

Machine learning in Czech libraries - OCR for early printed and handwritten documents:

The PERO OCR project

Michal Hradiš

Brno University of Technology



Petr Žabička

Moravian Library in Brno



Digital libraries in the Czech Republic

- Kramerius digital library system (open source)
 - based on Solr, jpeg2000, MODS
 - OCR:ALTO, txt
 - Open API, IIF gateway available
- >40 libraries, >300 mil. unique pages
 - registr.digitalniknihovna.cz
 - www.digitalniknihovna.cz
- Common index: Czech digital library (in progress)



Goal: Make the content of digitized documents searchable

The screenshot displays a digital library interface. At the top, the logo for 'Moravská zemská knihovna' is visible. A search bar contains the text 'In öffentlichen digitalen Bibliotheksdokumenten suchen'. Navigation links include 'Sammlungen', 'Durchsuchen', 'Informationen', and 'Anmelden'. A German flag is shown in the top right corner. On the left, there are icons for PDF, image, and text, along with a search bar 'Im Dokument suchen' and a page indicator '2 of 39 Seiten'. A grid of document thumbnails is shown, with the second thumbnail in the first row highlighted with a blue border. The main area shows a large image of a handwritten document page. A red rectangular box highlights a section of the text. To the right, a sidebar provides metadata for the document, including the collection name, author, and publication details.

Moravská zemská knihovna

In öffentlichen digitalen Bibliotheksdokumenten suchen

Sammlungen Durchsuchen Informationen Anmelden

Im Dokument suchen

2 of 39 Seiten

Sammlung von verschiedenen Stücken für das Piano-Forte

Geistige(r) Schöpfer
[Kovařík, Franciska](#)

Mit Unterstützung von
[norway grants](#)

Publikationsangaben
1826

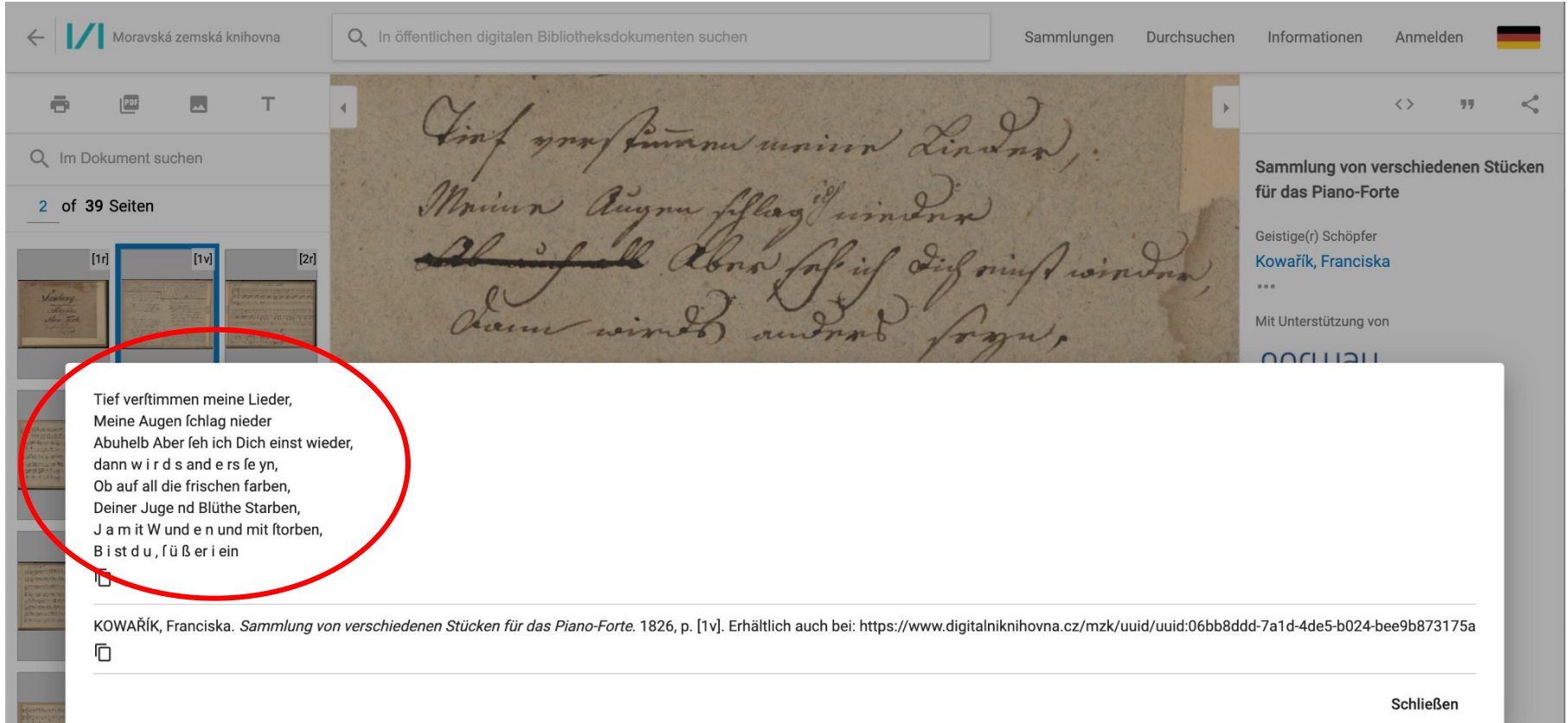
Objektkategorie
[Musiknoten](#)

Schlagwörter
[Hudba pro klávesové nástroje](#)
[Tance \(klavír\)](#)
[Sonáty \(klavír\)](#)
[Písně \(zpěv, klavír\)](#)

Ein grossmännlicher Mann,
Mein Auger schlag' wieder
~~Ab auf all die freigeu hander,~~
denn wieder andrad' freye,
Ab auf all die freigeu hander,
denn Auger schlag' wieder,
Ja mit Schindern und mit Hander,
Leist' ich, frey' und rein.

A Klavi' miltwar' mit pifur. Die ma' fa' zum huan' fofur
Lohnkim' p'w' f'z' ad' j'ra' hifur' Wank' is' W'ib' f'f'ig
F'ig' d'ly' d' y'z'w'ang' h'w'ad' d' h'f' y' h'w'ic'and' p'f'z' g'ra
M'ad' d'f' f' h' f'w'ud'ua' h'w'ic'and' h'w'ic'and' h'w'ic'and'

Goal: Make the content of digitized documents searchable



The screenshot shows a digital library interface for the Moravská zemská knihovna. The main content is a handwritten document page with the following text:

Tief verstimmen meine Lieder,
Meine Augen schlag nieder
Abuhalb Aber seh ich Dich einst wieder,
dann w i r d s a n d e r s e y n,
Ob auf all die frischen farben,
Deiner Juge nd Blüthe Starben,
J a m i t W u n d e n u n d m i t f o r b e n,
B i s t d u , f ü ß e r i e i n

A red circle highlights the German text. Below the document, there is a search overlay with the following text:

Tief verstimmen meine Lieder,
Meine Augen schlag nieder
Abuhalb Aber seh ich Dich einst wieder,
dann w i r d s a n d e r s e y n,
Ob auf all die frischen farben,
Deiner Juge nd Blüthe Starben,
J a m i t W u n d e n u n d m i t f o r b e n,
B i s t d u , f ü ß e r i e i n

Below the search overlay, there is a citation:

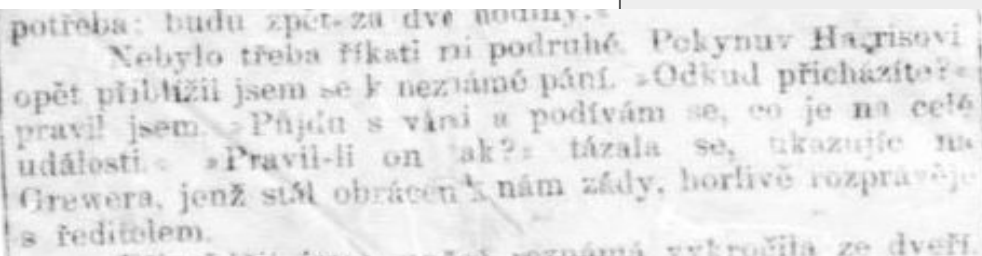
KOWAŘÍK, Franciska. *Sammlung von verschiedenen Stücken für das Piano-Forte*. 1826, p. [1v]. Erhältlich auch bei: <https://www.digitalniknihovna.cz/mzk/uuid/uuid:06bb8ddd-7a1d-4de5-b024-bee9b873175a>

At the bottom right, there is a button labeled "Schließen".

