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*.

Cieľ 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

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 diferenciace
- různé rukopisné styly a autoři
- regionální sounáležitost

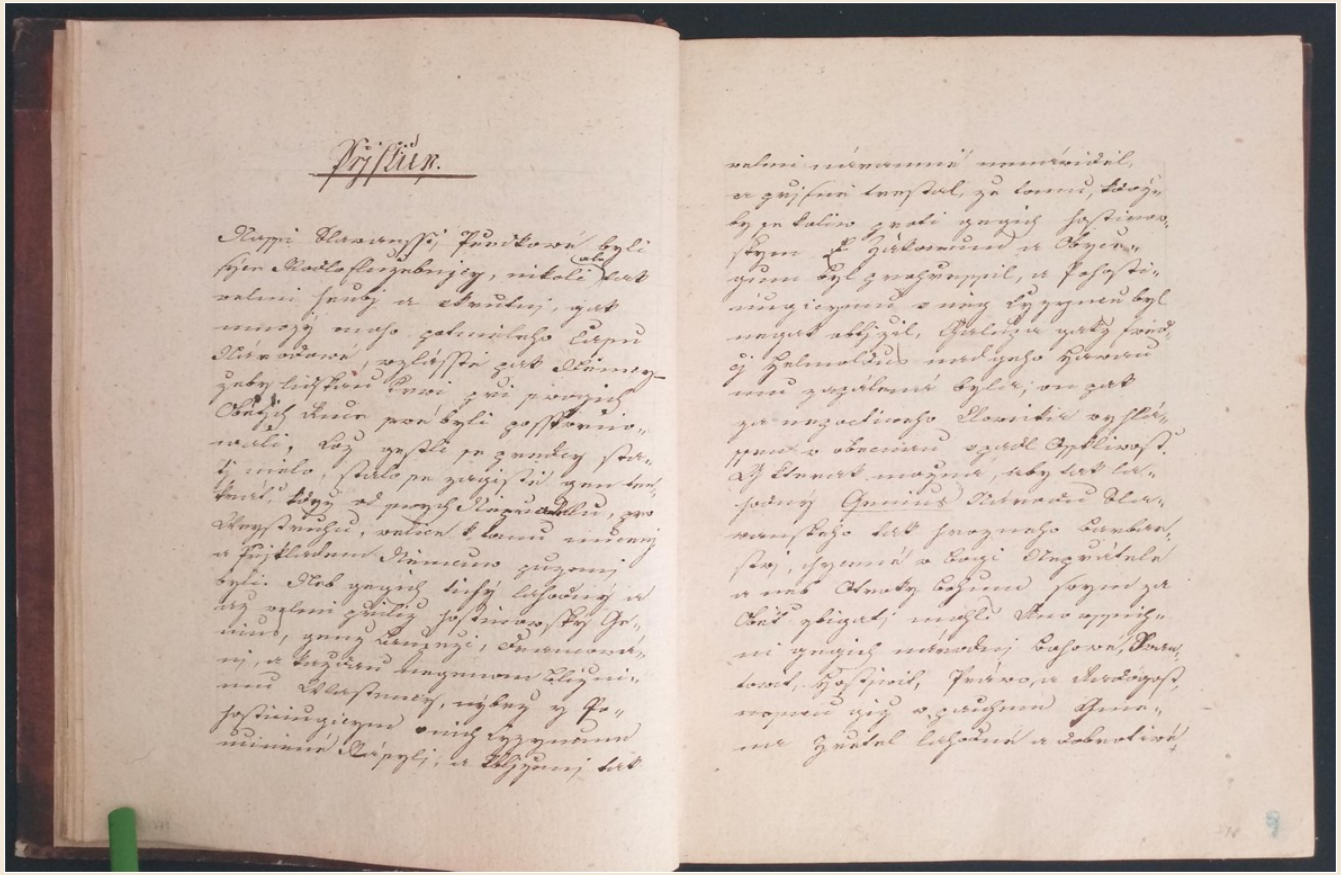
Tvorba modelu automatického rozpoznávání bohemikálního rukopisného textu v platformě Transkribus

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



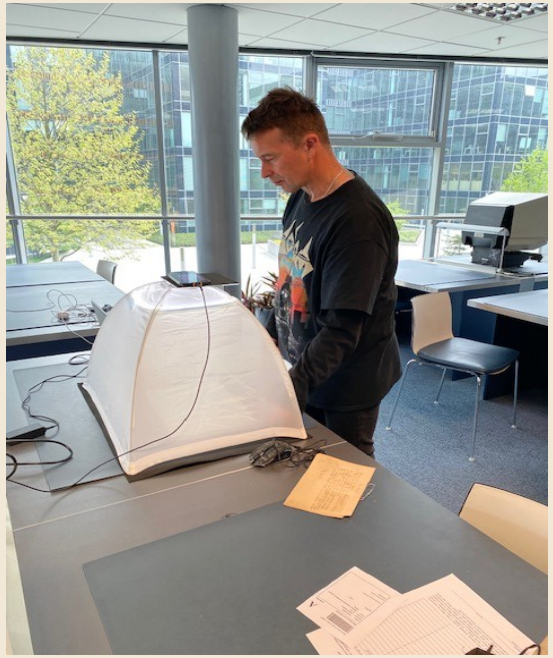
Výzvy:

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

Východisko:

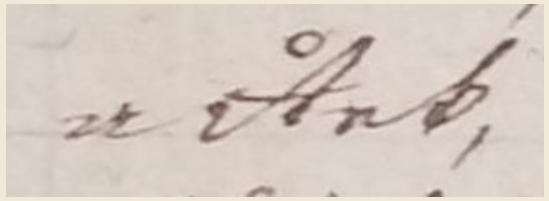
- Dílčí model s chybovostí 8,31 %
- Agregovaný model s chybovostí 6,55 %, který odstranil řadu nejednoznačností.

Mýtické povídky o bozích a bohyních moravských Slovanů

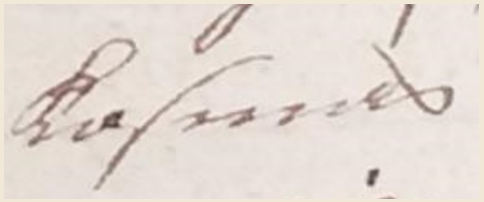



Zachraňujeme historické kulturní dědictví

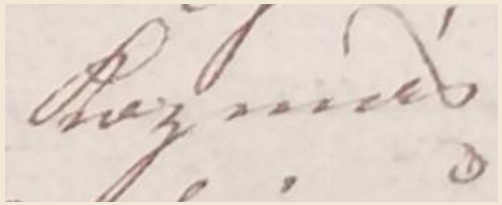
Mýtické povídky o bozích a bohyních moravských Slovanů



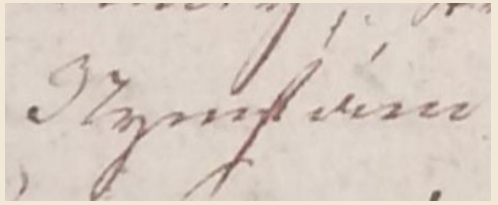
a Řek,



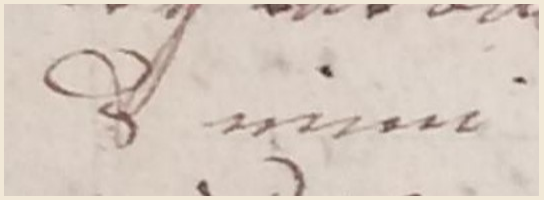
Kosmás



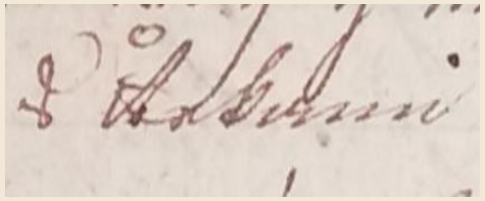
Kozmás



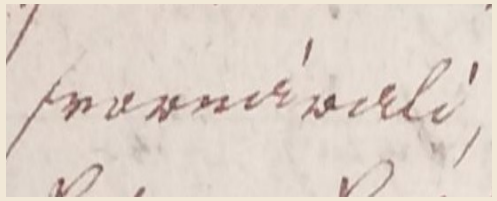
Nymfám



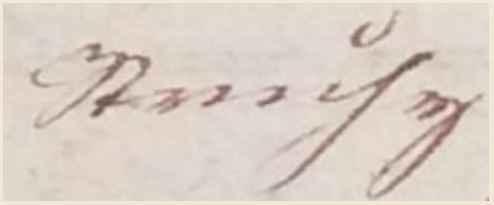
s nimi



s Řekami

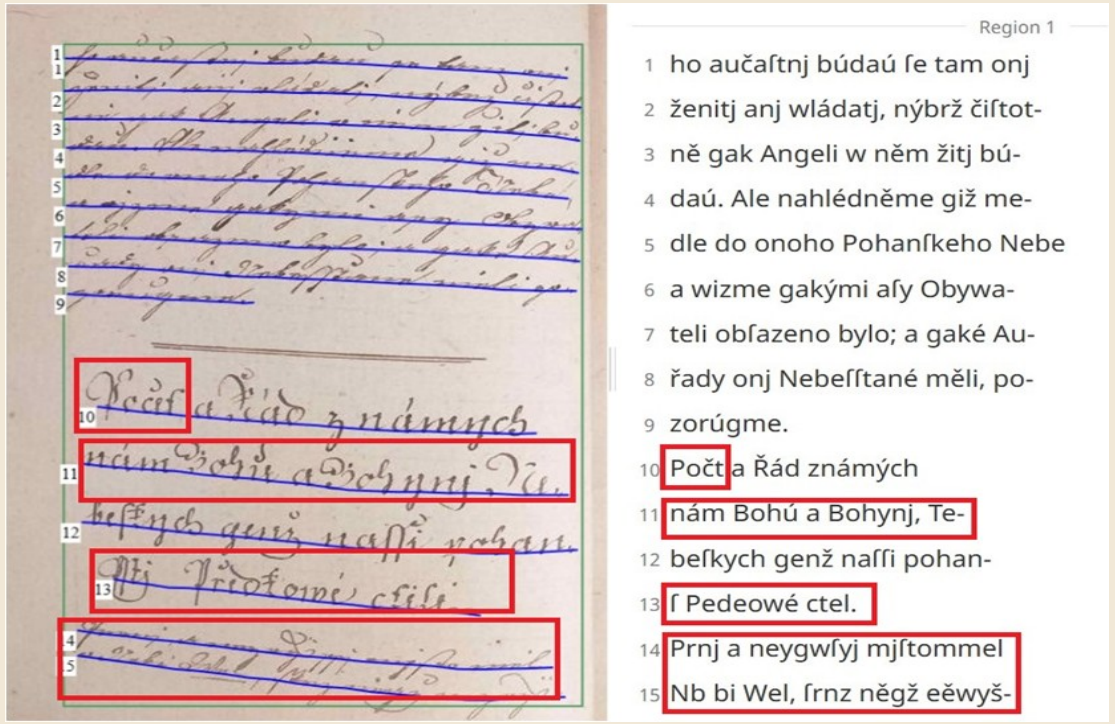


srownáwali ,



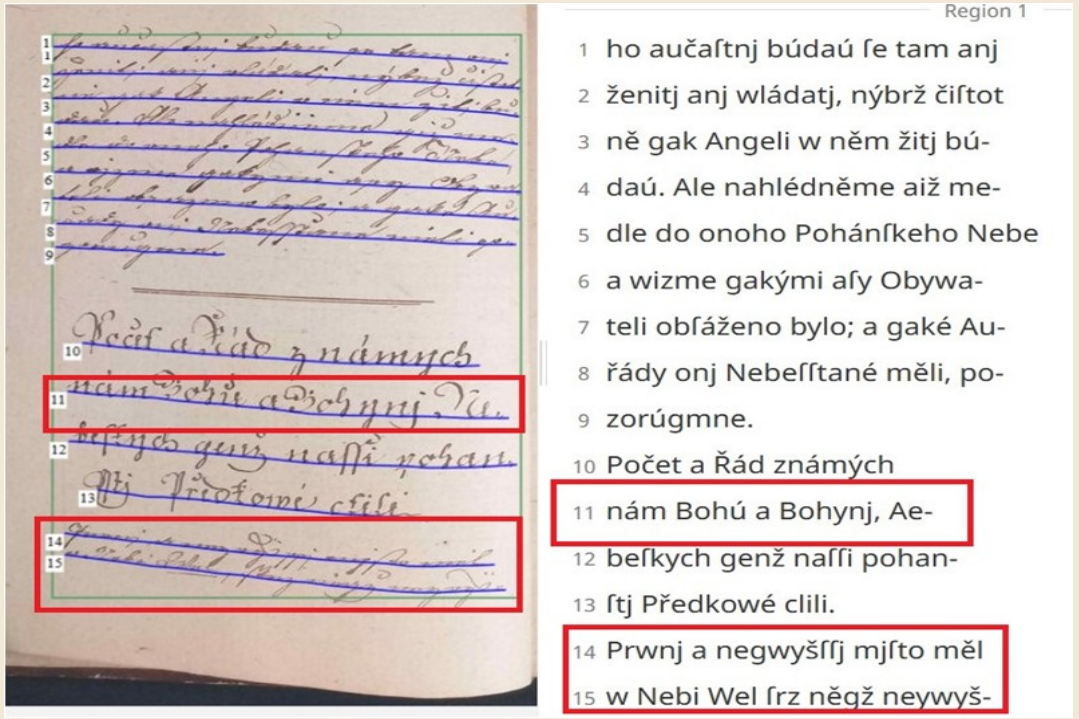
Strúhy

Mýtické povídky o bozích a bohyních moravských Slovanů



Region 1

ho aučaſtnj búdaú ſe tam onj
ženitj anj wládatj, nýbrž čiſtot-
ně gak Angeli w něm žitj bú-
daú. Ale nahlédněme giž me-
dle do onoho Pohanſkeho Nebe
a wizme gakými aſy Obywa-
teli obſazeno bylo; a gaké Au-
řady onj Nebeſſtané měli, po-
zorúgme.
Počt a Řád známých
nám Bohú a Bohynj, Te-
beſkych genž naſſi pohan-
ſ Pedeowé ctel.
Prnj a neygwſyj mjſtommel
Nb bi Wel, ſrnz nëgž eëwyš-



