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

The PERO OCR project

Michal Hradiš
Brno University of Technology

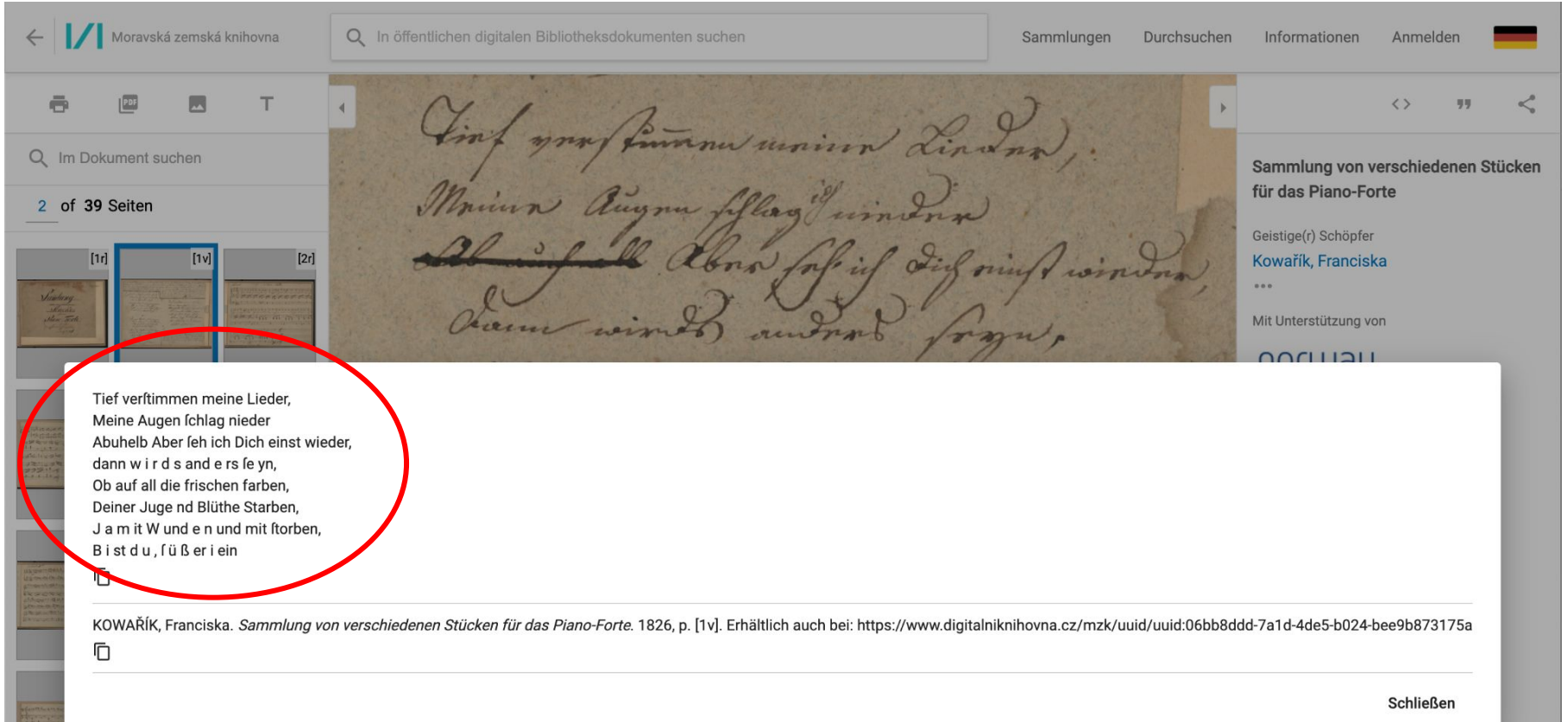
Petr Žabička
Moravian Library in Brno

Digital libraries in the Czech Republic

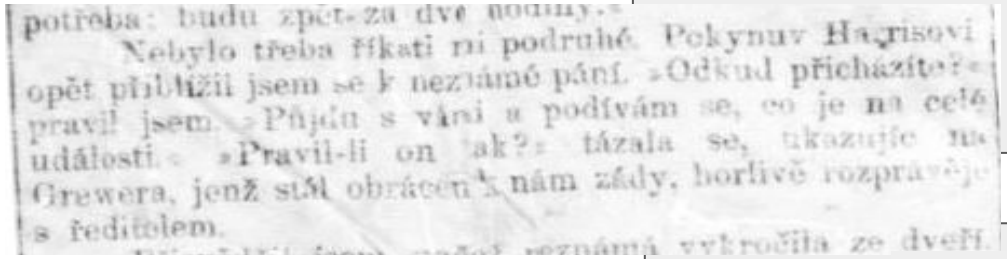
- Kramerius digital library system (open source)
  - based on Solr, jpeg2000, MODS
  - OCR:ALTO, txt
  - Open API, IIIF gateway available
- >40 libraries, >300 mil. unique pages
  - registr.digitalniknihovna.cz
  - www.digitalniknihovna.cz
- Common index: Czech digital library (in progress)

Goal: Make the content of digitized documents searchable

Goal: Make the content of digitized documents searchable

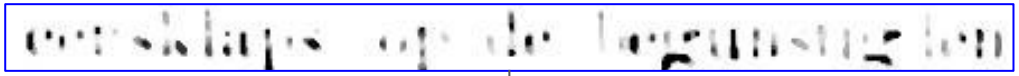


OCR Quality

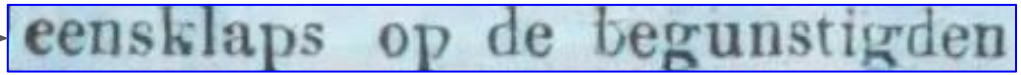

potřeba: budu zpět za dvě minuty.«
Nebylo třeba říkati ni podruhé. Pokynuv Hagrisovi