OCR Quality

Previous
OCR

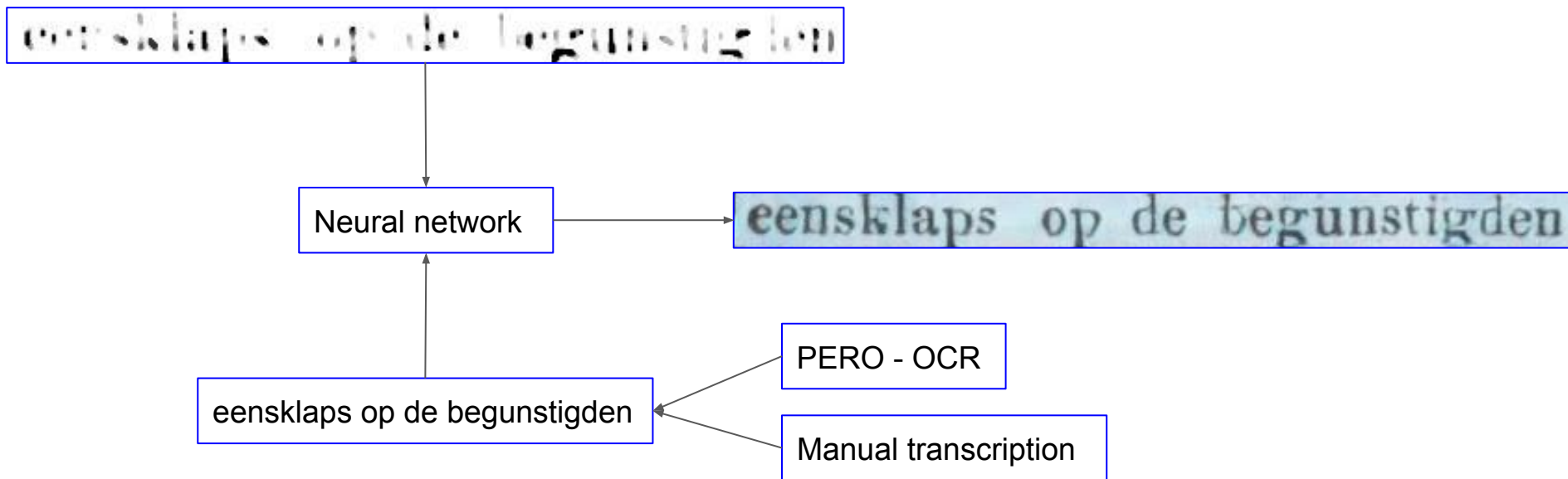
-livlo i ludi i |H>d nl.é l'- V r. Haj-isov; přichazít:- ■ i i- ii«
ceh dál osti i i Pra vil- nž stál , nera l



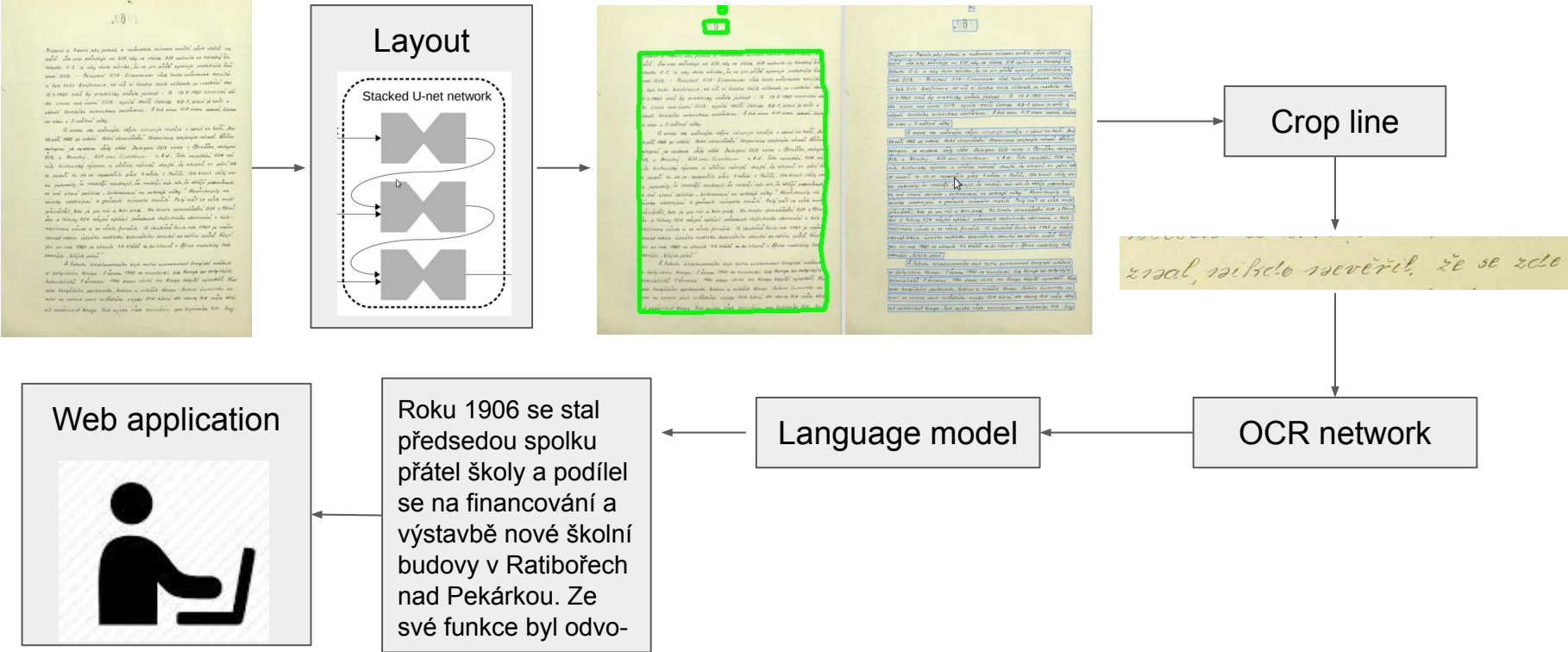
PERO

Nebylo třeba říkati ni podruhé. Pokynuv Hagnisovi
opět přiblížil jsem se k nez**t**ámé pání. »Odkud přicházíte?«
pravil jsem. »Půjdu s v**an**i a podívám se, co je na celé
události.« »Pravil-li on **ak**? tázala se, ukazujíc na
Grewera, jenž stál obráčen**k** nám zády, horlivě rozprávěje
s ředitělem.