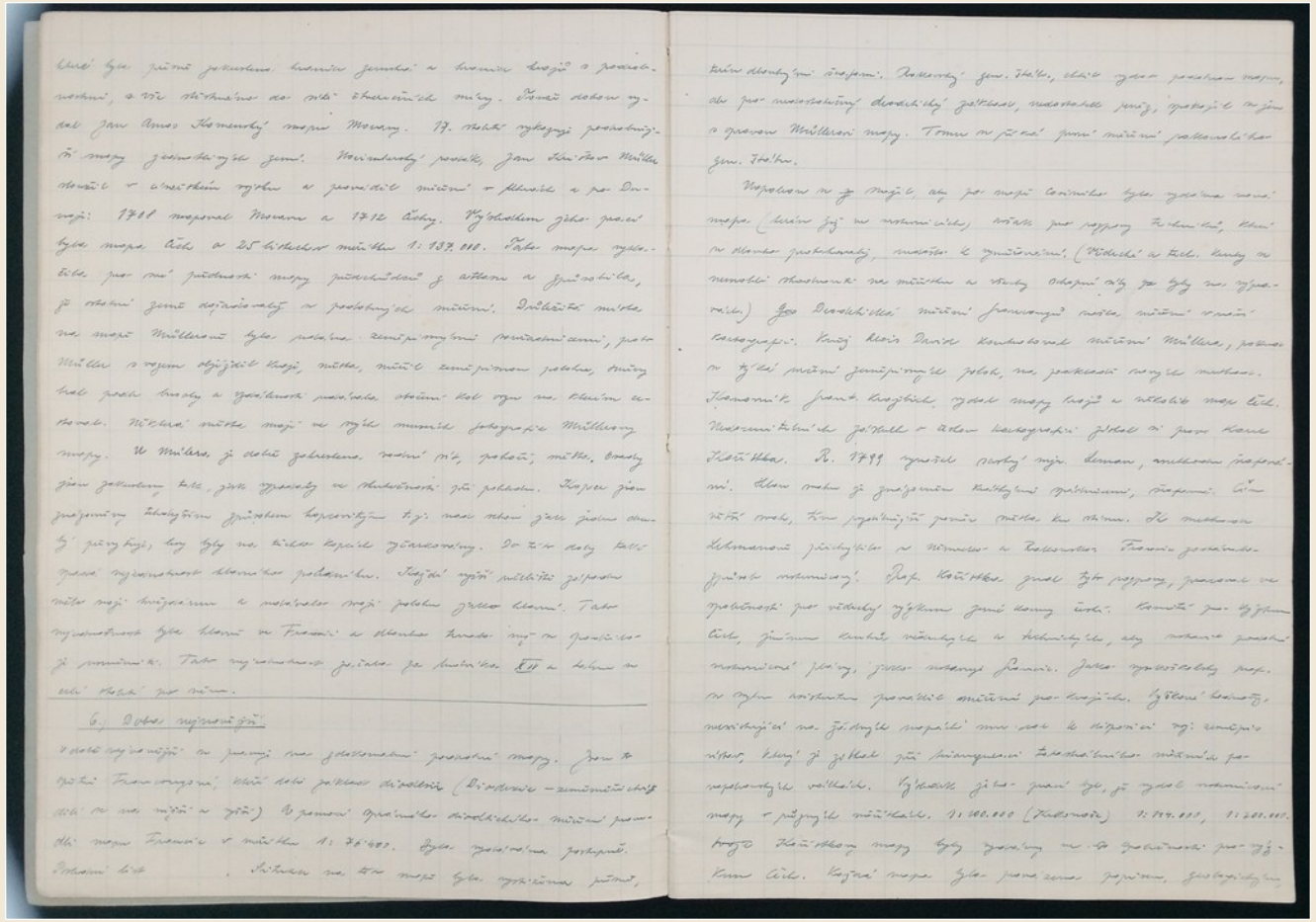
Region 1

ho aučaſtnj búdaú ſe tam anj
ženitj anj wládatj, nýbrž čiſtot
ně gak Angeli w něm žitj bú-
daú. Ale nahlédněme aiž me-
dle do onoho Pohánſkeho Nebe
a wizme gakými aſy Obywa-
teli obſáženo bylo; a gaké Au-
řády onj Nebeſſtané měli, po-
zorúgmne.
Počet a Řád známých
nám Bohú a Bohynj, Ae-
beſkych genž naſſi pohan-
ſtj Předkowé clili.
Prwnj a negwyšſſj mjſto měl
w Nebi Wel ſrz nëgž neywyš-

Model Mystic Absolut s chybovostí 8,31 %

Agregovaný model s chybovostí 6,55 %

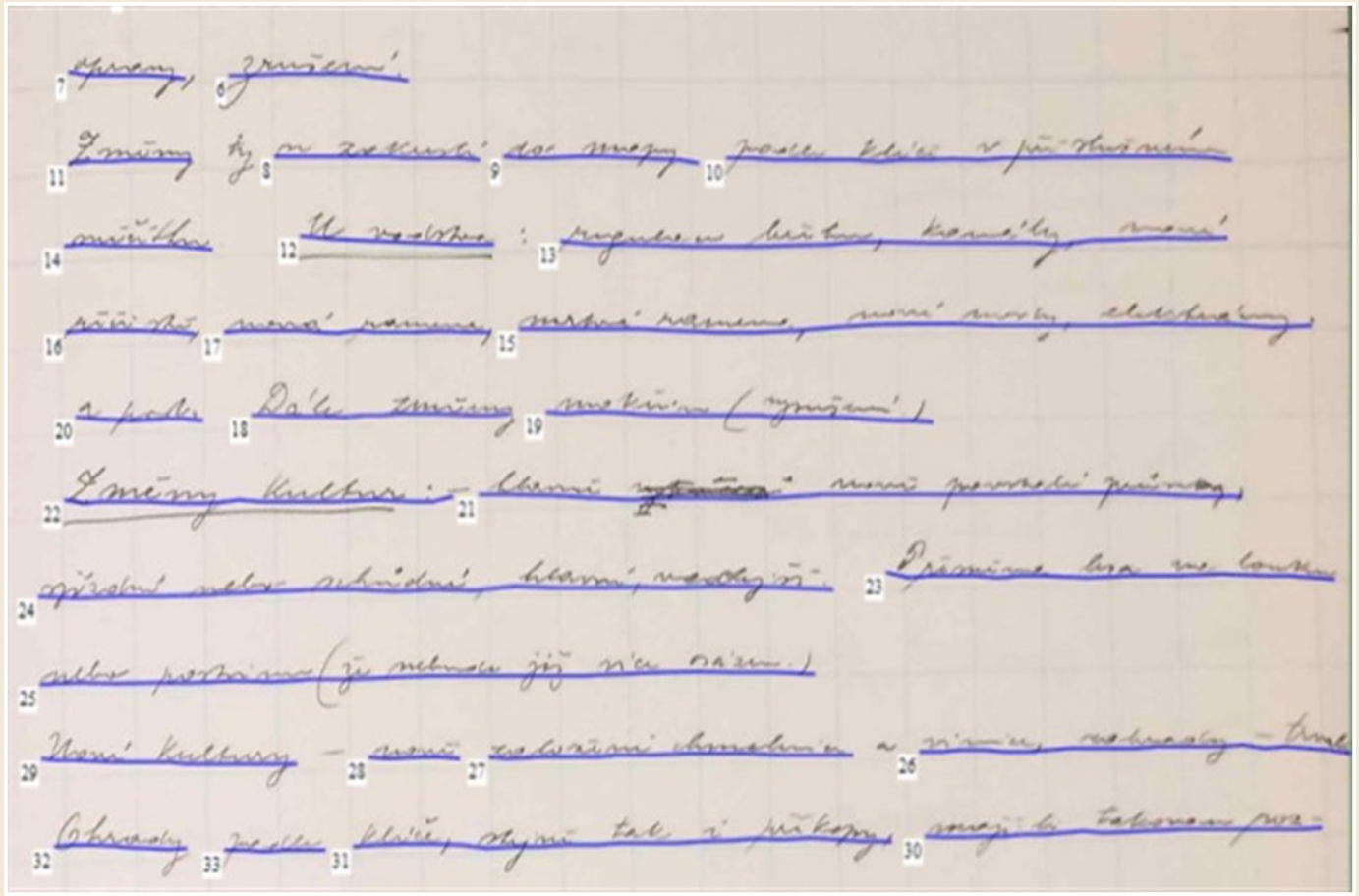
Otakar Jaroš: Nauka o terénu (školní sešit)



Výzvy:

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

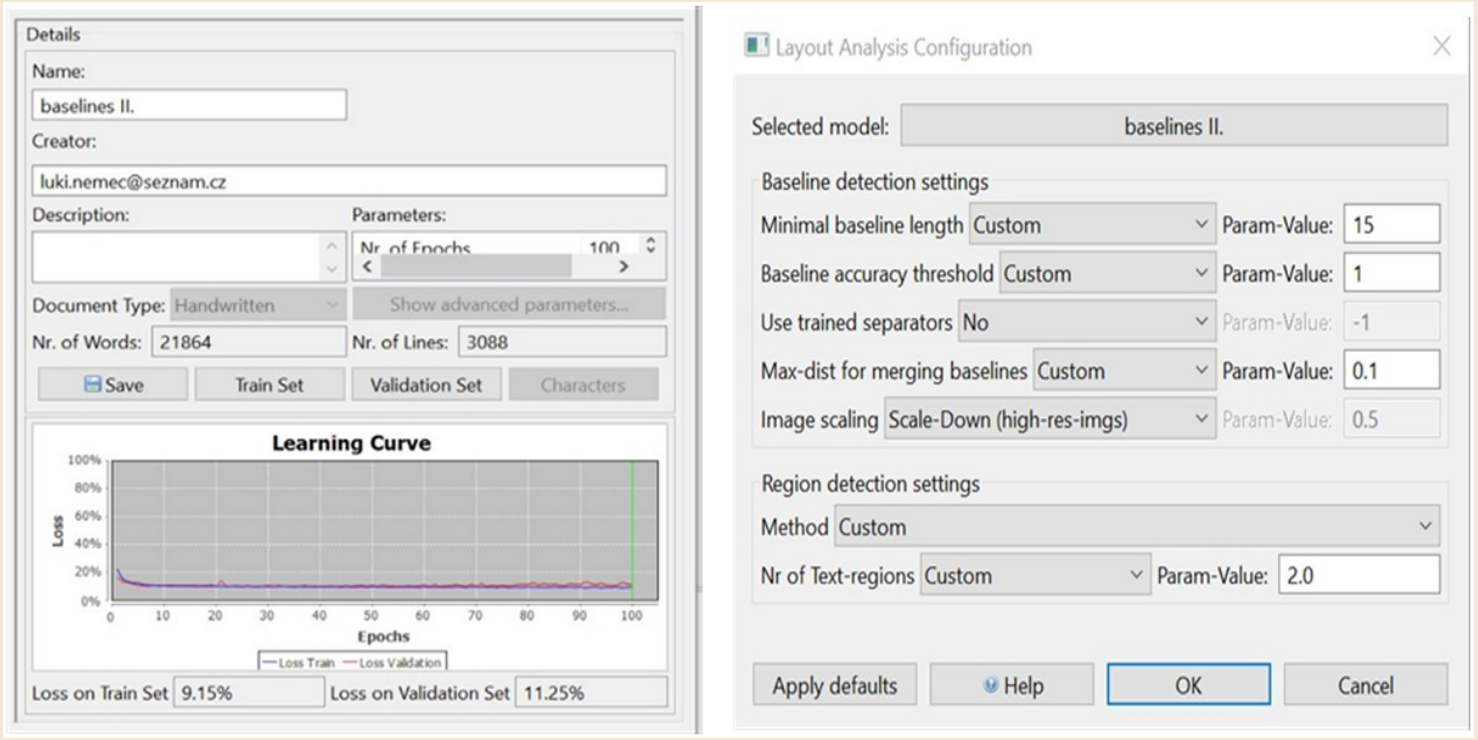
Otakar Jaroš: Nauka o terénu (školní sešit)



Výzvy:

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

Otakar Jaroš: Nauka o terénu (školní sešit)



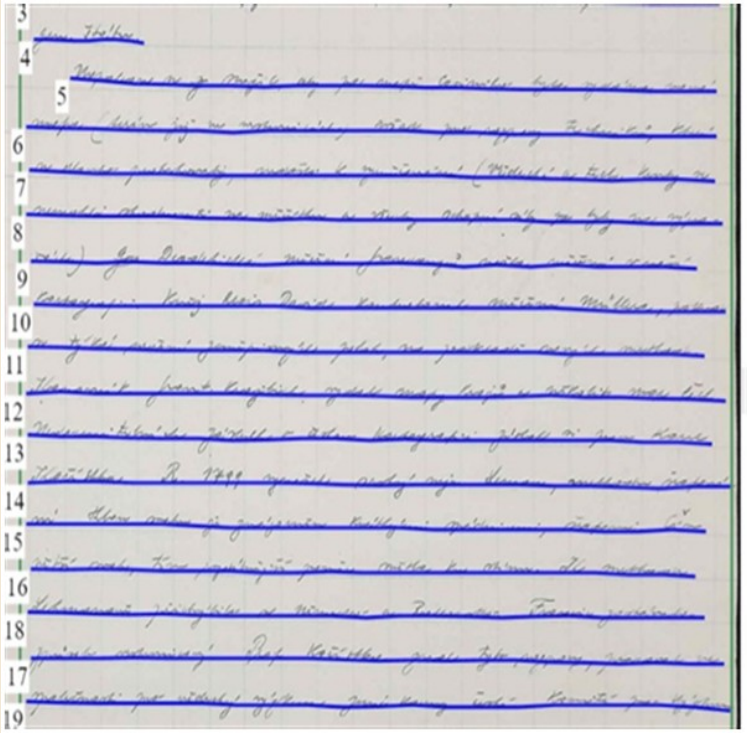
Východisko:

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

Otakar Jaroš: Nauka o terénu (školní sešit)

Odborné vědecké soustředění v Jazernici…víkendový pracovně oddechový pobyt účastníků projektu.

Otakar Jaroš: Nauka o terénu (školní sešit)

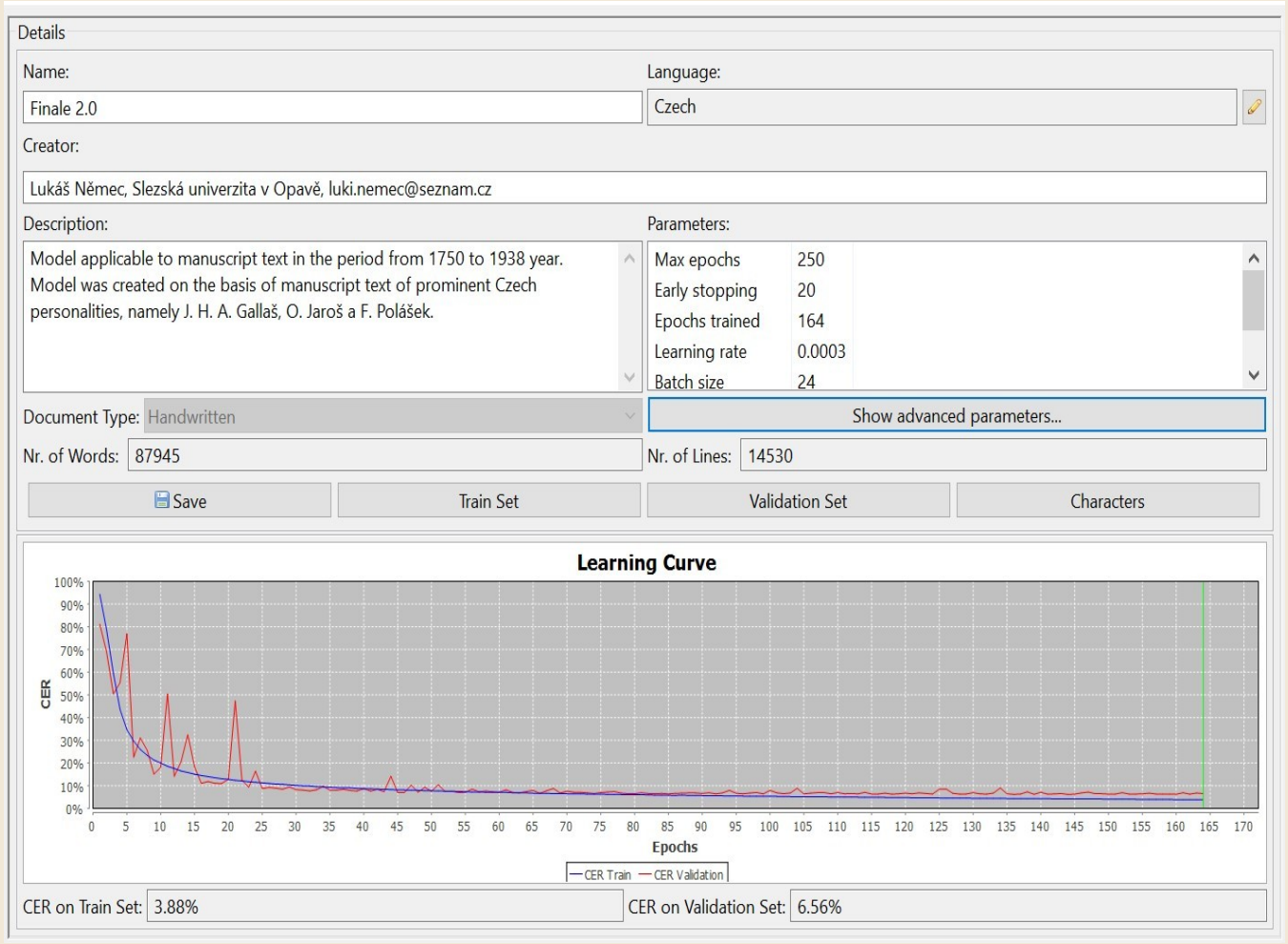


4 gen. štábu.
5 Napoleon se je snažil, aby po mapě lasiného byla vydána nová
6 mapa (terén již ve vrstevnicích, avšak pro rozpory techniků, které
7 se dlouho protahovaly, nedošlo k vyměřování (Vědecké a tech. kruhy: se
8 nemohli shodnouti na měřítku a všechy schopné síly pe byly na výpra-
9 vách.) Jes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

- Ukázka experimentálního ověření naši 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



Doba trénování: 11 hod. 26 min.

