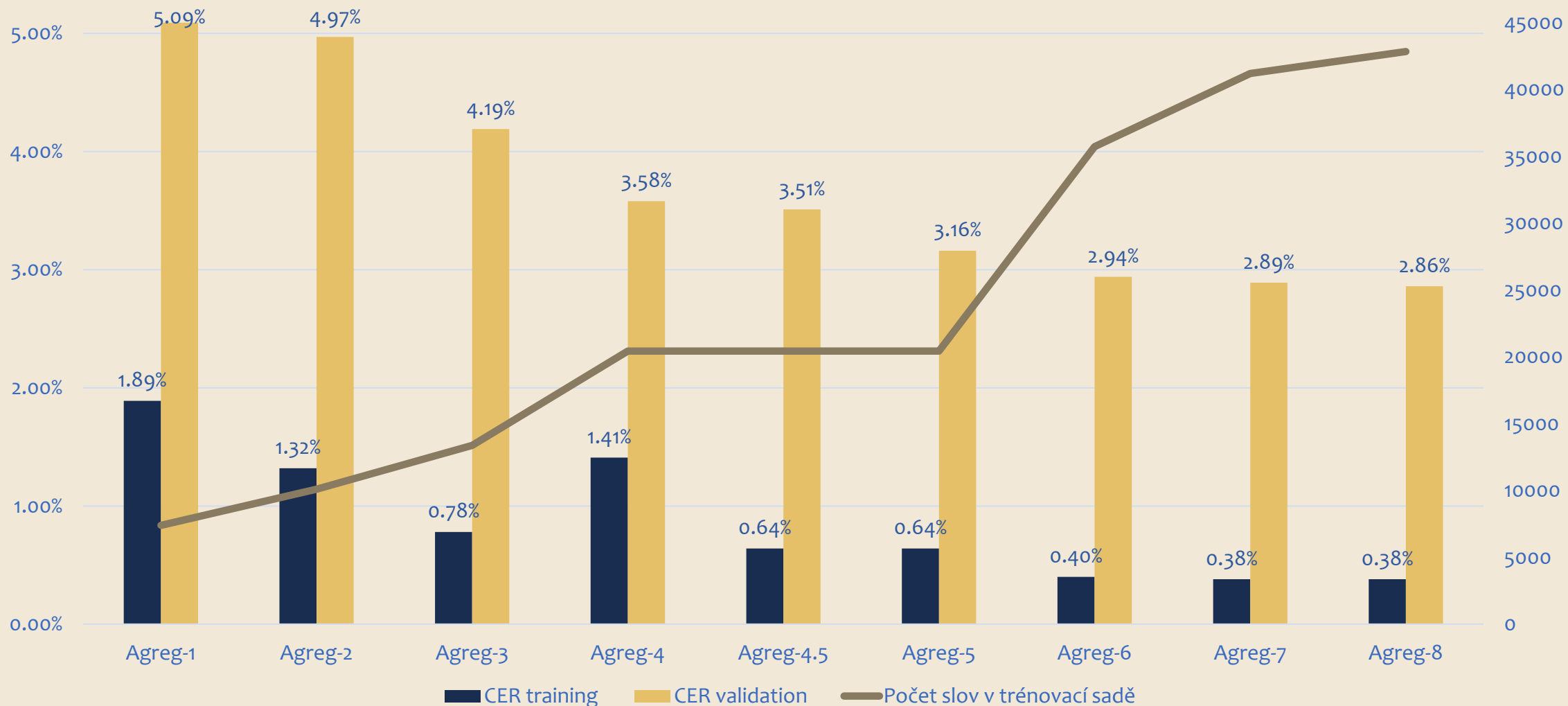
1. Geometrické transformace
 - *Rotace, škálování*
2. Fotometrické úpravy
 - **Změna kontrastu (+40, -20), jasu (+10, -10)**
 - *ostrosti, šumu apod.*
3. Deformace
 - **Změna šířky či výšky obrazových dat (+- 10, 20%)**
 - *Práce s oříznutím obrázku*
 - *Přidávání stínů a rozmazávání*

Augmentace obrazových dat modelu u ČMK-20

Úspěšnost dílčích modelů



Vývoj úspěšnosti a počtu slov modelu AGREG



prof. PhDr. Dušan Katuščák, PhD.
dusankatuscak@gmail.com

Lukáš Němec
luki.nemec@seznam.com

Vojtěch Říha
vojtech.riha149@gmail.com



**SLEZSKÁ
UNIVERZITA**
FILOZOFICKO-
PŘÍRODOVĚDECKÁ
FAKULTA V OPAVĚ