Experiments: Image Improvement



PERO: Full OCR pipeline



Detection of paragraphs and text lines

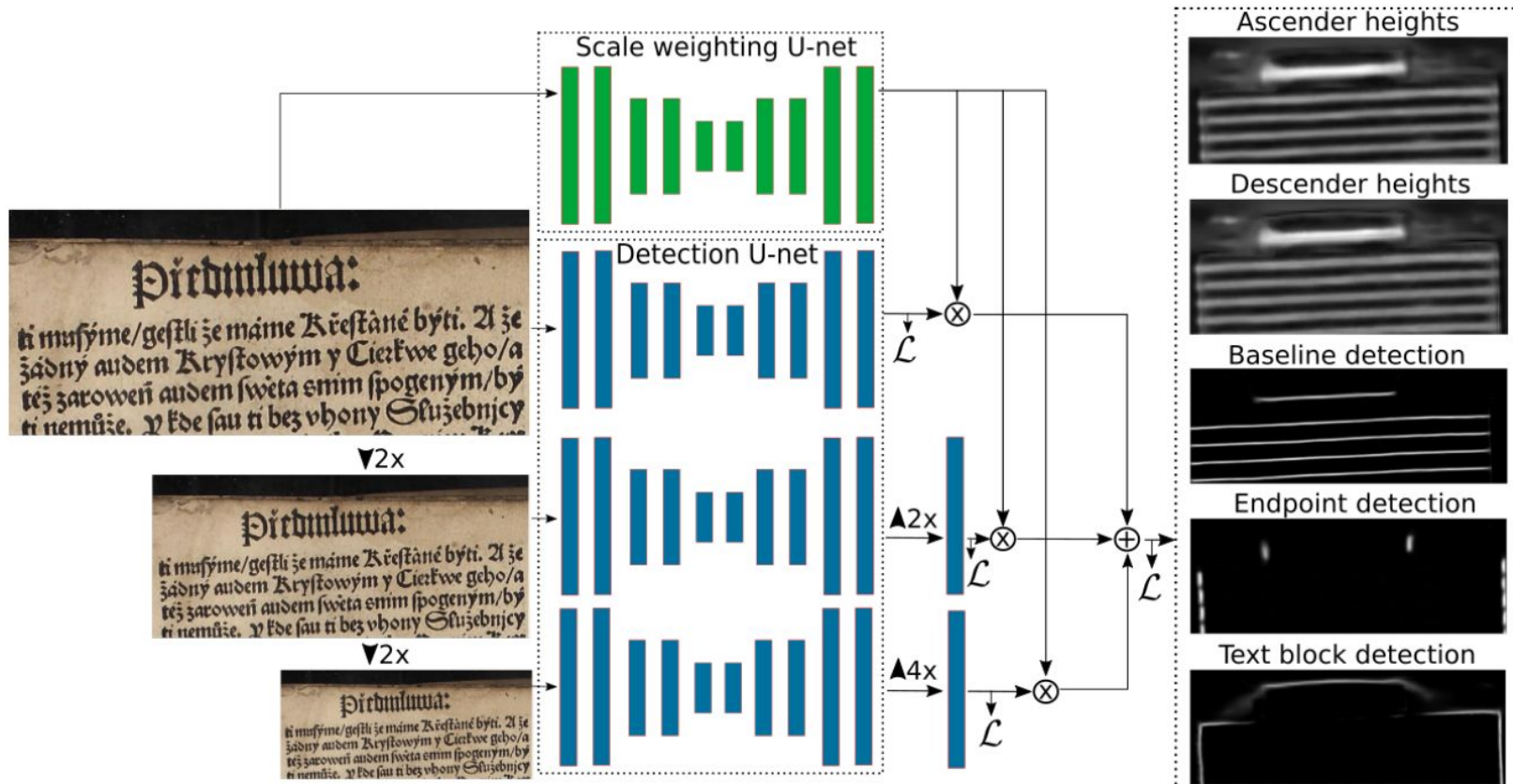
The image displays a grid of 12 small advertisements, each with red bounding boxes around individual text lines and blue arrows pointing to paragraph boundaries. The advertisements are as follows:

- Koupím regály.** hledáte-li, zaplatíte a inzerát do "Lidových Novin" a nemá udělat, kdo jste, co umíte a jaká místa hledáte neboť Lidové Noviny? Hlas obchodníci, zamestnanci, hospodáři, státníci Vám práci mohou dát.
- Učene.** hodného chlapce z řádné rodiny přijme do učení na knihárství za výhodných podmínek. Ant. Altrichter, knihář ve Františkově n. R. Morava 9512.
- Inteligentní slečna.** učitelka vyučuje hru na piano. Lask. nab. pod adr. Půlnost' poštovní hlavní pošta Brno 2155.
- Přijme se dělník.** a hoch, do učení, na nábytkovou práci, u Ig. Diblíka, Solnoúřední ul. 15. I. 2158.
- Reznický.** a zručný obchodník blízko města, se prodá. Adresa, v. admin. 1. 2151.
- Zemědělský polir.** asi s. 10, zedníky přijme se ihned pro dení, po případě, pro akordní práci. K. doručení ve státní kanceláři Josefa Rohana, Místek 2139.
- ky, s. kmínkou, máton peprou, vanilkou výtečně, chuti a jako led ochlazené, jediné, u Ig. Isakiewicze, Nádražní nám. 8901.**
- Olomoucké syrečky.** výtečné, velké, čtyřnášákové, 1 kopek za 1 kor. 32 hal. dohrátek v bednách a 8 kopek zašlá Fr. Noha, Olomouc, II. Pfl. 60 kopek franko! Krásné plakáty zdarma! 1859.
- Zámečnickví.** portální a stavební, rolety, plechové žvýer (meir. 9 K), plachty párové, plechové puťenky pro stavitele, železo na sporáky vždy na skladě, niklovyápi, mosazování, elektrickým proudem jediné, u 744.
- Jonáše v Brně.** Pekařská ul. 28.
- Učene.** hledáte-li, zaplatíte a inzerát do "Lidových Novin" a nemá udělat, kdo jste, co umíte a jaká místa hledáte neboť Lidové Noviny? Hlas obchodníci, zamestnanci, hospodáři, státníci Vám práci mohou dát.
- Zaměstnavatelé! Pozor!** Přijme, přesniky, tovaryše, dělníky a sluháckého pasážistů, nejlépe, prostřednictvím inzerátů "Lidových Novinach". Učazíte sábiak velké početí neboť "Lidové Noviny" jsou největším, českým, deníkem na Moravě.
- Přijme se.** schopný, kovářský pomocník, do stálé práce u p. Leoše Mlčocha v Těšnovicích u Kroměříže. 2120.
- Snazlivý.** mladý bednář k výrobě škopků, se ihned přijme u Jana Krudkého ve Staré Kárávě. 2152.
- Učene na pekařství.** z řádné rodiny přijme Math. Dufek v Mor. Olešnici. 9494.
- Zámečnického.** dělníka, a jednoho učně přijme ihned Jos. Hníbk, strojní a star. z.
- Bryčka na pérách.** hjata se lacino prodá. Ant. Skřara Brno, Nová ul. 81. 2128.
- Výprodej.** 9183 hospodářských strojů v útřích v továrně.
- 20 železných oken.** 100x1 m. svítlost, kamenné schody, prodá Pacák, Ryčlova ul. 8. 7. Brno, (Křánová) 7407.
- 4kolový.** dětský vozík se lacino prodá, Brno, Sýkova ul. (Schüttgasse) 8. 4 u domovněka. 2140.
- Med v plástech.** kg. za 2 kor. vymetání, kg. za kor. 1-60 dodává
- a předložení splátkových sřizenek obdržít skvostné hodinky.** (Časopis Prázdniny) Brno, Velké náměstí 28.
- Breje, - skřipce** a veskerézbožíoptické
- S. Vašíček** optik a mechanik, BRNO, Ferdinandova ulice 26. I. poschodí. 26.
- Provoznické zboží.** Cíaha družh, má na skladě Frant. Kopecký, provoznické mistr, Dominikánské nám. 8. 2. Zároveň přijme hocha do učení z řádné rodiny na 3 roky s celým zaopáčením. 1807.

Detetion of paragraphs and text lines is not easy



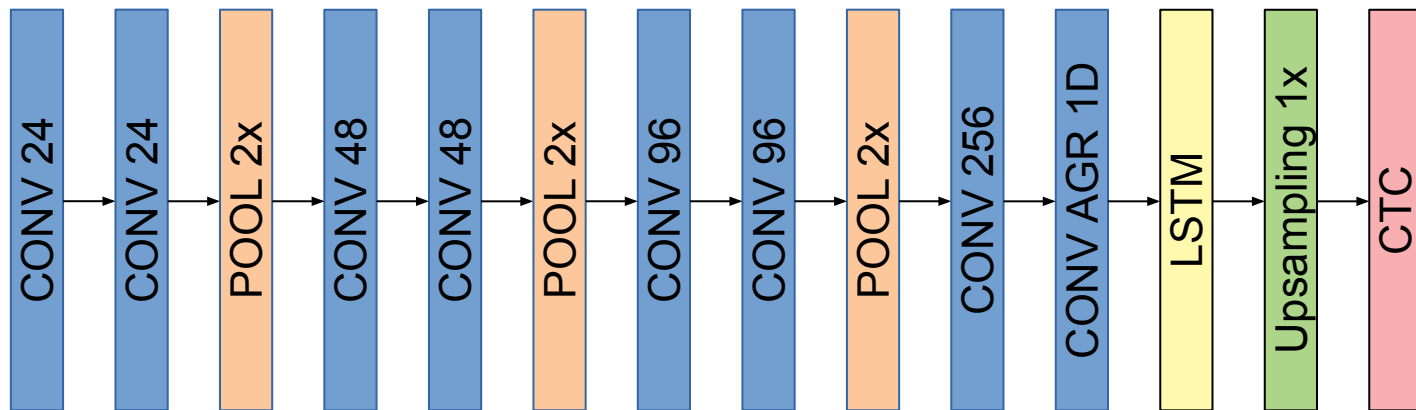
Text line detection - Fully convolutional networks



Text transcription - convolutional/recurrent networks + CTC loss

fer ran delicado que se deshaze en la boca, no se senti-

ser ran delicado que se des heze en la boca, co se seti-



Language models

- Recurrent networks working with characters.