Počet epoch: 164 (bylo nastaveno 250)

Počet stran v kvalitě GT: 514

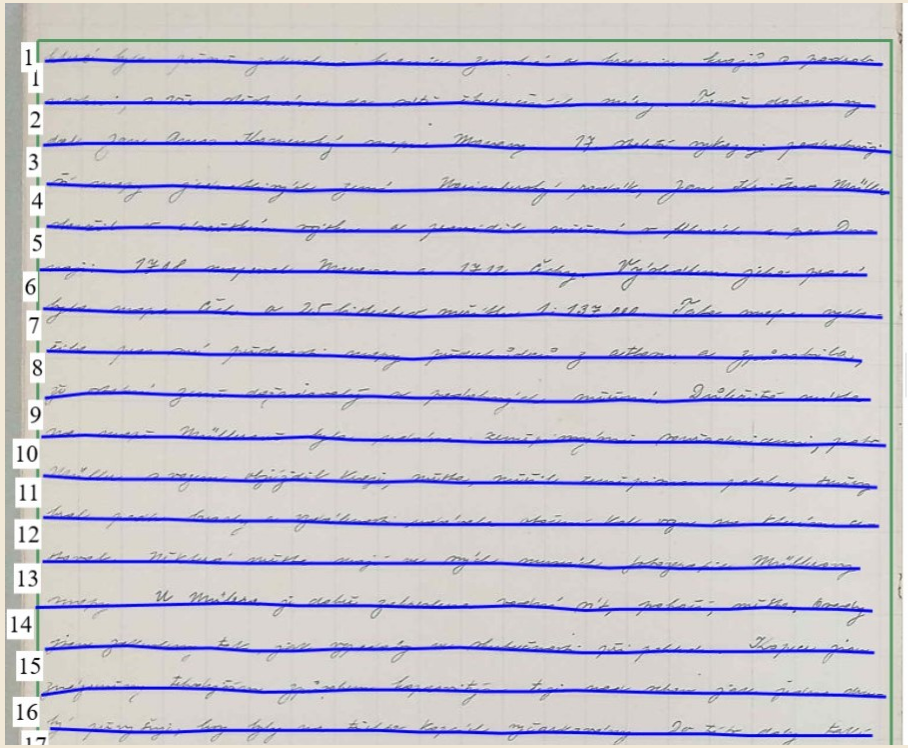
Počet slov trénovací sady/ ověřovací sady: 87.945 / 10.808

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

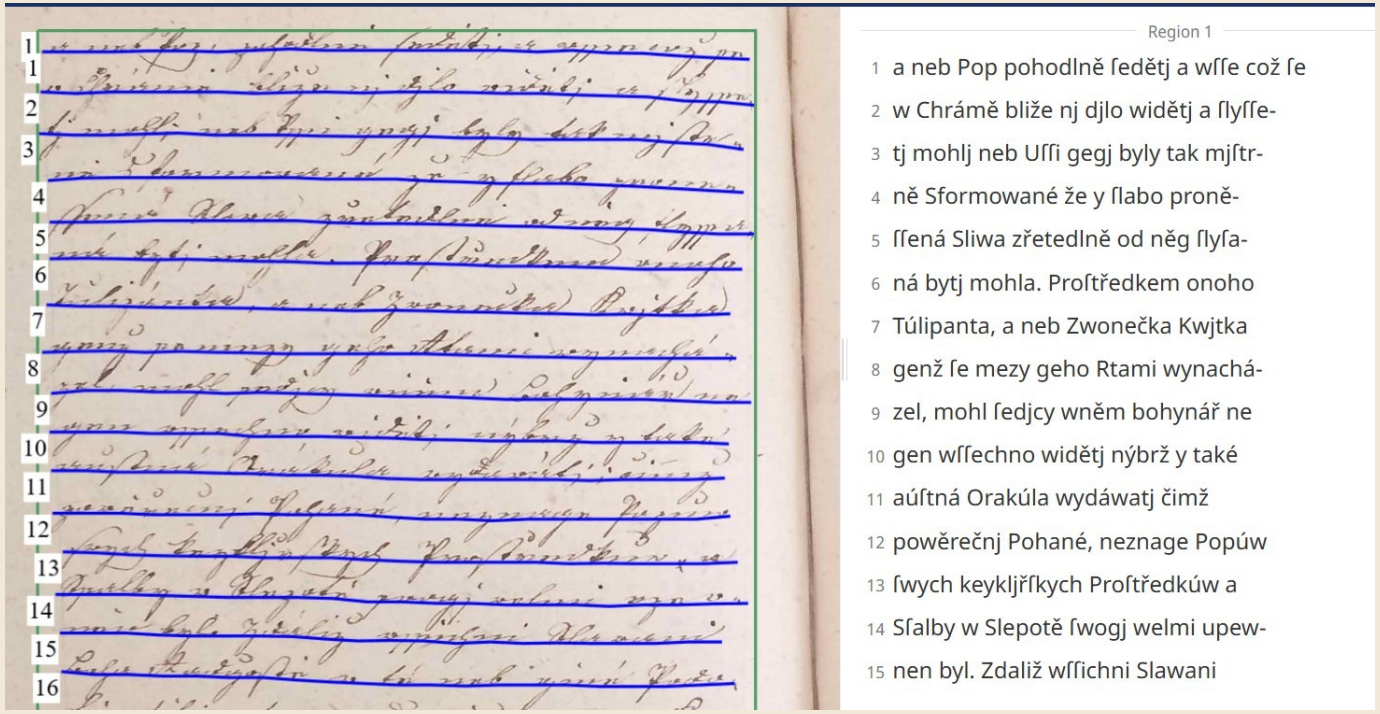


Region 1

1 které byla přesně zakreslena hranice zemská a hranice krajů s podrob-
2 nostmi, s vše stěstnáno 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 Krištov Müllur
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 dožadovalj se podobných měření. Důležitá místa
10 na mapě Müllerově byla udána zeměpisnými souřadnicemi, proto
11 Můller 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 sít, 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

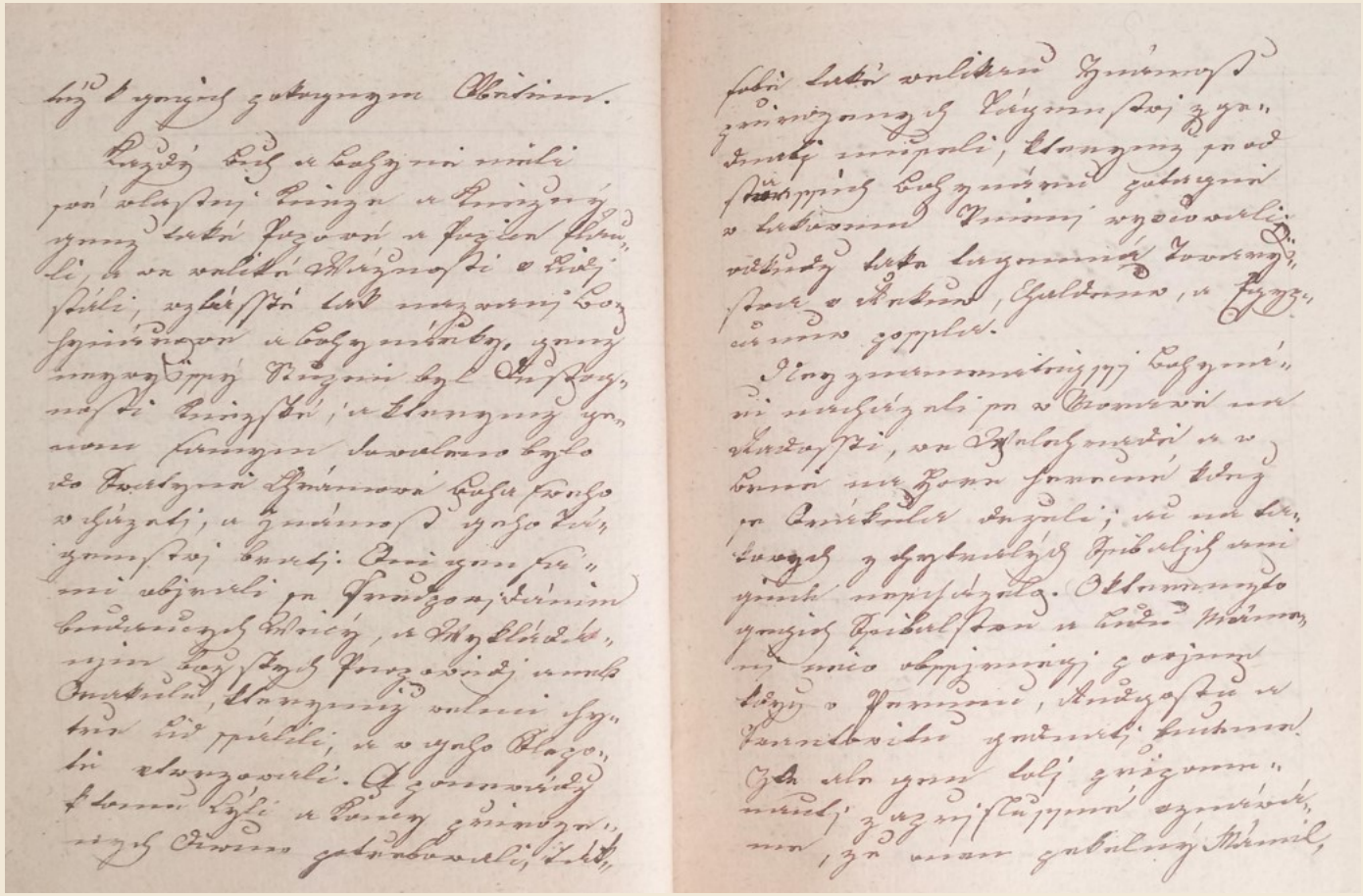


Region 1

a neb Pop pohodlně ſedětj a wſſe což ſe
w Chrámě bliže nj djlo widětj a ſlyſſe-
tj mohlj neb Uſſi gegj byly tak mjſtr-
ně Sformowané že y ſlabo proně-
ſſená Sliwa zřetedlně od něg ſlyſa-
ná bytj mohla. Proſtředkem onoho
Túlipanta, a neb Zwonečka Kwjtka
genž ſe mezy geho Rtami wynachá-
zel, mohl ſedjcy wněm bohynář ne
gen wſſechno widětj nýbrž y také
aúſtná Orakúla wydáwatj čimž
powěrečnj Pohané, neznage Popúw
ſwych keykljřſkych Proſtředkúw a
Sſalby w Slepotě ſwogj welmi upew-
nen byl. Zdaliž wſſichni Slawani

Agregovaný model Finale 2.0

Ukázky rukopisných vzorků:



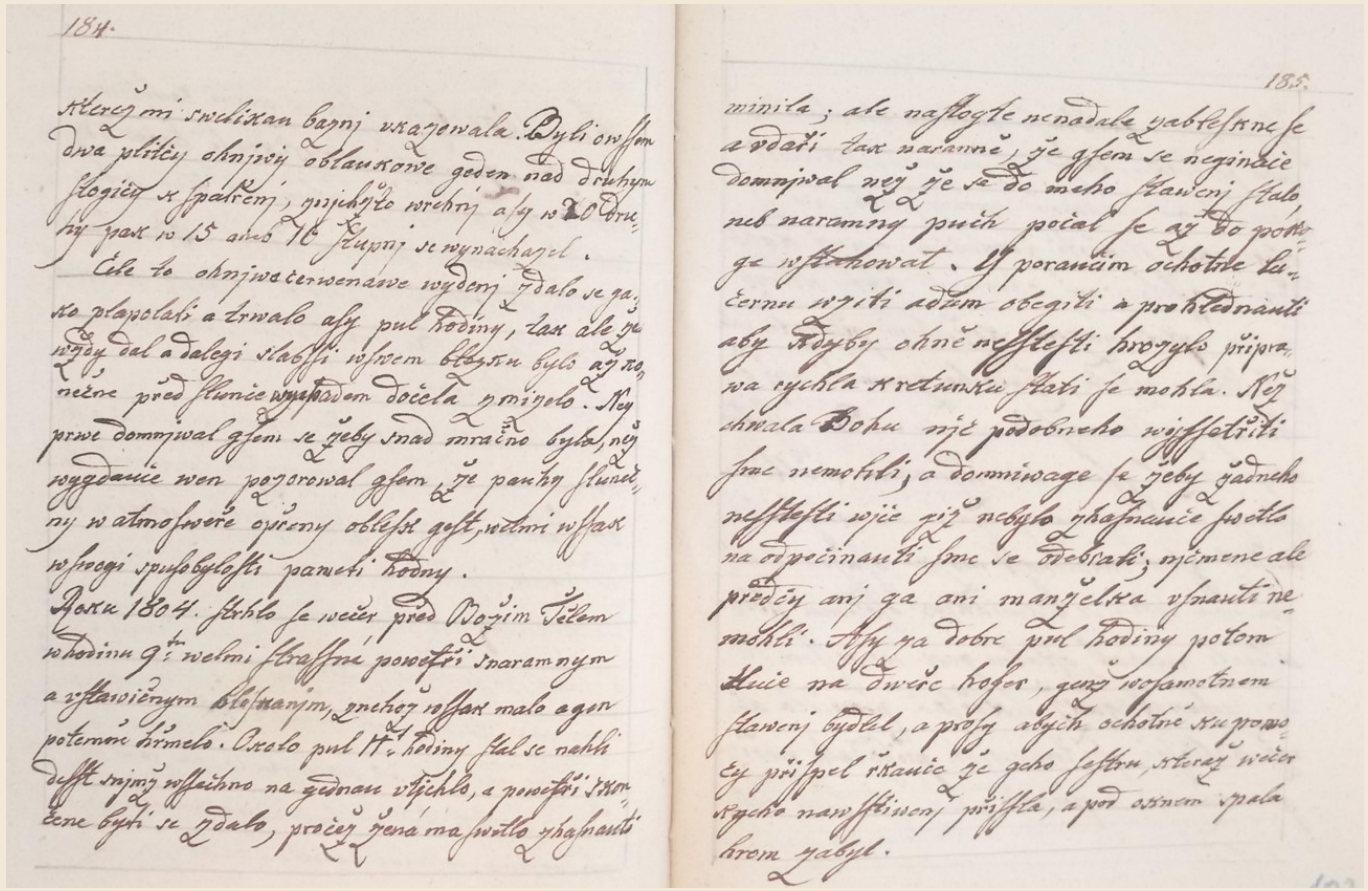
Josef Heřman Agapit Gallaš
Mythické povídky o bozích a bohyních moravských Slovanů

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:



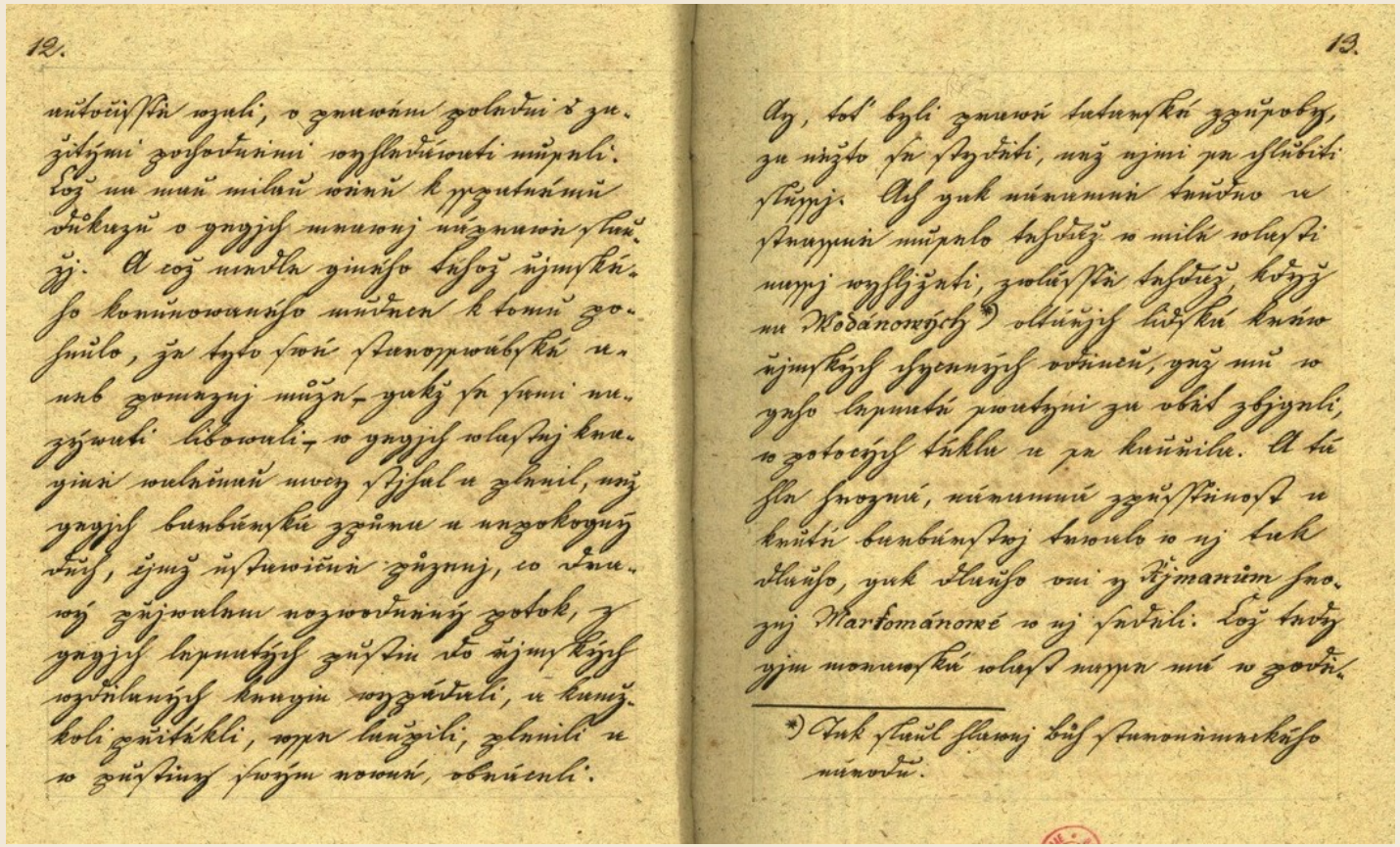
Josef Heřman Agapit Gallaš
Fyzické památky města Hranice a okolí.