Similar to working with words:

People like to go to ????????

library

cinema

pubs

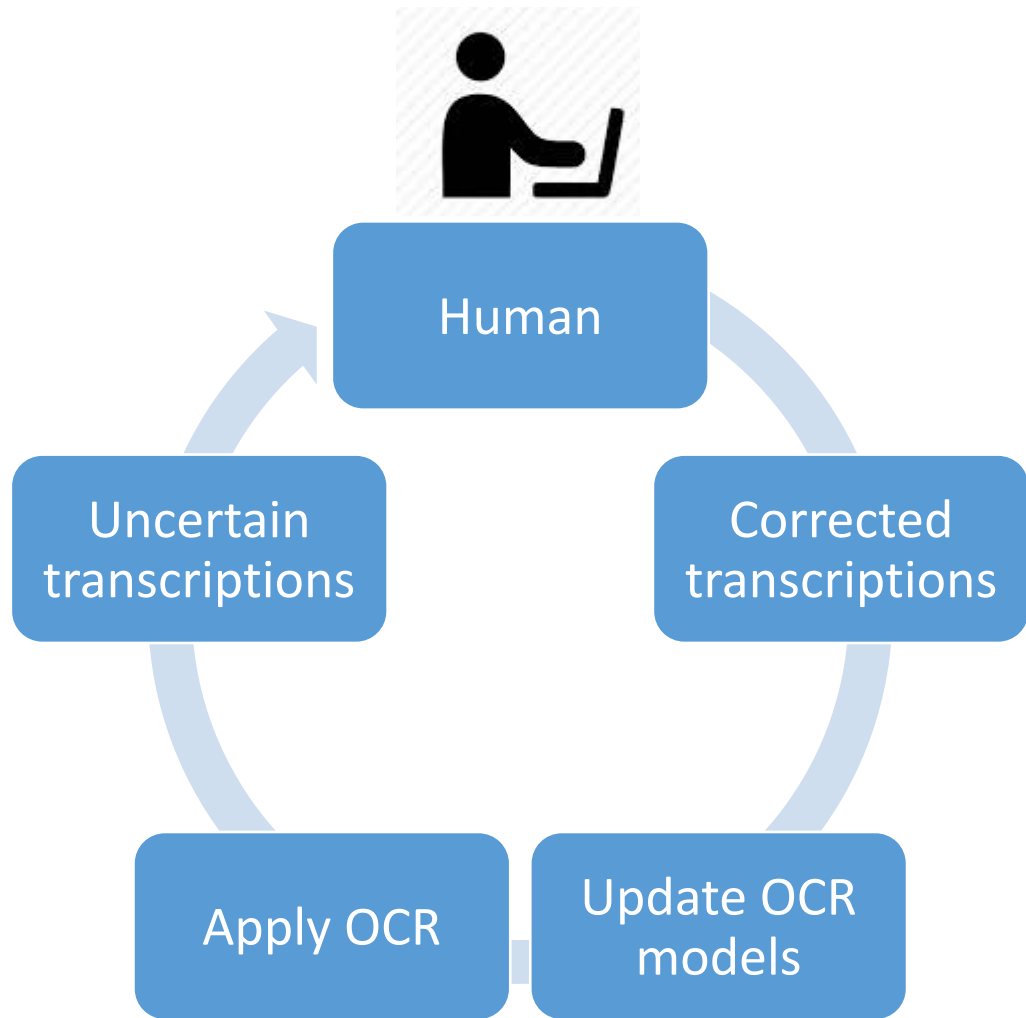
restaurants

....

work

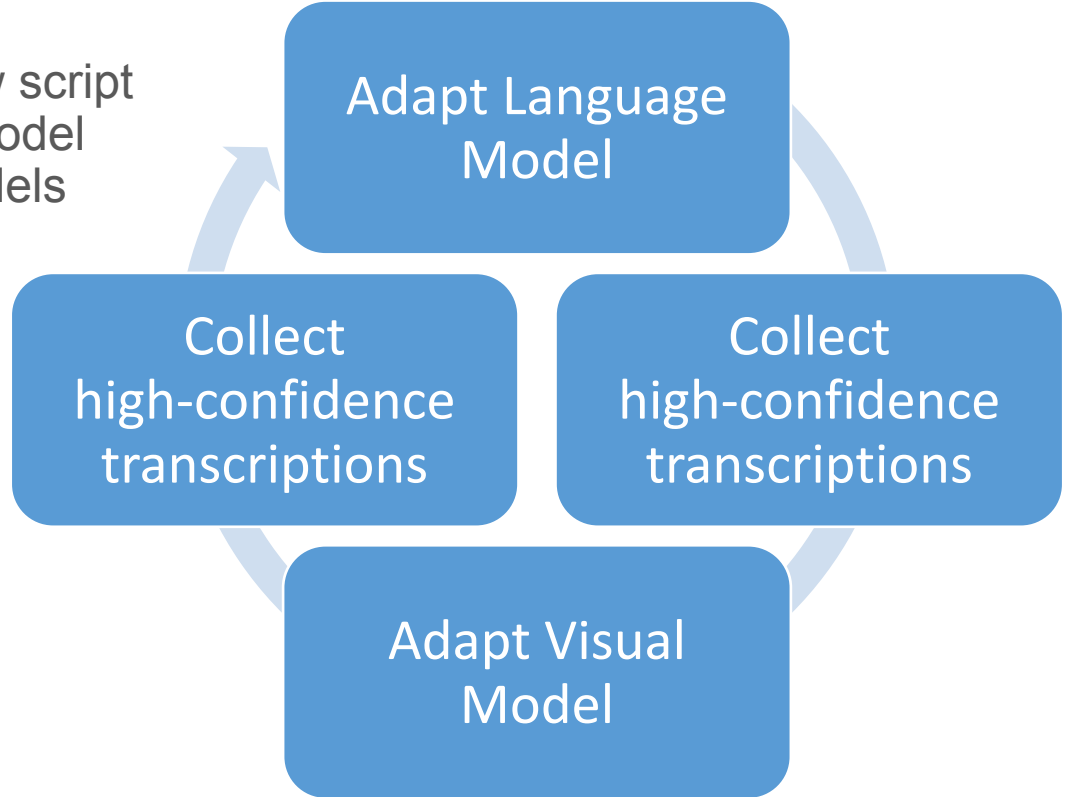
People are needed

f horrors and desolation.	29s to 31s,		
f horrors and desolation.	29s to 31s		
en He   gouwen woonach	ИМОТЬ СЕ		
en He   gouwen woonach	ИМОТЬ СЕ		
kse-tiende	пощавнувало,	this Evening Tuesday.	
kse-tiende	 ошавнувало	this Evening Tuesday	
niets	are the working	weet men too ve!	AMINER will
niets	are the working	weet men too ve 	AMINER will

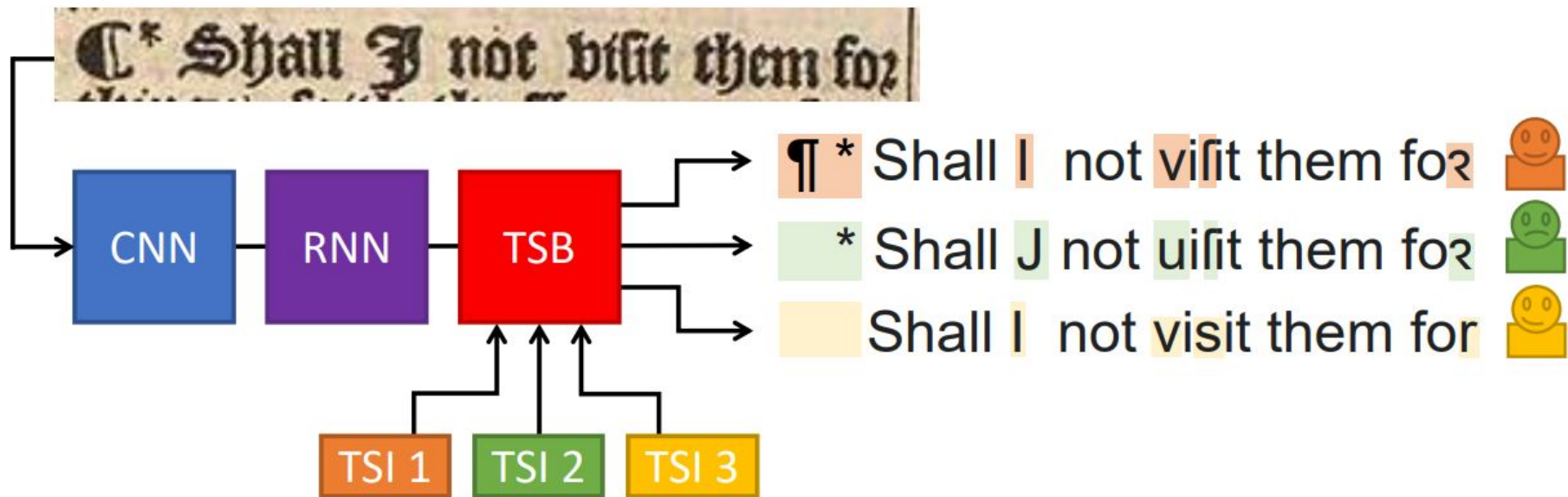


Adapting to a specific document

For transcription of a new script
Starting with a general model
Mutual adaptation of models



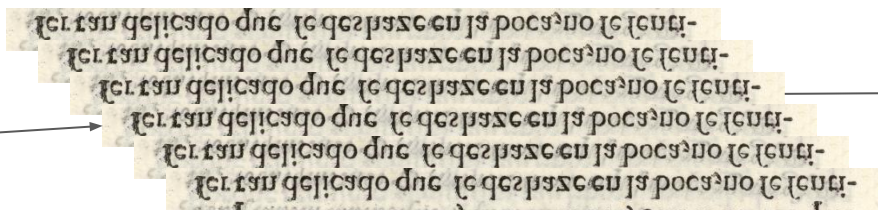
Transcription styles - transcription vs. transliteration



Transcription style selection

Manual

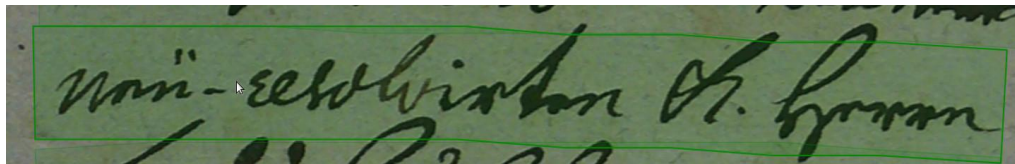
Automatic based on lines already transcribed and corrected



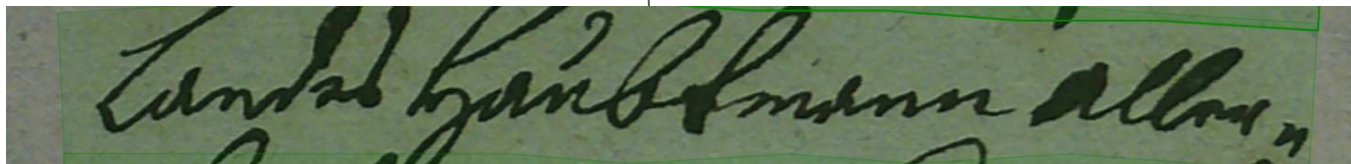
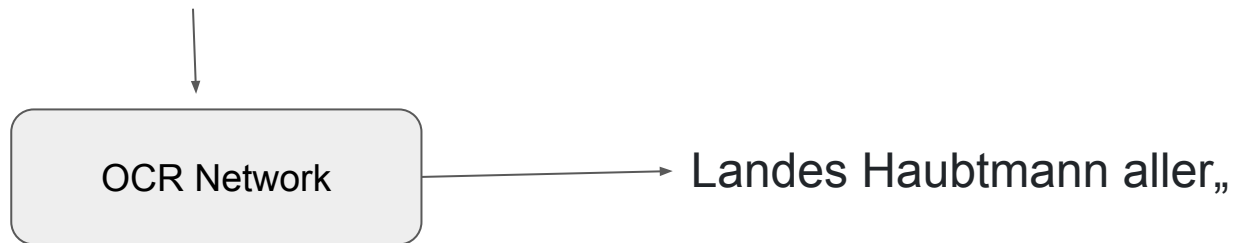
Transcription styles

Immediate transfer of transcription corrections

Corrected line:

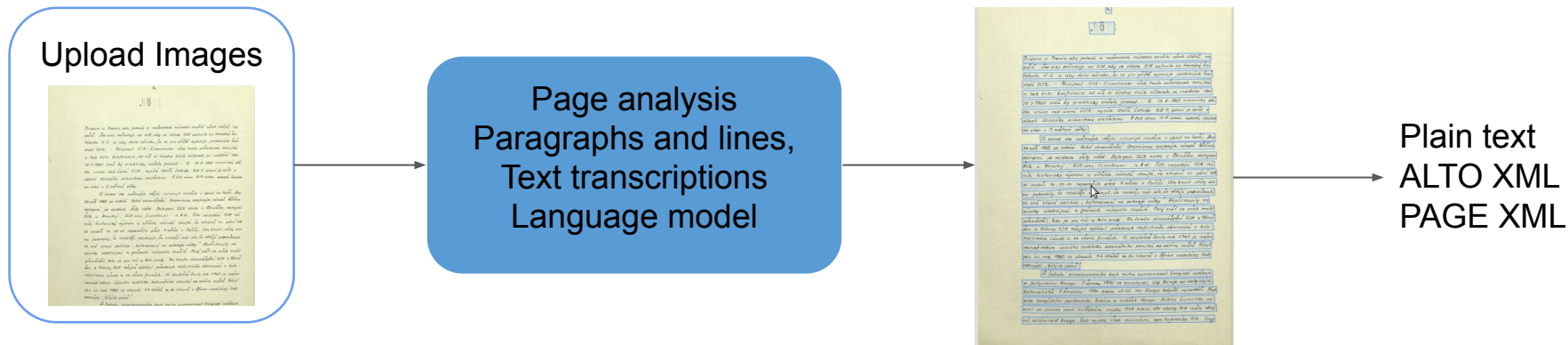


neü-resolvirten d. Herrn



API

- Batch upload page images of a whole book - wait - download
- Example python scripts - uploads directory, downloads the results
 - https://github.com/DCGM/pero-ocr-api/tree/master/user_scripts
- Documentation
 - app.swaggerhub.com/apis-docs/LachubCz/PERO-API/1.0.4



External datasets

- [IMPACT resources](#) – Older european prints
 - 1.2M lines, 9 languages, 10 scripts
 - 0.5 % error rate per character
- Deutsches Textarchiv (~6000 prints)
 - Mostly german fraktur
 - 0.3 % error rate per character
- Czech prints
 - Newspaper digitized from mikrofilms
 - Older prints
- Bentham
 - Manucrtipts of several authors - 21,403 pages
 - Necessary to align the text with manuscript lines
- Czech letters 20th century - 2000 letters

Real number of lines for training

- Medieval
 - 15.302 lines / 8.268 from pero-ocr
- Handwriting
 - 717.040 lines / 386.906 from pero-ocr
- Kurrent script
 - 208.439 lines / 17.949 from pero-ocr
- Prints
 - 1.517.028 lines / 222.575 from pero-ocr
- German Fraktur
 - 2.003.467 lines / 0 from pero-ocr
- Fraktur, including Czech
 - 2.108.993 lines / 105.526 from pero-ocr

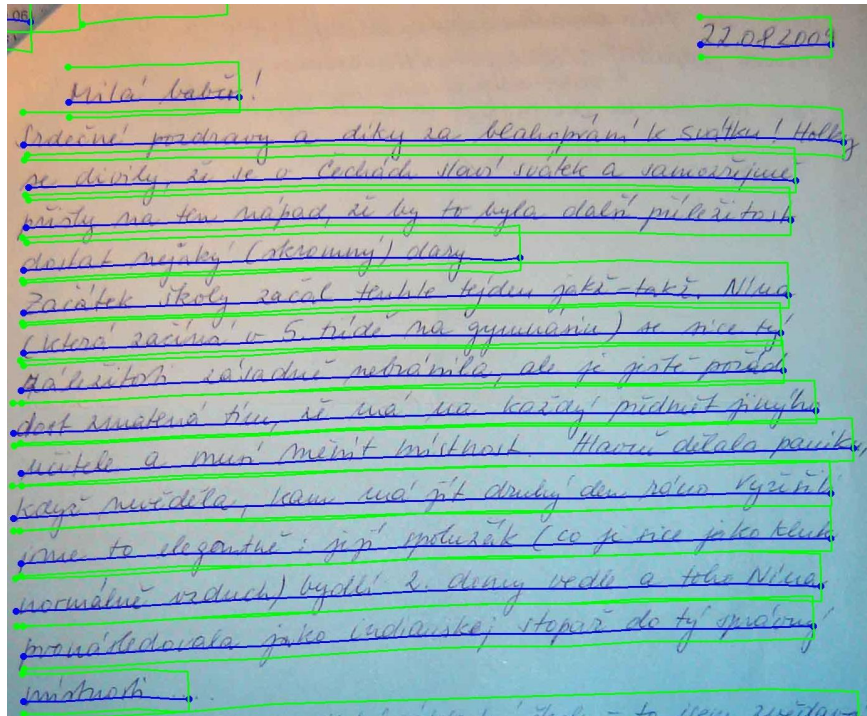
Alignment of text with lines - from any sources

22. 08. 2009

Milá babčo!