Uloženo v:

MZA Brno, G 11 Sbírka rukopisů Františkova muzea Brno, sign. 658, čeština, latina, papír, rukopisná kniha, originál, vázáno v tvrdých deskách, šířka 215 mm, výška 270 mm, pův. pag. 236, nová fol. 128; stará sign.: Schr. 223, pův. 288, červ.

Agregovaný model Finale 2.0

Ukázky rukopisných vzorků:



Josef Heřman Agapit Gallaš
Walaši v kraji Přerovském

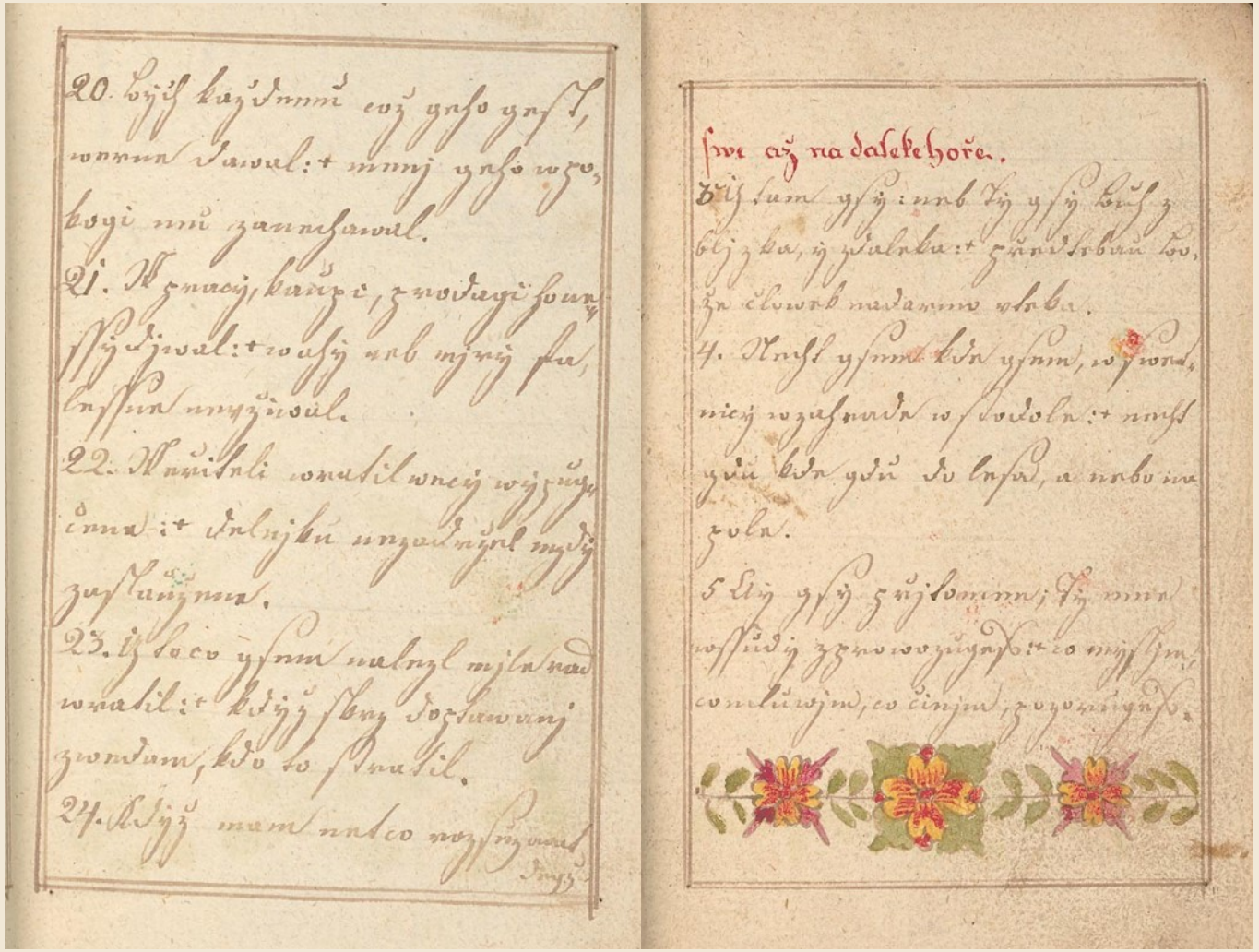
Uloženo v:

Knihovna Národního muzea, Praha, sign. II F 12, čeština, datace 1801-1840, codex, 228 stran, 12 obrazových příloh.; šířka 170 mm, výška 215 mm, digitalizovaný dokument, dostupný z:

https://www.manuscriptorium.com/apps/index.php?direct=record&pid=KNM-NMPII_F_12_1PU4RI1-cs

Agregovaný model Finale 2.0

Ukázky rukopisných vzorků:



František Polášek
Pravé poznání boha...

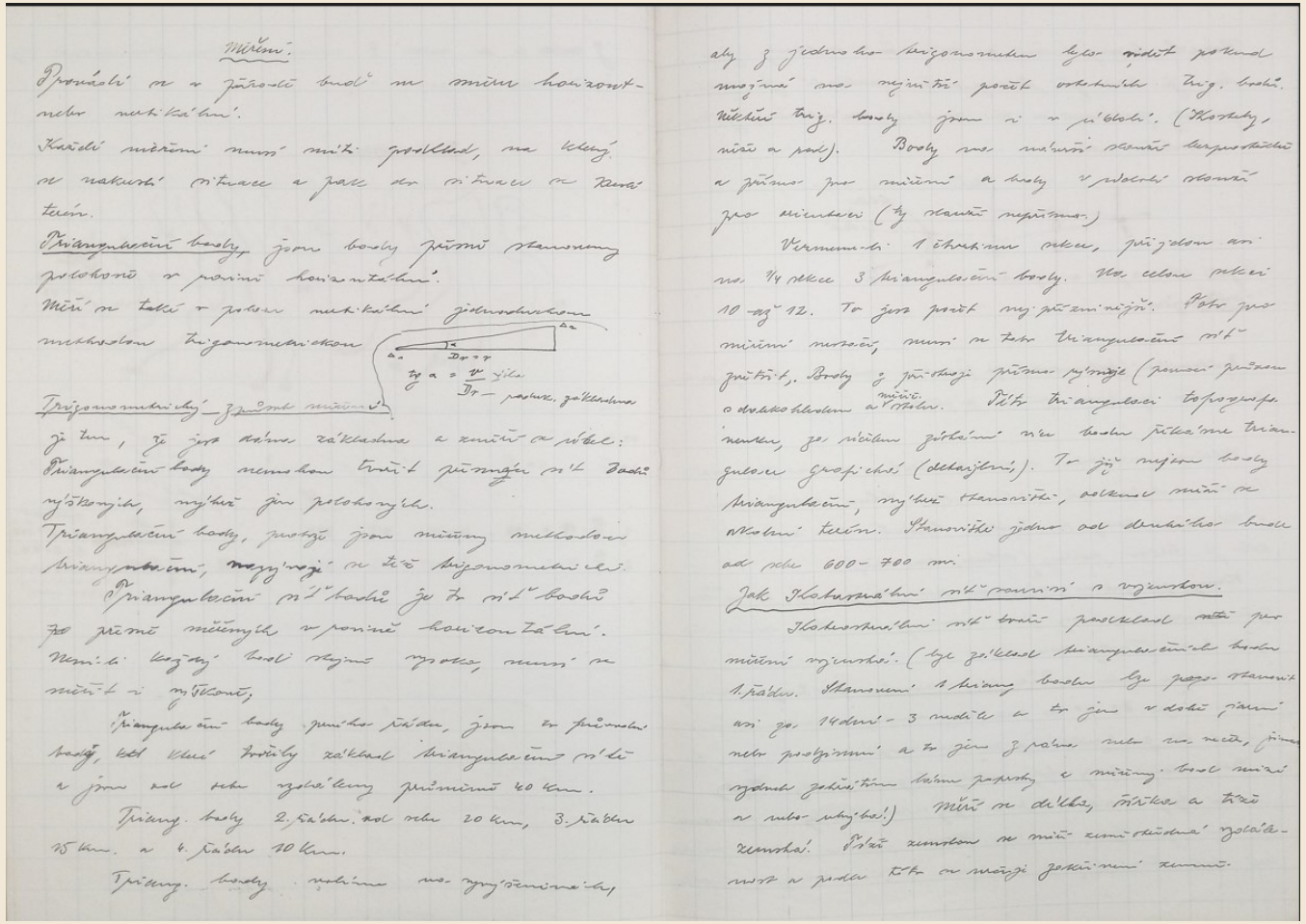
Uloženo v:

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

https://new.manuscriptorium.com/hub/catalog/default/detail/single/manuscriptorium%7CVMO-VMO K_24073 0U6ABL2-cs?lang=cs

Agregovaný model Finale 2.0

Ukázky rukopisných vzorků:



Otakar Jaroš
Nauka o terénu, školní sešit, čtverečkovaný papír

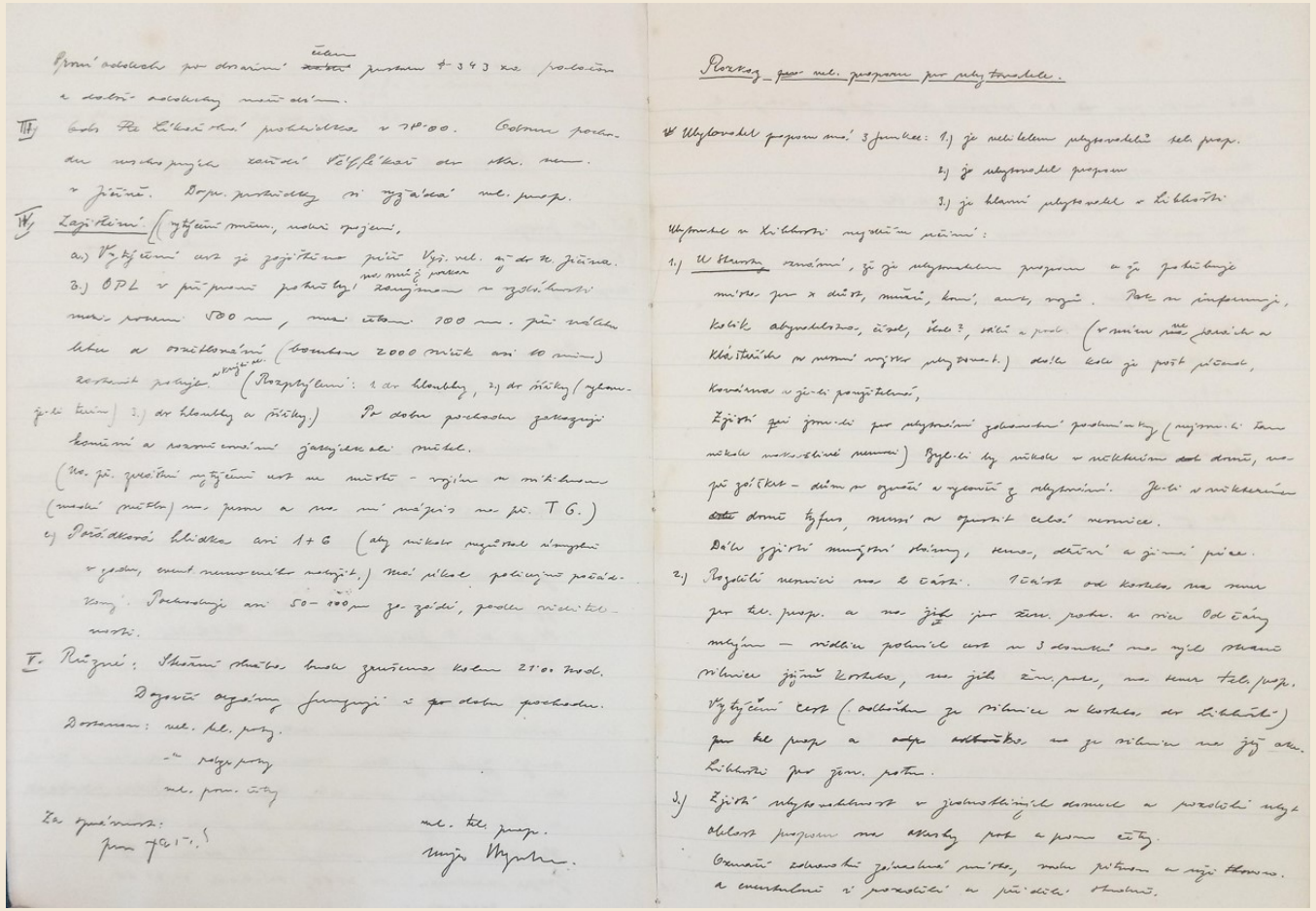
Uloženo v:

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

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

Agregovaný model Finale 2.0

Ukázky rukopisných vzorků:



Otakar Jaroš
Nauka o terénu, školní sešit, linkovaný papír

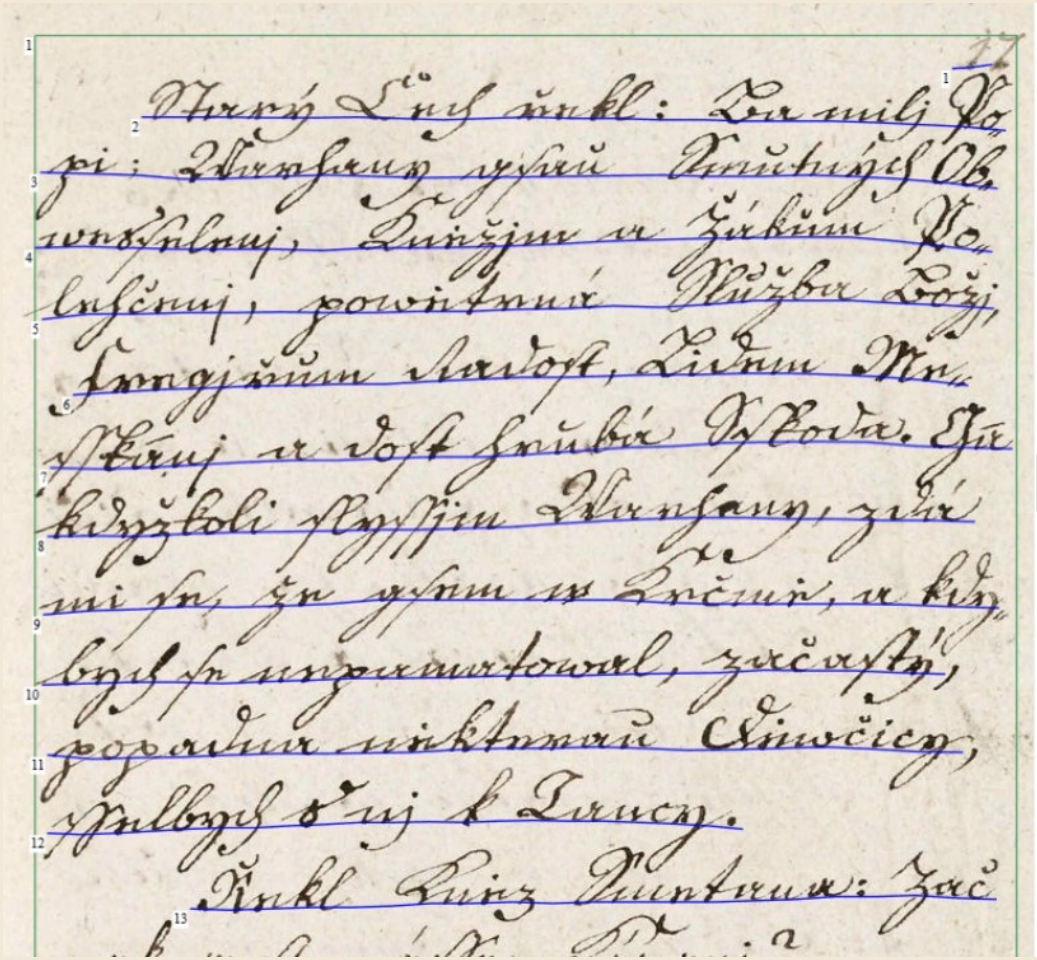
Uloženo v:

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

17
Stárý Čech řekl: Ba milj Po-
pi: Wárhaný gſaú Gmútných Ob-
webſelegj, Kněžjm a Žákúm Po-
lehčenj, powětrú Slúžba Božj,
Přepjřúm Radoſt, Lidem Me-
ſſkánj a doſt hrúbá Sſkoda. Gu
kdyžloli ſlyſſjm Wárhený, zdá
mi ſe že gſem w Krčině, a kdy-
bych ſe nepamatowal, začaſký
popadna některoú děwčicy
ſſelbych s nj k Čaúcy.
Řekl Kněz búletanx : Zač
pak geſt náſſe Kázágj?
Stary Čech řekl: Rázánj wá-

Zdroj:

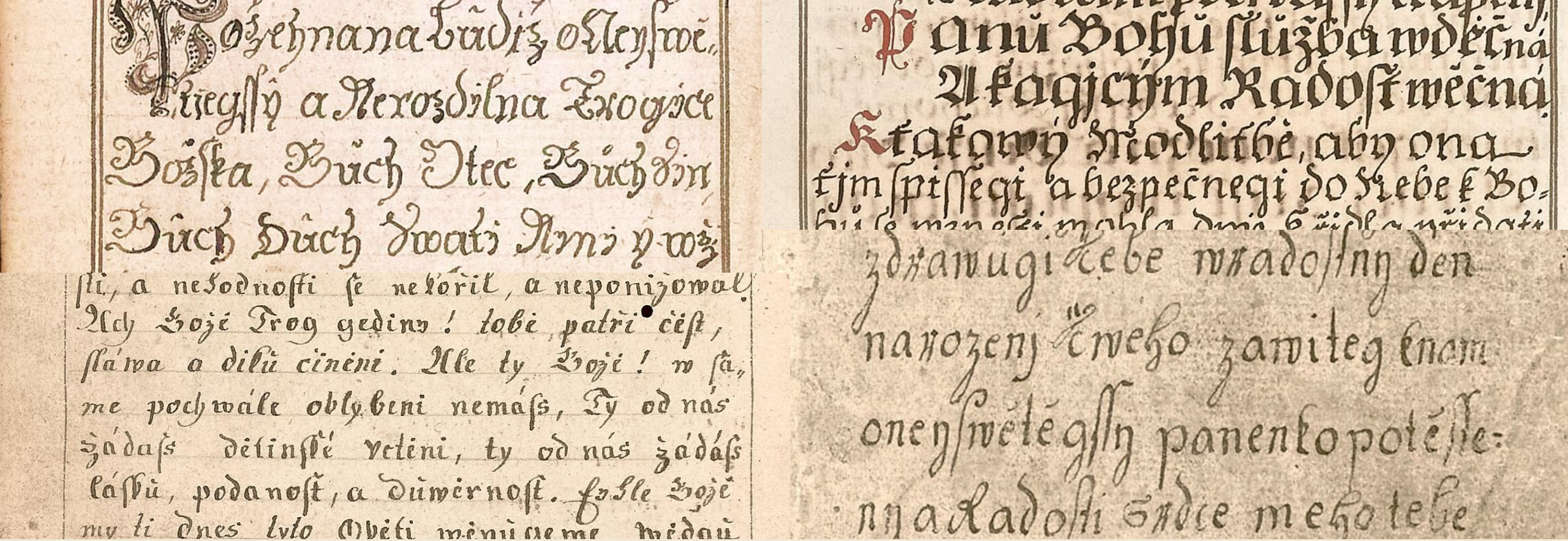
Rozličné písně starožitné

Uloženo v:

Moravská zemská knihovna v Brně
RKP-0048.022; Rkp 59

AUTOMATICKÁ TRANSKRIPCE POMOCÍ PLATFORMY TRANSKRIBUS

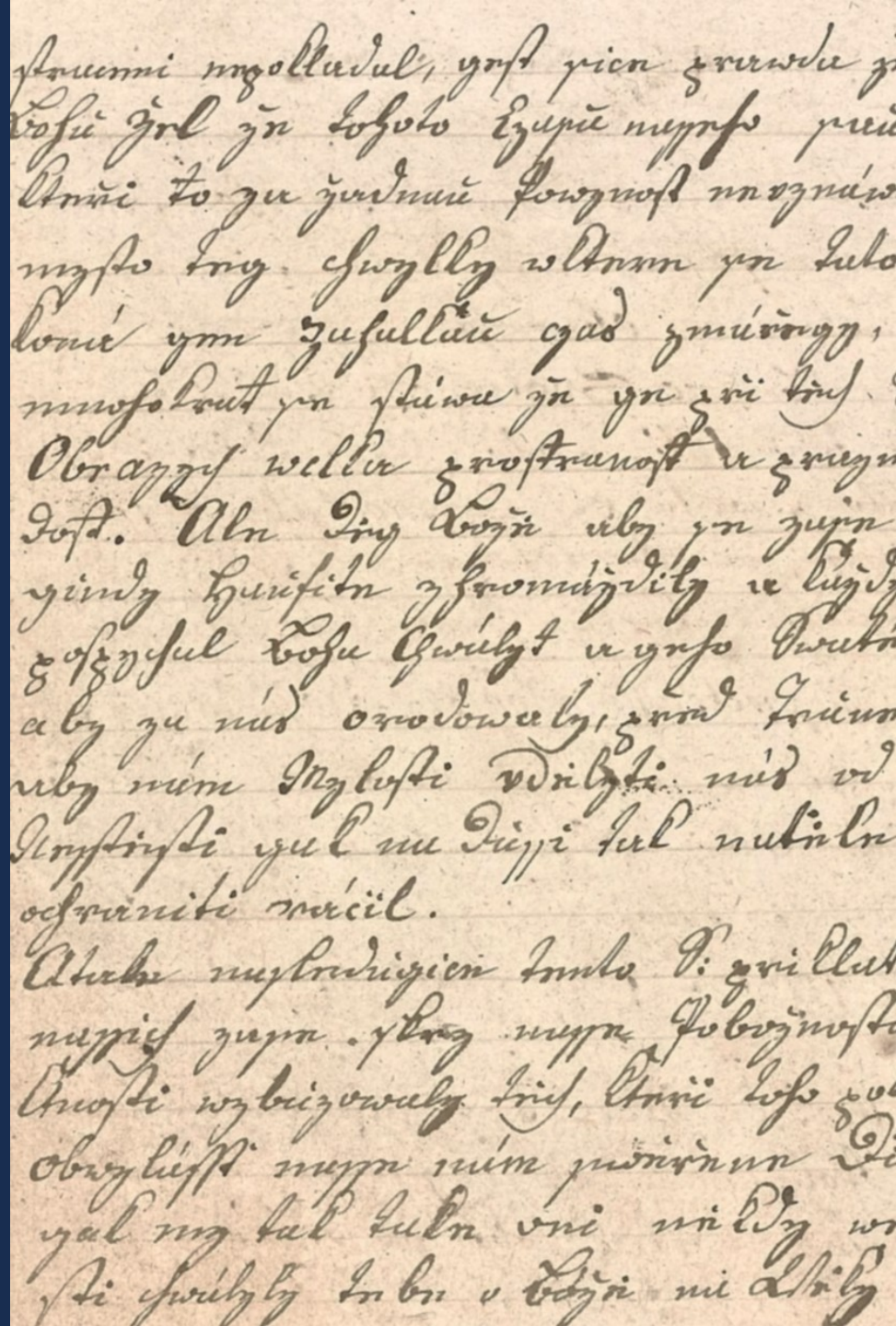
VOJTĚCH ŘÍHA



MODEL AGREG-8

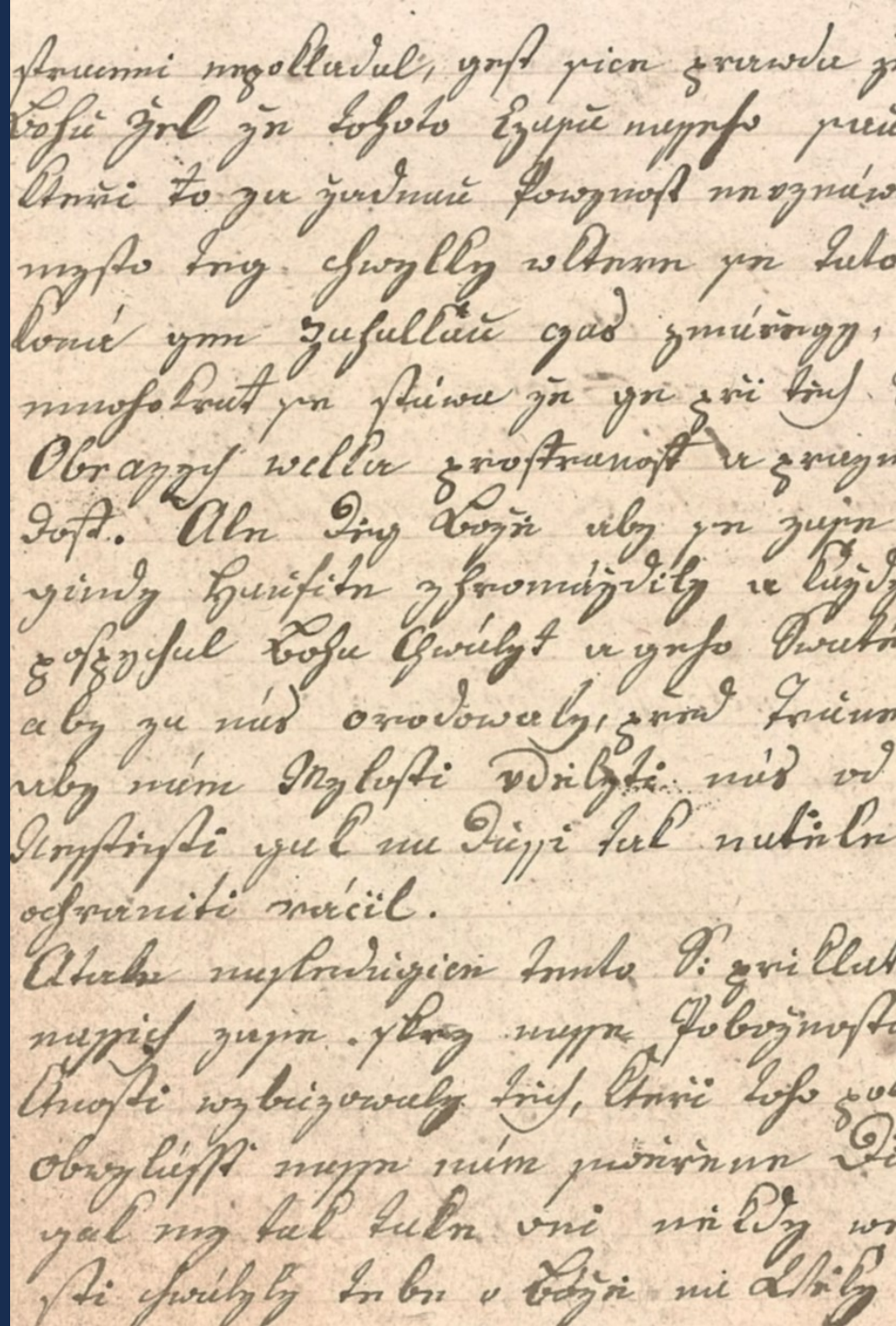
Model Agreg-8

- 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)*



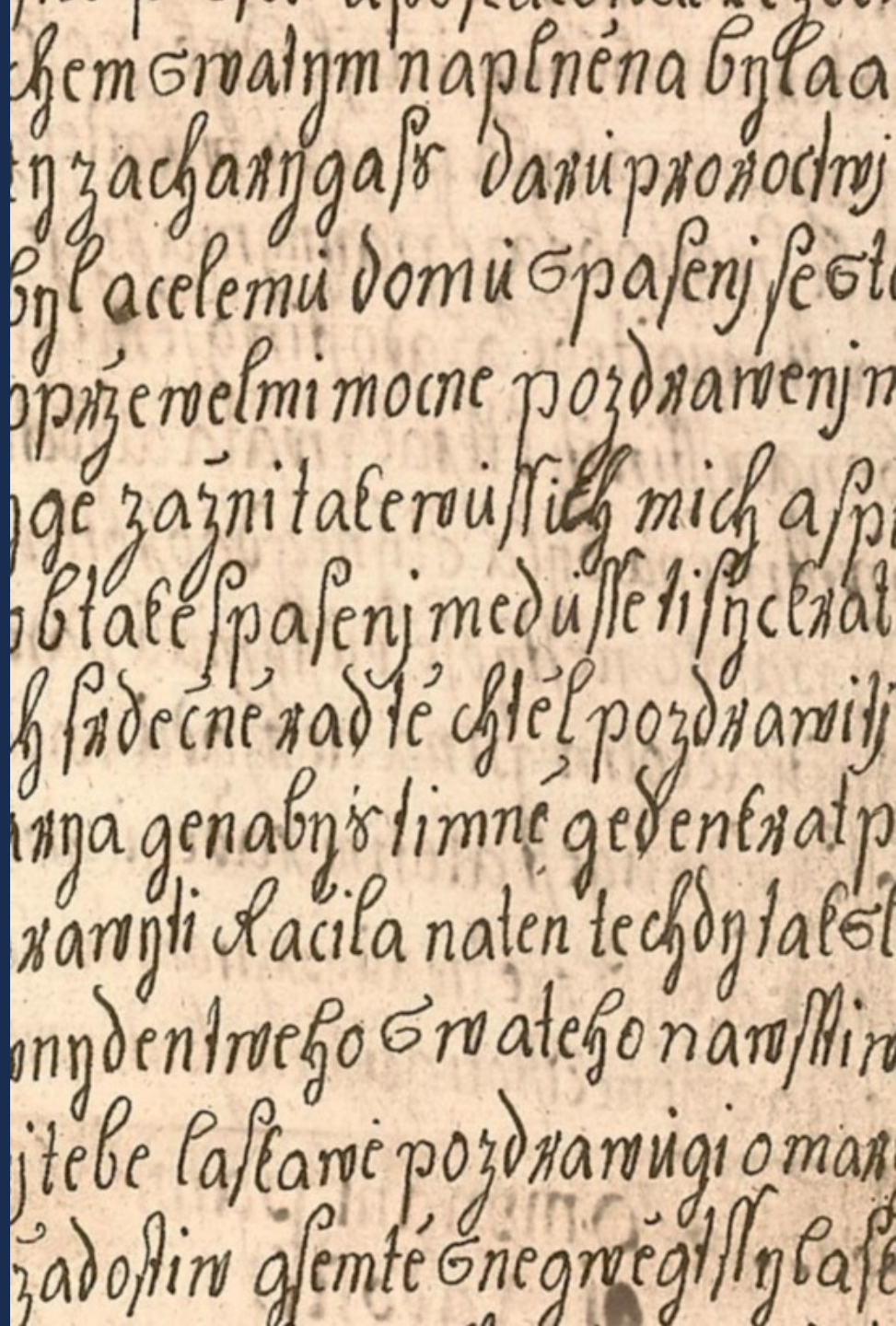
Model Agreg-8

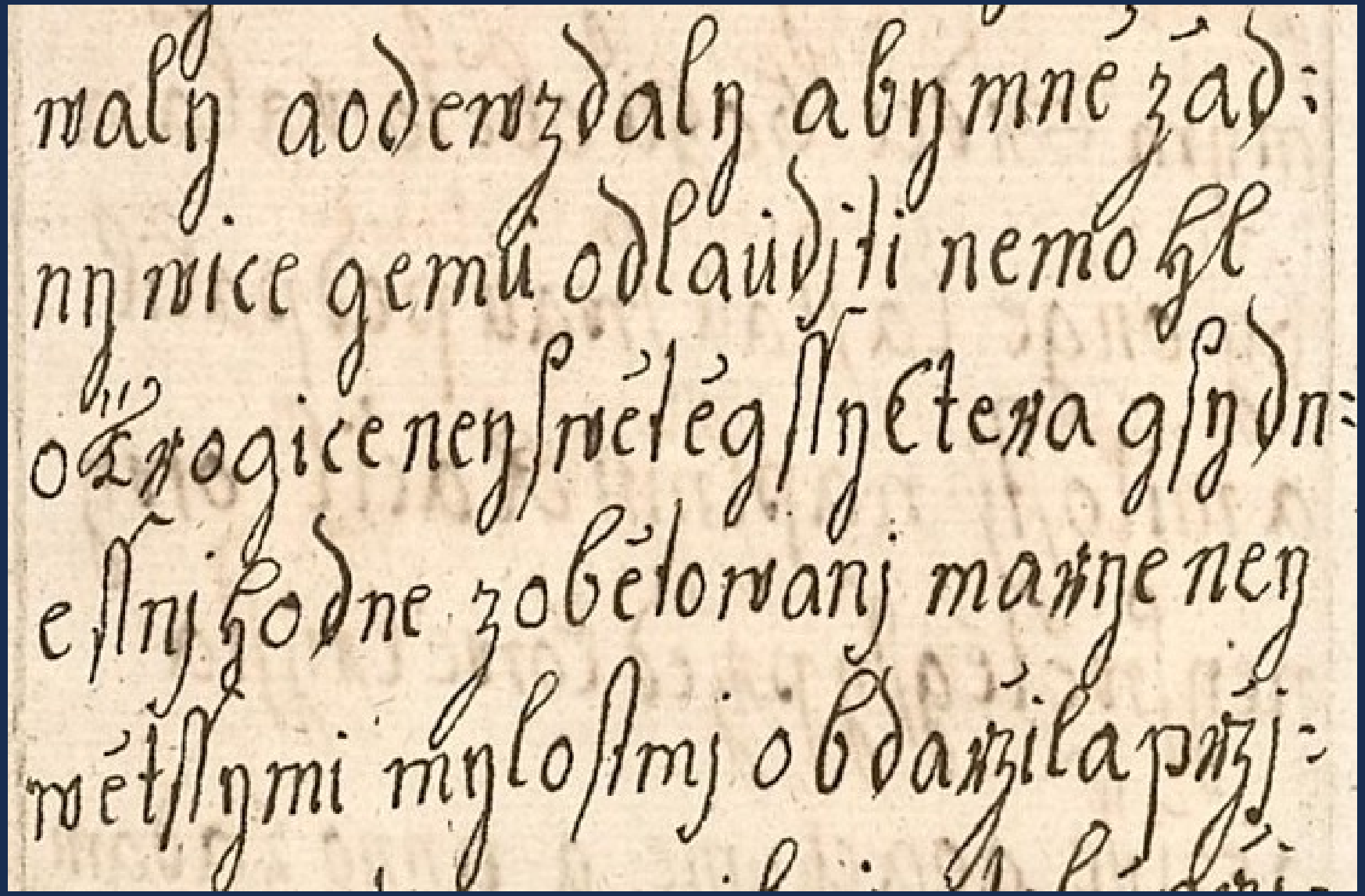
- HTR model kombinující 5 rukopisných dokumentů
  - *Česká modlitební kniha*
  - *Cesta Svatocellenská*
  - *Radostná cesta*
  - *Modlitby, písně a litanie*
  - *Modlitební knížka*



1 . Česká modlitební kniha

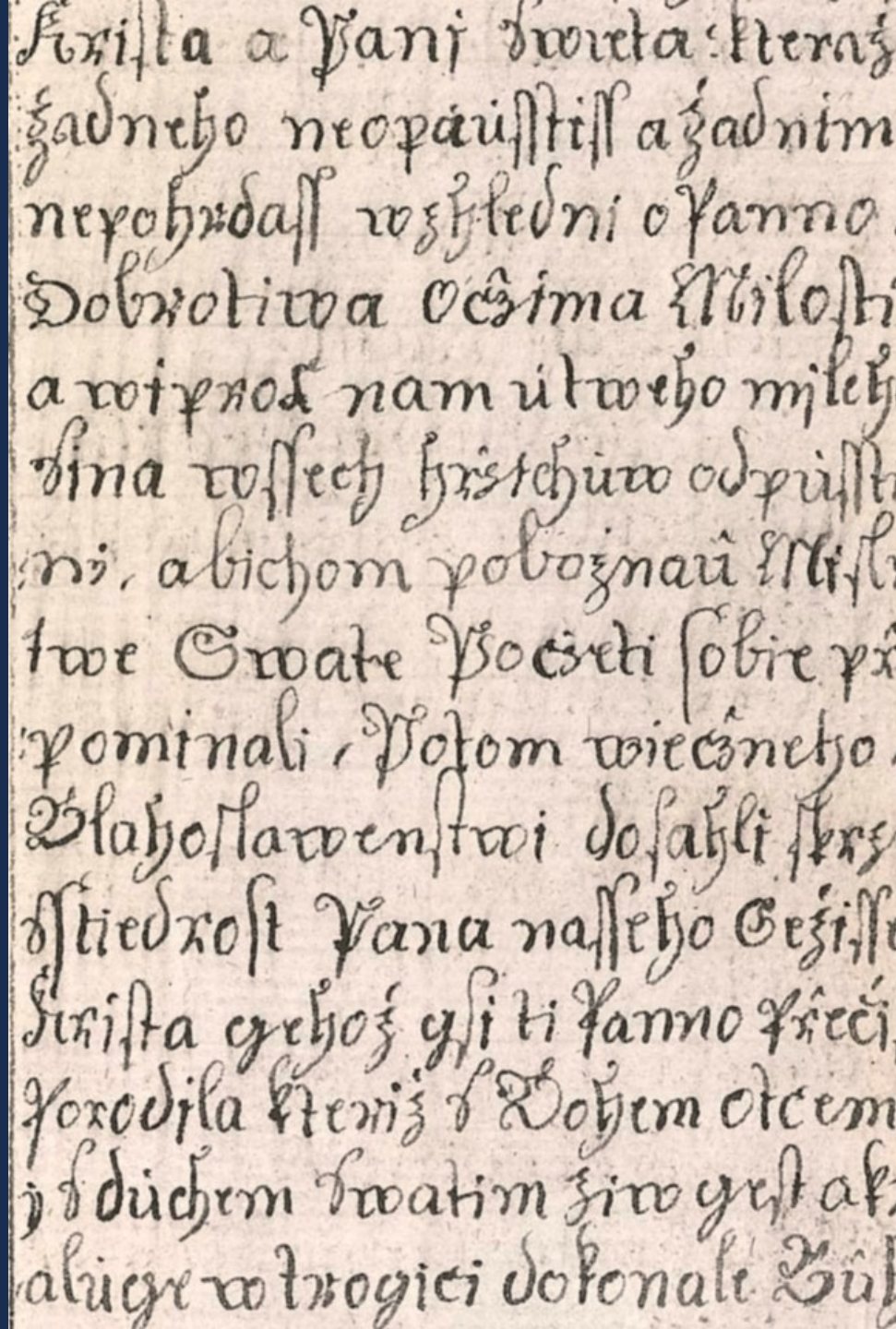
- Datace: 1733-1766
- Instituce: Muzeum Jindřichohradecka
- Signatura: RK 071
- Rozměry: 14 cm x 10,5 cm
- **Problematika zasahování do sousedících řádků**

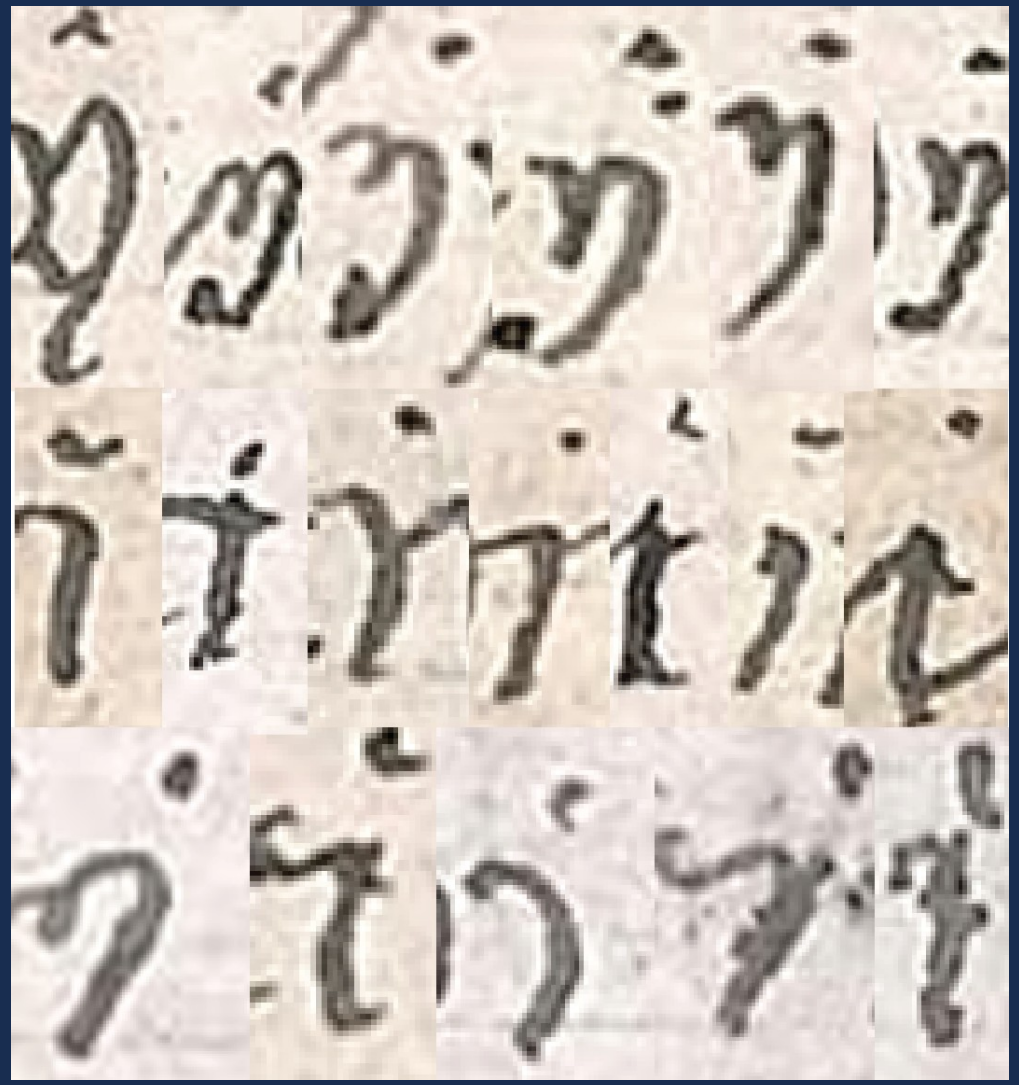




2 . Cesta Svatocellenská

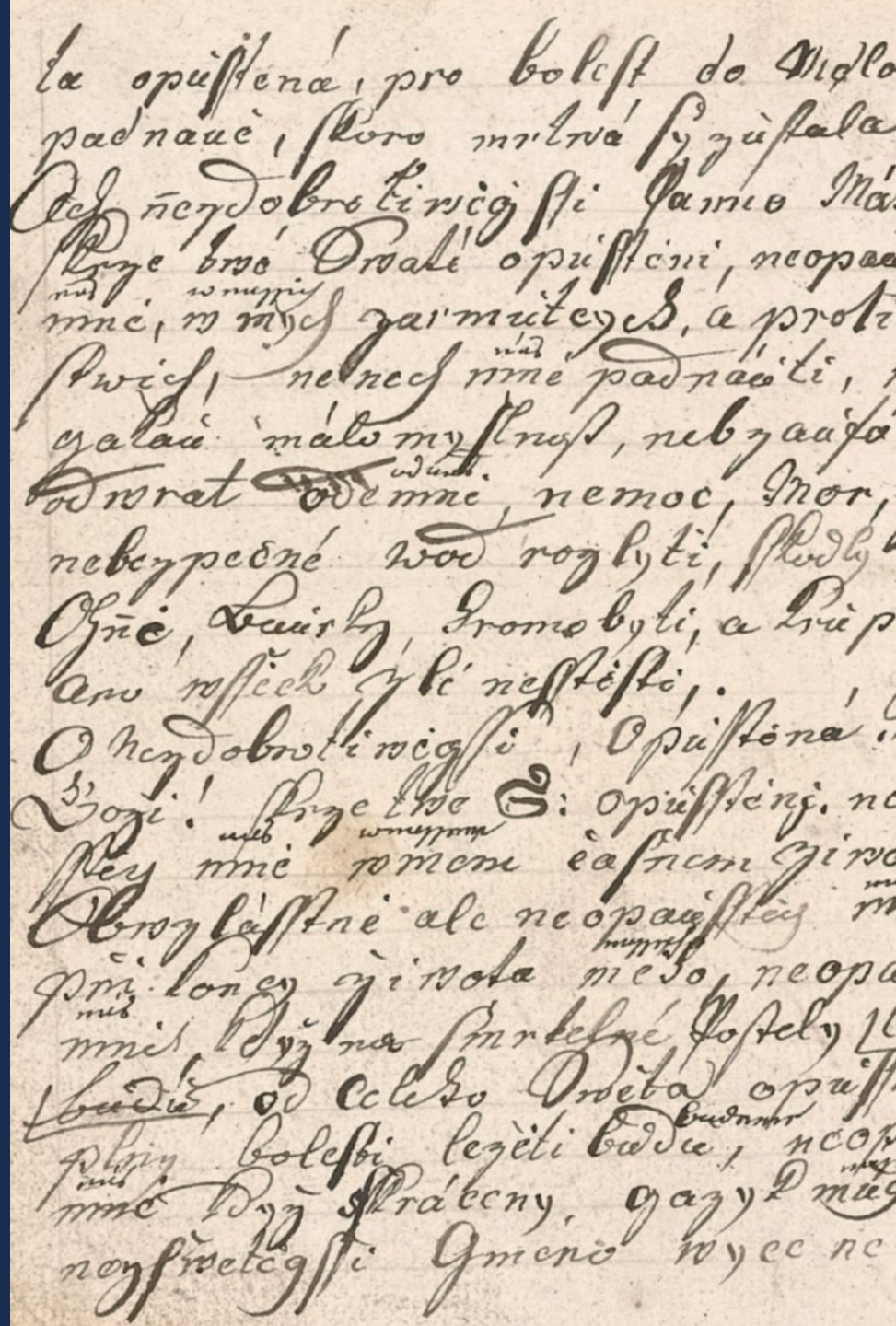
- Datace: 1733-1766
- Instituce: Muzeum Jindřichohradecka
- Signatura: RK 037
- Rozměry : 18 cm x 11,5 cm
- **Problematika podobnosti i/y**
- **Používání obrovských iniciál (segm.)**