Srdečné pozdravy a díky za blahopřání k svátku! Holky se divily, že se v Čechách slaví svátek a samozřejmě přišly na ten nápad, že by to byla další příležitost dostat nějaký (skromný) dary.

Začátek školy začal tenhle týden jakž-takž. Nína (která začíná v 5. třídě na gymnáziu) se sice tý záležitosti zásadně nebránila, ale je ještě pořád dost zmatená tím, že má na každý předmět jinýho učitele a musí měnit místnost. Hlavně dělala paniku, když nevěděla, kam má jít druhý den ráno. Vyřešili jsme to elegantně: její spolužák (co je sice jako kluk normálně vzduch) bydlí 2. domy vedle a toho Nína pronásledovala jako indianskej stopař do ty správný místnosti...



Brno Handwritten Dataset

Přepsáním následujícího textu přispějete k vývoji automatického přepisu českých historických dokumentů. Naším cílem je, aby bylo v budoucnu možné online vyhledávat například v obsahu kronik, matričních a pozemkových knih i historické korespondence. Prosim, přesně přepište vytištěný text kamkoliv na tuto stránku se zachováním řádků i případných chyb. Vyplněním tohoto listu souhlasíte s jeho zveřejněním a volným využitím pro libovolné účely. Vyplněný list prosím zašlete na adresu: Michal Hradiš, Fakulta informačních technologií, VUT v Brně, Božetěchova 1/2, 612 66 Brno. Případně list naskenujte na stolním skeneru a zašlete na adresu: ihradis@fit.vutbr.cz s předmětem „Brno HWR dataset”. —

6253419

„Téhle loutně se chce hrát jenom v moll,“ usmál se na něho Arren z okna. „Chce se jí plakat. A co byste chtěli slyšet, vážení hostitelé?“

„Já jim to řeknu, Zebbie. Normální hladinu.“ Hilda počkala, dokud jí Zeb nedal signál, a pak zvolala: „Chlapci! Plukovník

Brumby potřebuje pomoc. Pojd'te si pro něj. Už vás nebudeme strašit řevem.“

„Tím vám vyhrožuje?“

13.

„Ne, pane.“

„To dělám i dnes,“ usmál se Alexijev. „Předně - tihle tu jsou déle, měli by se uzdravit dřív. A pokud jste slyšeli, co jsem říkal

6253419

„Téhle loutně se chce hrát jenom v moll,“ usmál se na něho Arren z okna. „Chce se jí plakat. A co byste chtěli slyšet, vážení hostitelé?“

„Já jim to řeknu, Zebbie

Brumby potřebuje pomoc. Pojd'te si pro něj. Už vás nebudeme strašit řevem.“

„Tím vám vyhrožuje?“

13.

„Ne, pane.“

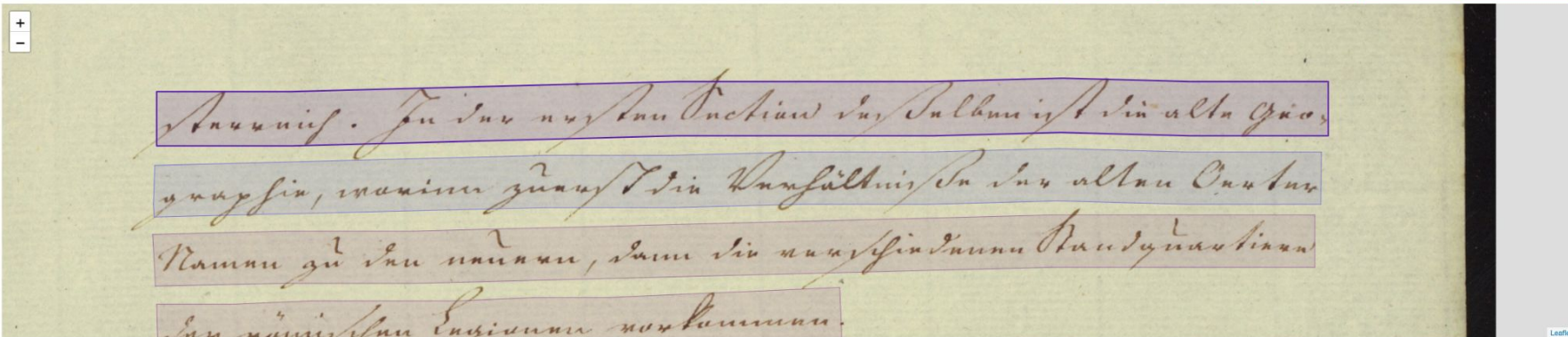
„To dělám i dnes,“ usmál se

Document name: WS21: 6. UKÁZKA MZK - KURENT



← Back Next →

Export PAGE Export ALTO Export TEXT Export IMAGE Show Suspect Lines



+
-

Leaflet

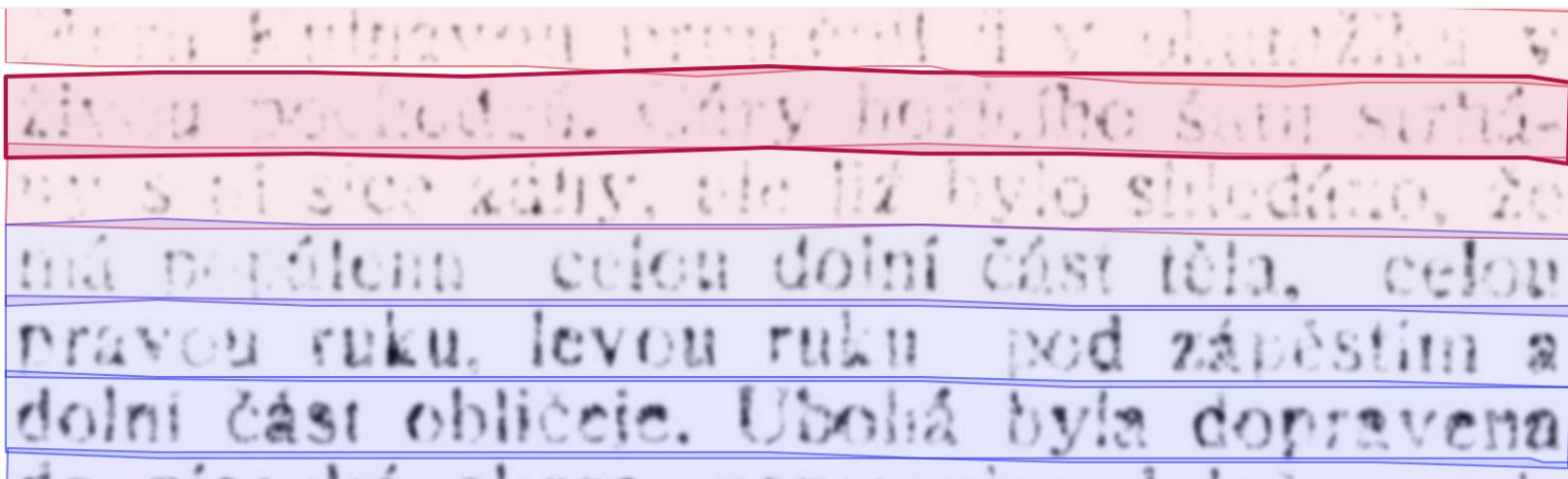
1
sterrich. In der ersten Section deßelben ist die alte Geographie, worinn zuerst die Verhältniße der alten Oerter Namen zu den neuern, dann die verschiedenen Standquartiere der römischen Legionen vorkommen.

Height 81 Pad 100

Save Delete line Ignore line Next suspect line →

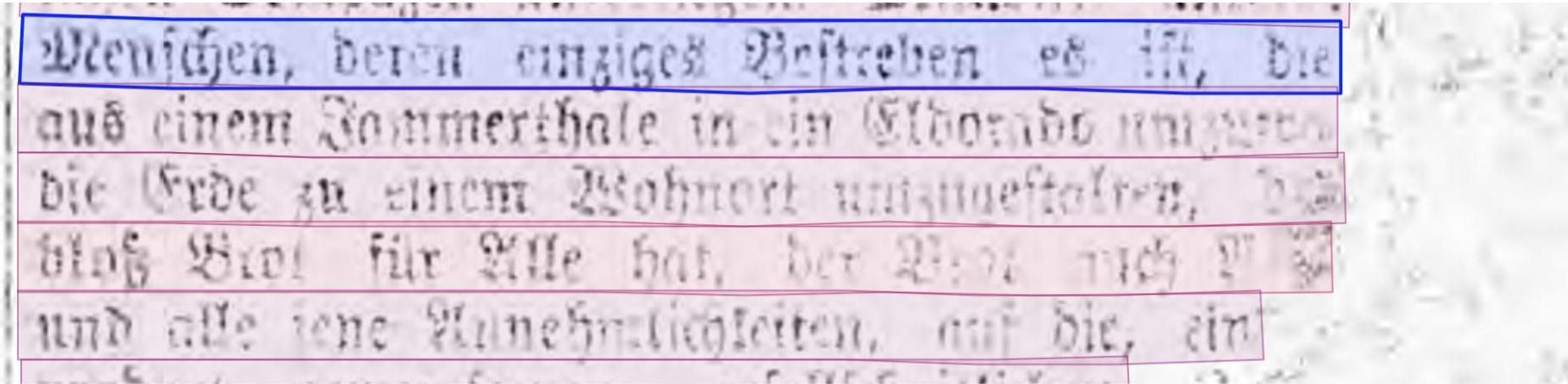


Scans of microfilmed newspaper



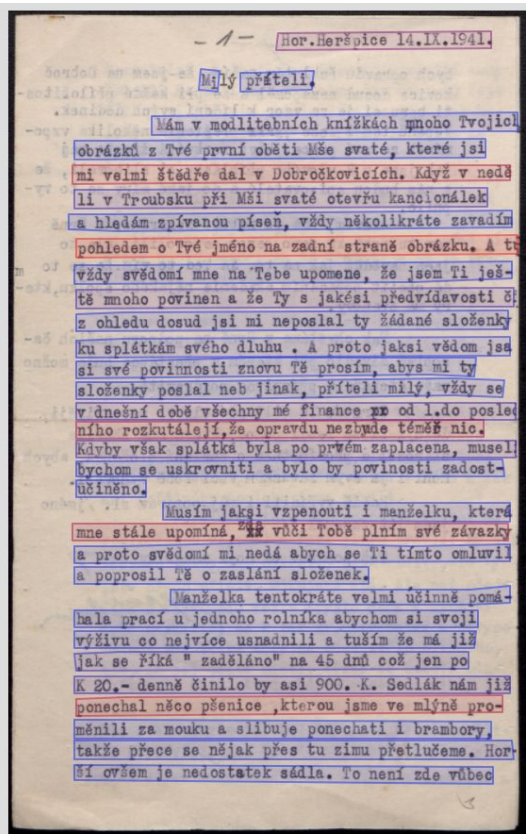
Pipra Kulnavou proměnil ji v okamžiku v živou pochodň. Čáry hořícího šatu strhány s ní sice záhy, ale již bylo shledáno, že má popálenou celou dolní část těla, celou pravou ruku, levou ruku pod zápěstím a dolní část obličeje. Ubohá byla dopravena

Scans of microfilmed newspaper - fraktur



Menschen, deren einziges Bestreben es ist, die
aus einem Jammerthale in ein Eldorado
die Erde zu einem Wohnort umzugestalten,
bins Brot für Alle hat, der Vet auch e
und alle jene Annehmlichkeiten, auf die, ein

Typewritten

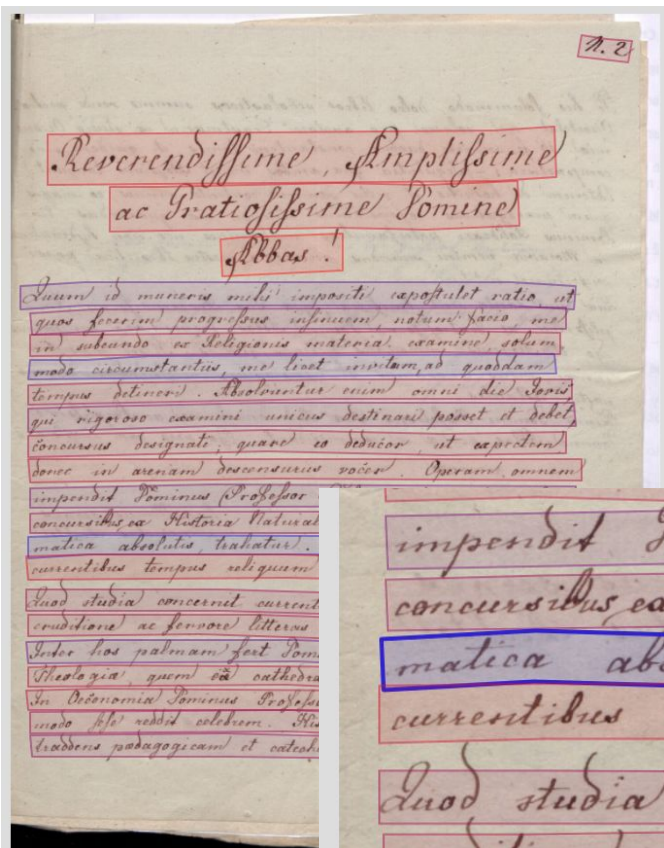


v dnešní době všechny mé finance od 1. do posled-
ního rozkutálejí, že opravdu nezbyde téměř nic.

Kdyby však splátka byla po prvním zaplácena, museli

finance xx od 1. do posled-
ního rozbyde téměř nic.
po prvním zaplácena, museli:

Manuscripts



impendit Dominus Professor Löcher ne mora amplius, concursibus ex Historia Naturali, Diplomatica, et Nemis matica absolutis, trahatur. Interea omne a studiis currentibus tempus reliquum materia revolvenda conseōro. Quod studia concernit currentia, nactus sum duces ingenio

impendit Dominus Professor Löcher ne mora amplius, concursibus ex Historia Naturali, Diplomatica, et Nemis matica absolutis, trahatur). Interea omne a studiis currentibus tempus reliquum materia revolvenda conseōro. Quod studia concernit currentia, nactus sum duces ingenio

mezi členy družstva tak potřebná při společném hospodaření.

Rovněž i v živočišné výrobě nebyla situace uspokojivá zvláště v mléčné

produkci mohlo být dosaženo daleko lepších výsledků kdyby ošetřova-

telé dojnic dbali více společného zájmu nežli svého prospěchu.