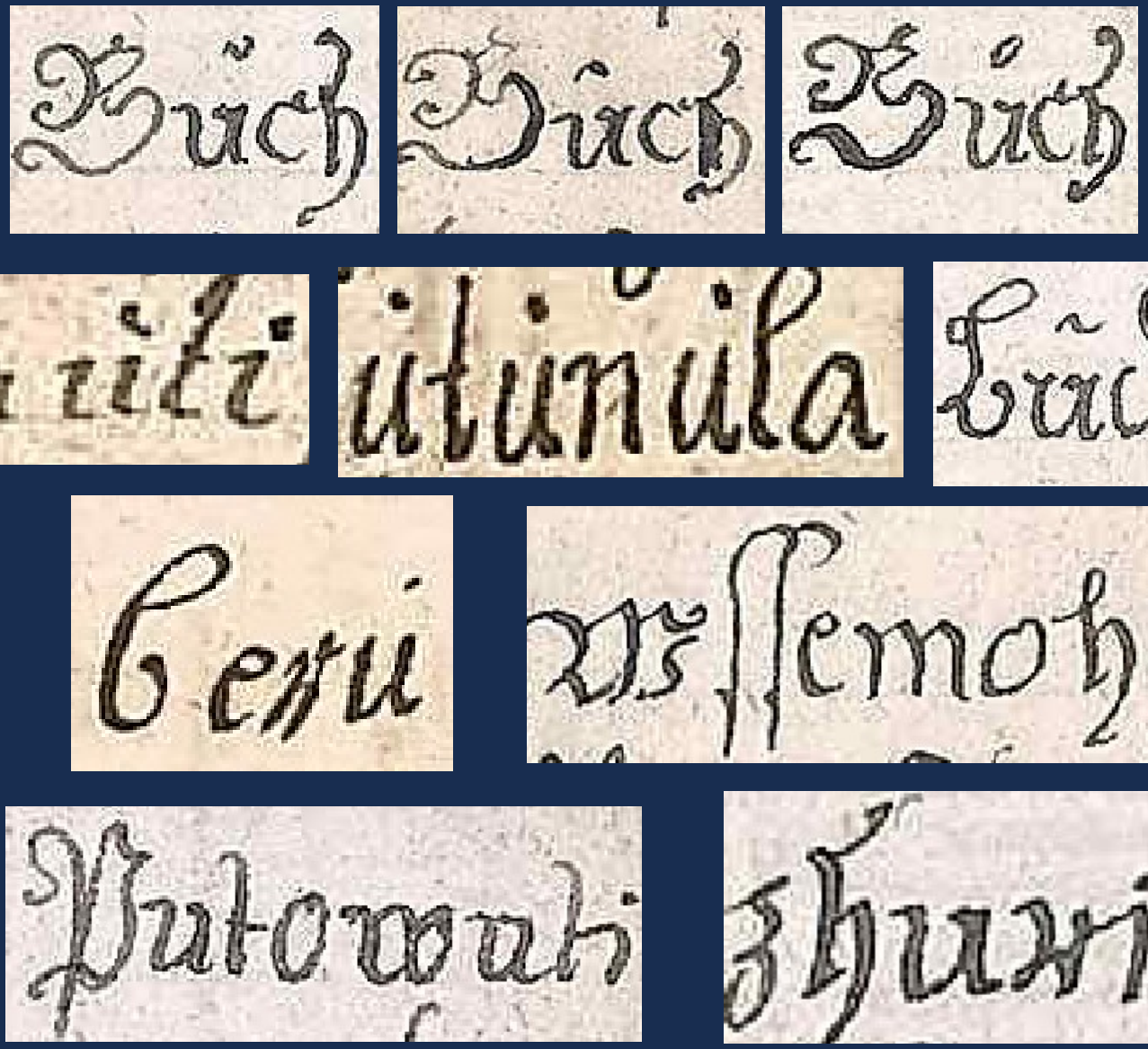




3. Radostná cesta

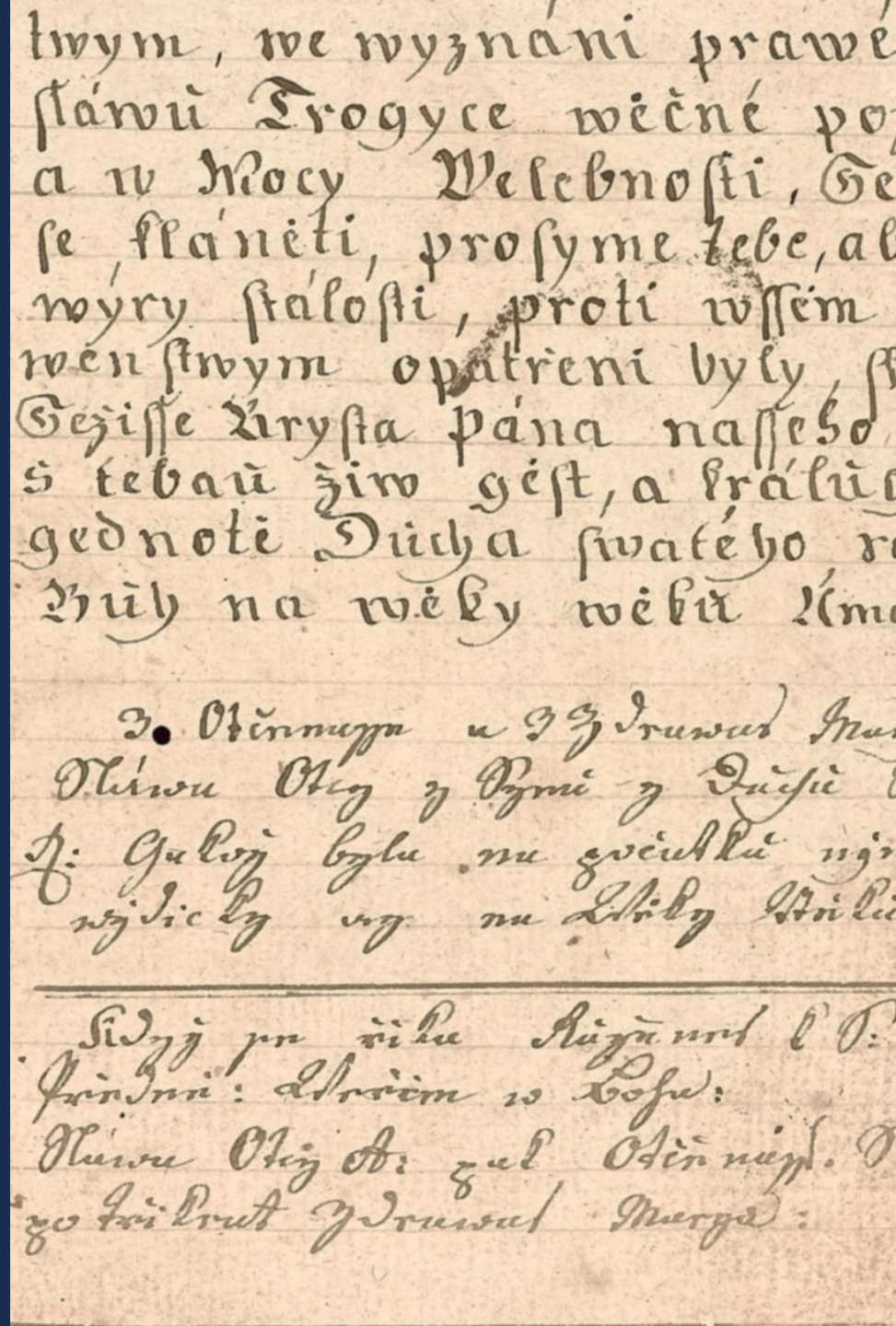
- Datace: 1829-1884
- Instituce: Moravské zemské muzeum
- Signatura: ST 2272
- Rozměry: 19,5 cm x 16,5 cm
- **Různost při psaní diakritiky**