Nutno ovšem doznati že rozptýlené ustájení dojnic na ~~řadě~~ místech,

mezi členy družstva tak potřebná při společném hospodaření. ✓

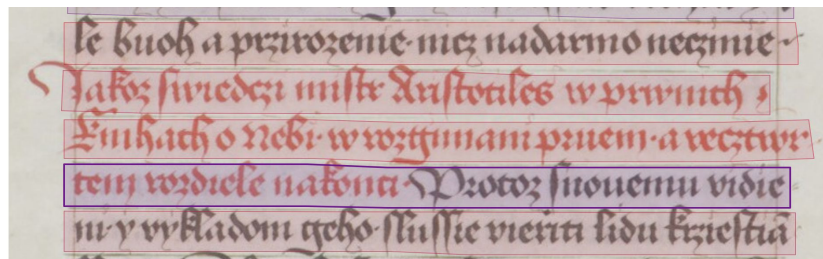
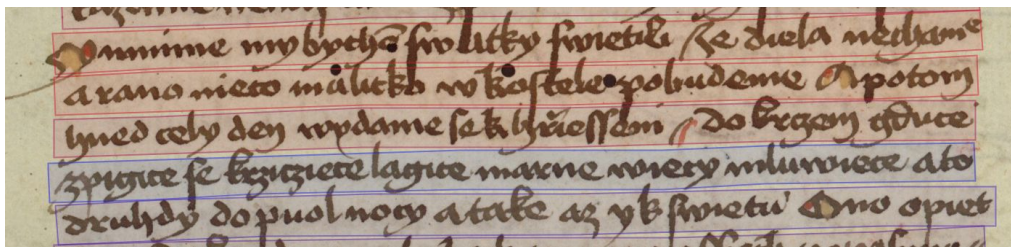
Rovněž i v živočišné výrobě nebyla situace uspokojivá zvláště v mléčné ✓

produkci mohlo být dosaženo daleko lepších výsledků kdyby ošetřova- ✓

telé dojnic dbali více společného zájmu nežli svého prospěchu. ✓

Nutno ovšem doznati že rozptýlené ustájení dojnic na ~~řadě~~ místech, ✓

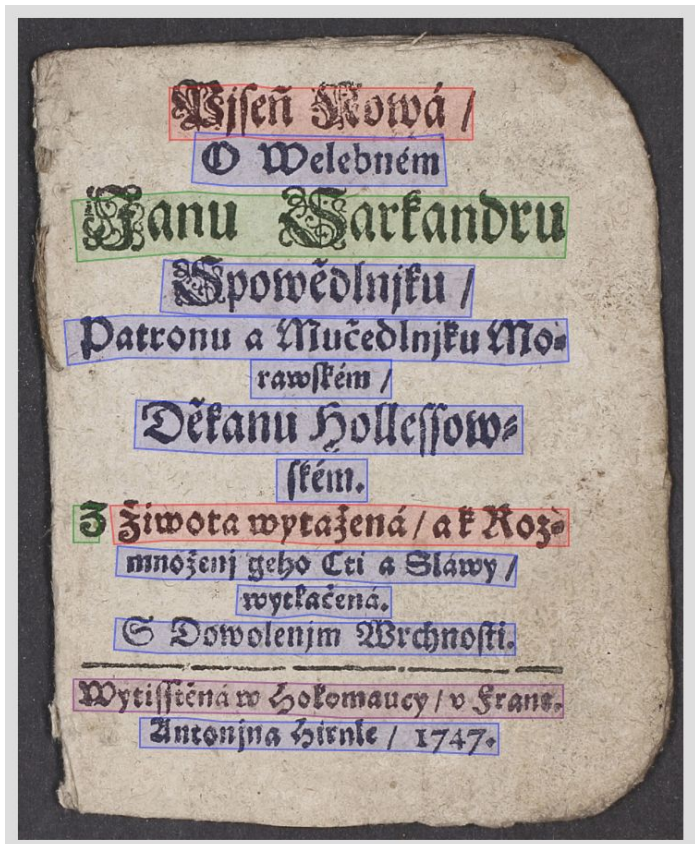
Czech medieval manuscripts



Imnime mybycha swatky fwietili se diela nedani
a rano nieto malitko w koltee pobudeme A potom
hned cely den wydame sek hriesslem Do krczem gduce
zpigice se krzicziece agite marne wiery mluwiece a to
druhdy do puo nocy a take az ys fwietu Ono opiet

le buoh a przirozenie niez nadarmo neczjie
Jako fwiedczy mistr Aristociles w prwnich
Eihach o Nebi w rozgiani pruem a vecztyr
tem rozdziele nakon Protoz snouemu vidie
ni y vykladom gehu sluffie wieriti lidu krzieftia

Broadside Ballads



Píseň **N**owá/

O Welebném

Janu Sarkandru

Spowědnjku/

Patronu a Mučedlnjku Mo=

rawském/

Děkanu Holleffow=

ském.

Žiwota wytažená/ a k Roz=

množenj geho Cti a Sláwy/

wytláčená.

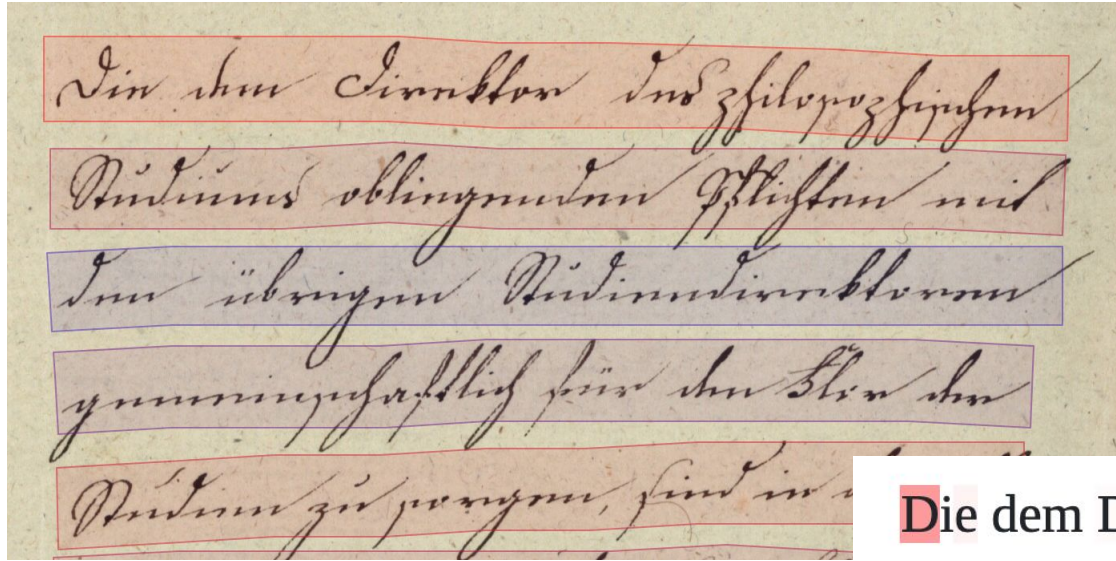
S Dowolenjm Wrchnosti.

Z

Wytisštěná w Holomaucy/ v Frane.

Antonjna Hirnle/ 1747.

Kurrent Script (German)



Die dem Direktor, des philosophischen
Studiums obliegenden Pflichten mit
den übrigen Studiendirektoren
gemeinschaftlich für den Fler der
Studien zu sorgen, sind in der all,

Kurrent Script (Czech)

Přjtomné Sirowiny čili sprosté zpě,
wy pro nassi chasu, neywjce toliko
přjležitně a w rozličných dobách
bez welikého duffewnjho napjmáej
a bez nepowědomé tehďáž mi čapo =

Přjtomné Sirowiny čili sprosté zpě,,
wy pro nassi chasu, neywjce toliko
přjležitně a w rozličných dobách
bez welikého duffewnjho napjmáej
a bez nepowědomé tehďáž mi čapo =

Cyrillic

Ці останні аж до нової доби, отже до появи капіталізму, оскільки репрезентувалися політично-пасивними громадськими шарами (селянством і робітництвом), національно перебували в історичній, так би мовити, летаргії. Вищі їхні верстви денационалізувалися, опинившись в лабетах чужої державности. З нею, поскільки держава згодом націоналізувалася, вони захоплювалися механічно чужою

Ці останні аж до нової доби, отже до появи капіталізму, оскільки репрезентувалися політично-пасивними громадськими шарами (селянством і робітництвом), національно перебували в історичній, так би мовити, летаргії. Вищі їхні верстви денационалізувалися, опинившись в лабетах чужої державности. З нею, поскільки держава

PERO

- CORE OCR - pero-ocr python package <https://github.com/DCGM/pero-ocr>
 - Python, PyTorch (running on both CPU/GPU)
- Web application for manual corrections - pero_ocr_web
 - Hosted for free at pero-ocr.fit.vutbr.cz
 - Source codes https://github.com/DCGM/pero_ocr_web
 - 250 users, 2.000 document, 100.000 pages, 700.000 corrected text lines
- OCR API for large volume document processing
 - Hosted at <https://pero-ocr.fit.vutbr.cz/api> (free for testing)
 - Source codes <https://github.com/DCGM/pero-ocr-api>
 - 160.000 processed pages
- Project info - <https://pero.fit.vutbr.cz/>