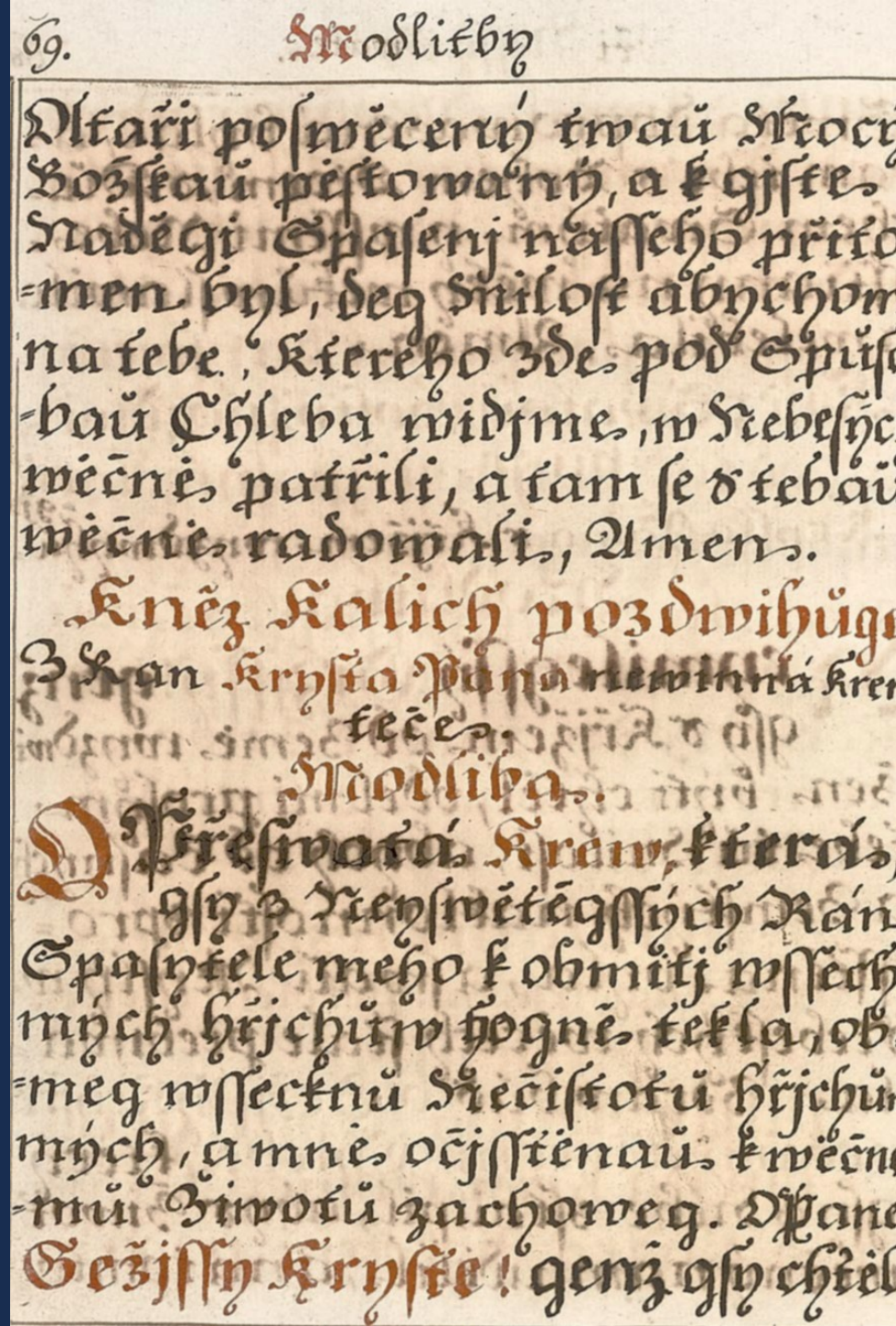
4 . Modlitby, písně a litanie

- 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)**



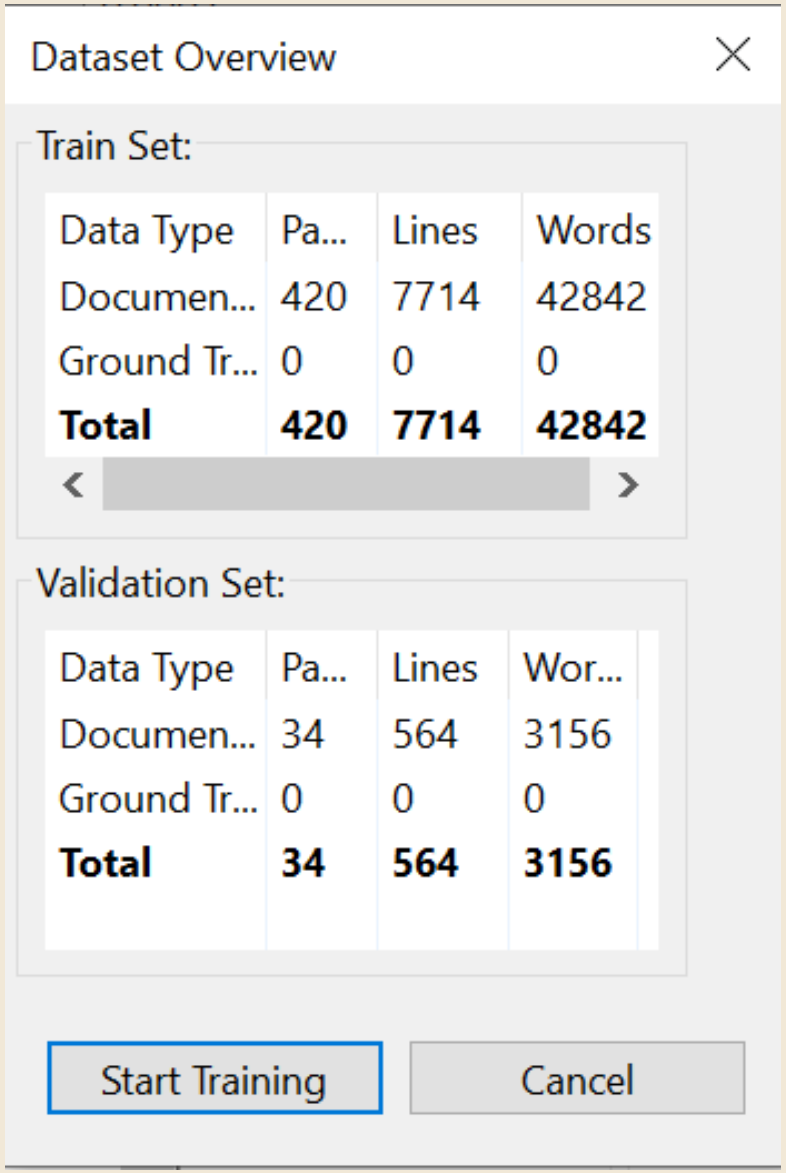
5 . Modlitební knížka

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



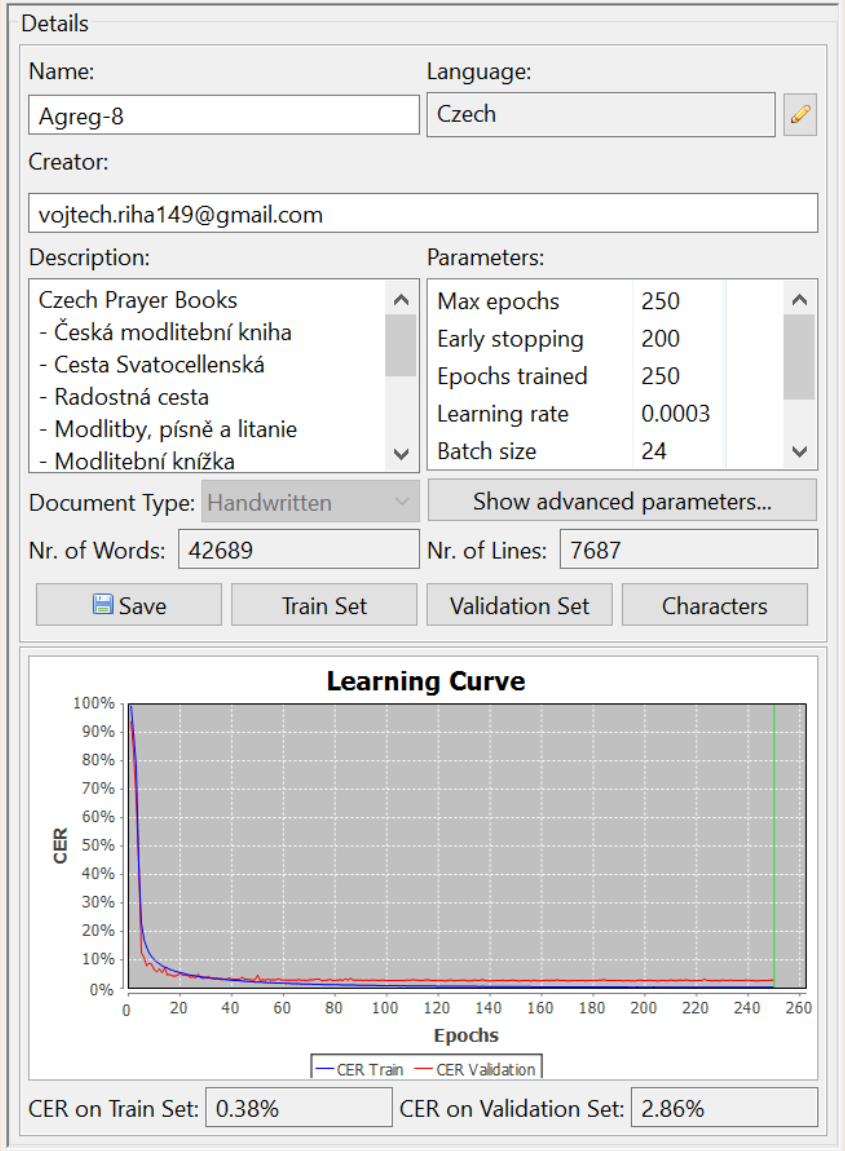
Model Agreg-8

- 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

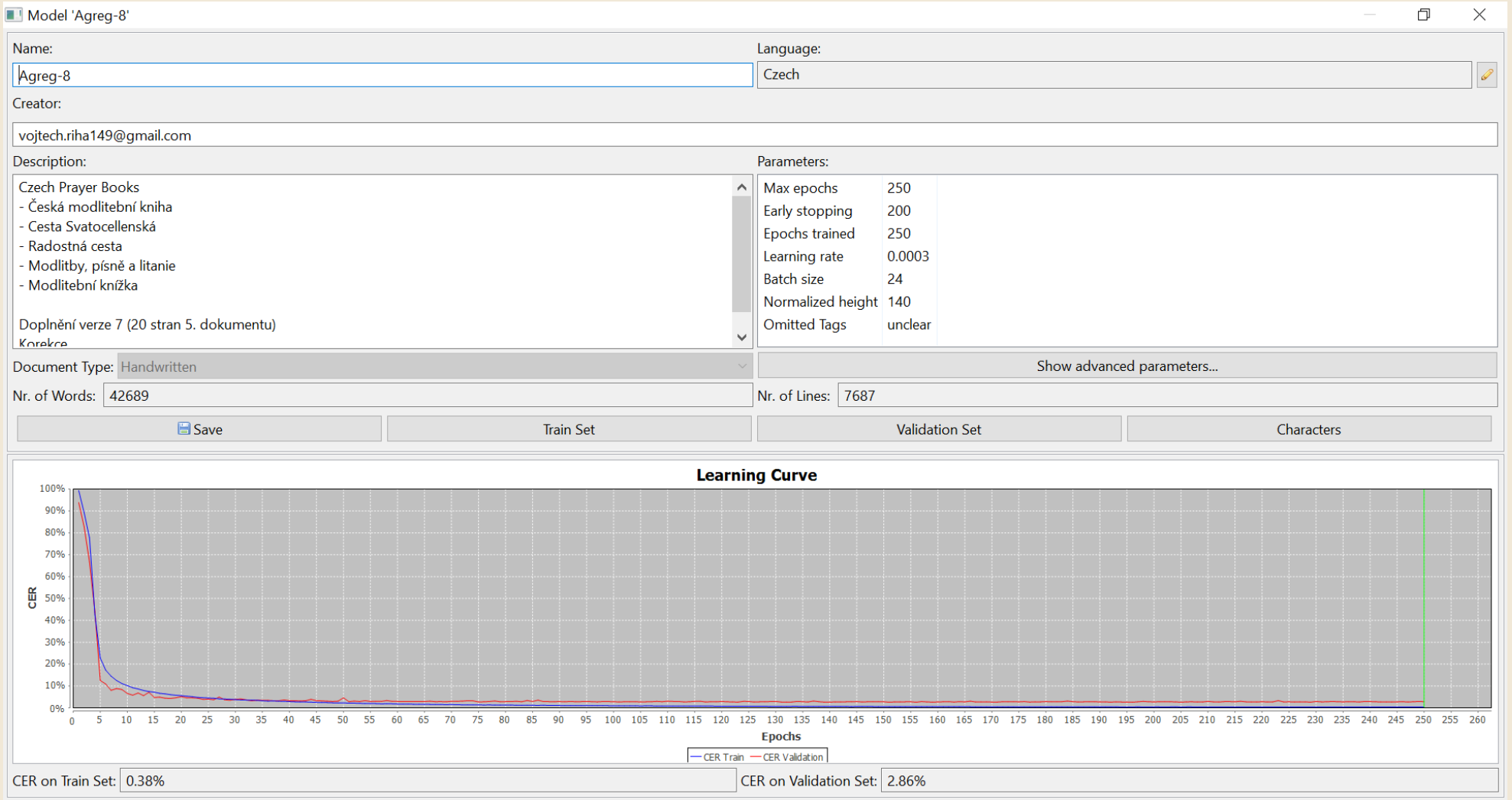


Model Agreg-8

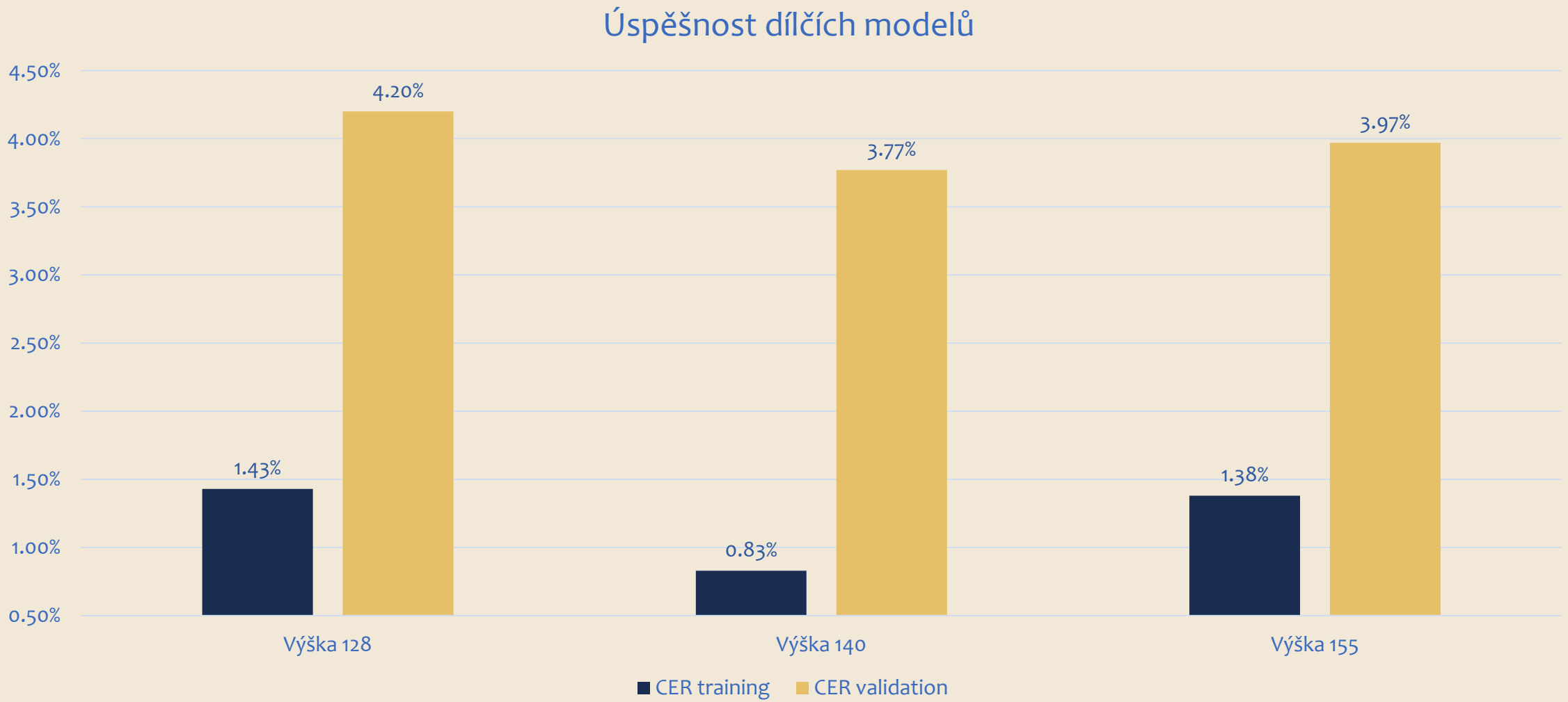
- Výsledek
  - *CER na trénovací sadě:* **0,38%**
  - *CER na validační sadě:* **2,86%**
  - *nejlepší epocha CER:* **2,60%**
  - *nejlepší epocha WER:* **16,57%**



Model Agreg-8



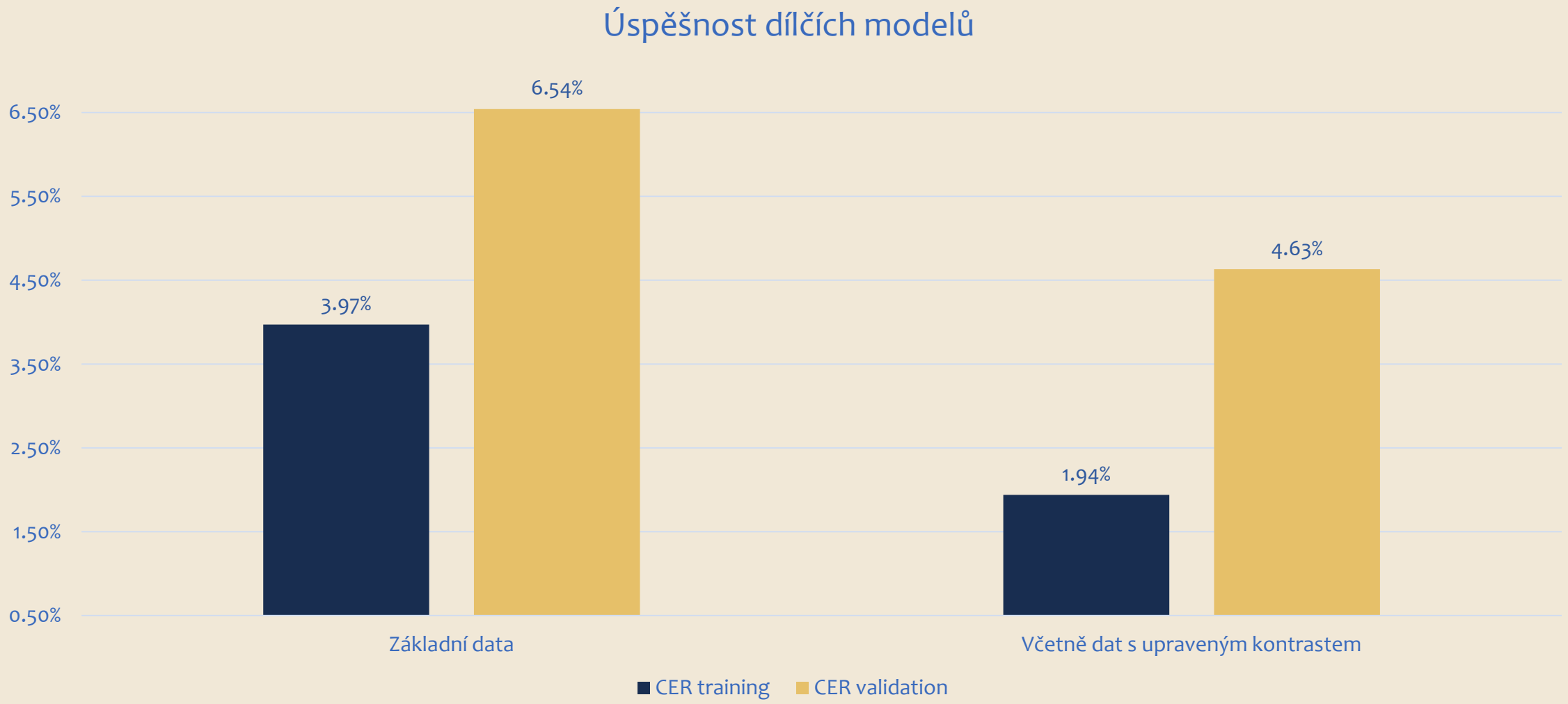
Korekce výšky řádku u modelu ČMK-70



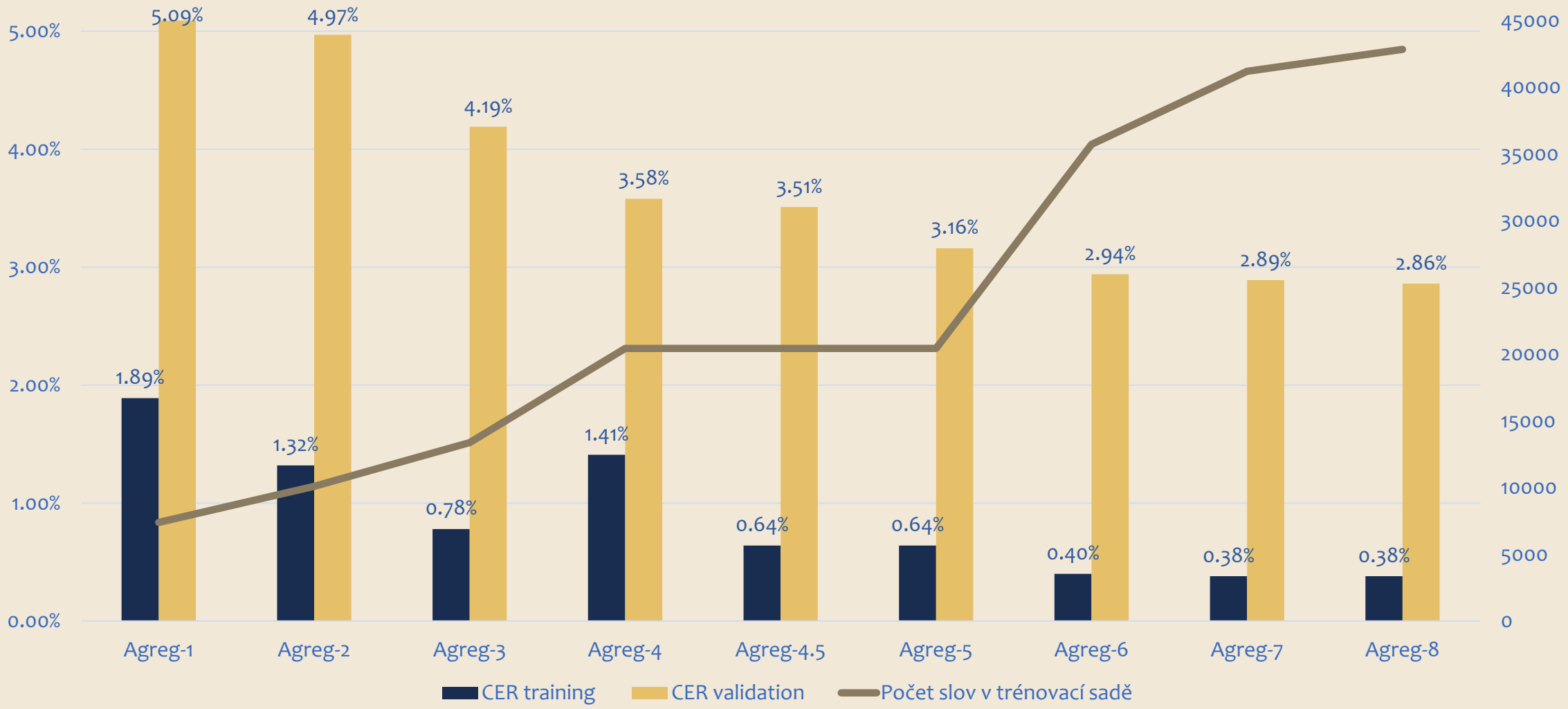
Augmentace obrazových dat

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



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