opět přiblížil jsem se k neznámé pání. »Odkud přicházíte?«
pravil jsem. »Půjdu s váni a podívám se, co je na celé
události.« »Pravil-li on ak?« tázala se, ukazujíc na
Grewera, jenž stál obrácenk nám zády, horlivě rozpravěje
s ředitelem.
vykročila ze dveří

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 Hagrisovi opět přiblížil jsem se k neztámé pání. »Odkud přicházíte?« pravil jsem. »Půjdu s váni a podívám se, co je na celé události.« »Pravil-li on ak? tázala se, ukazujíc na Grewera, jenž stál obrácenk nám zády, horlivě rozpravěje s ředitelem.

Experiments: Image Improvement



Neural network

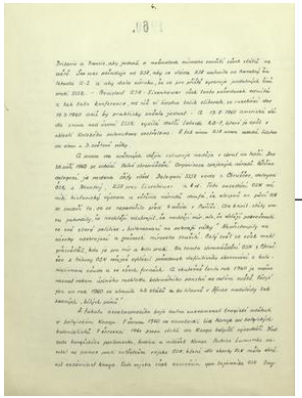


eensklaps op de begunstigden

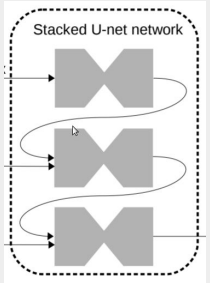
PERO - OCR

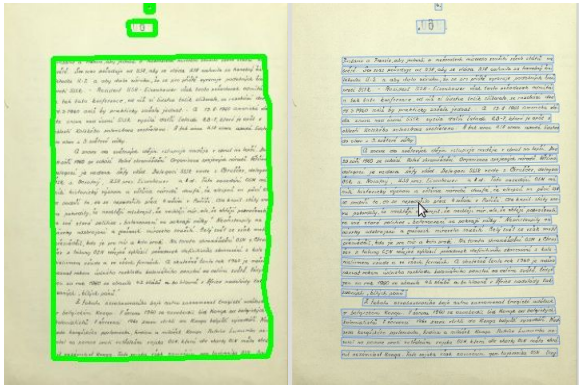
Manual transcription

PERO: Full OCR pipeline

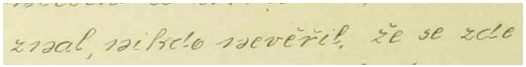


Layout





Crop line



OCR network

Language model

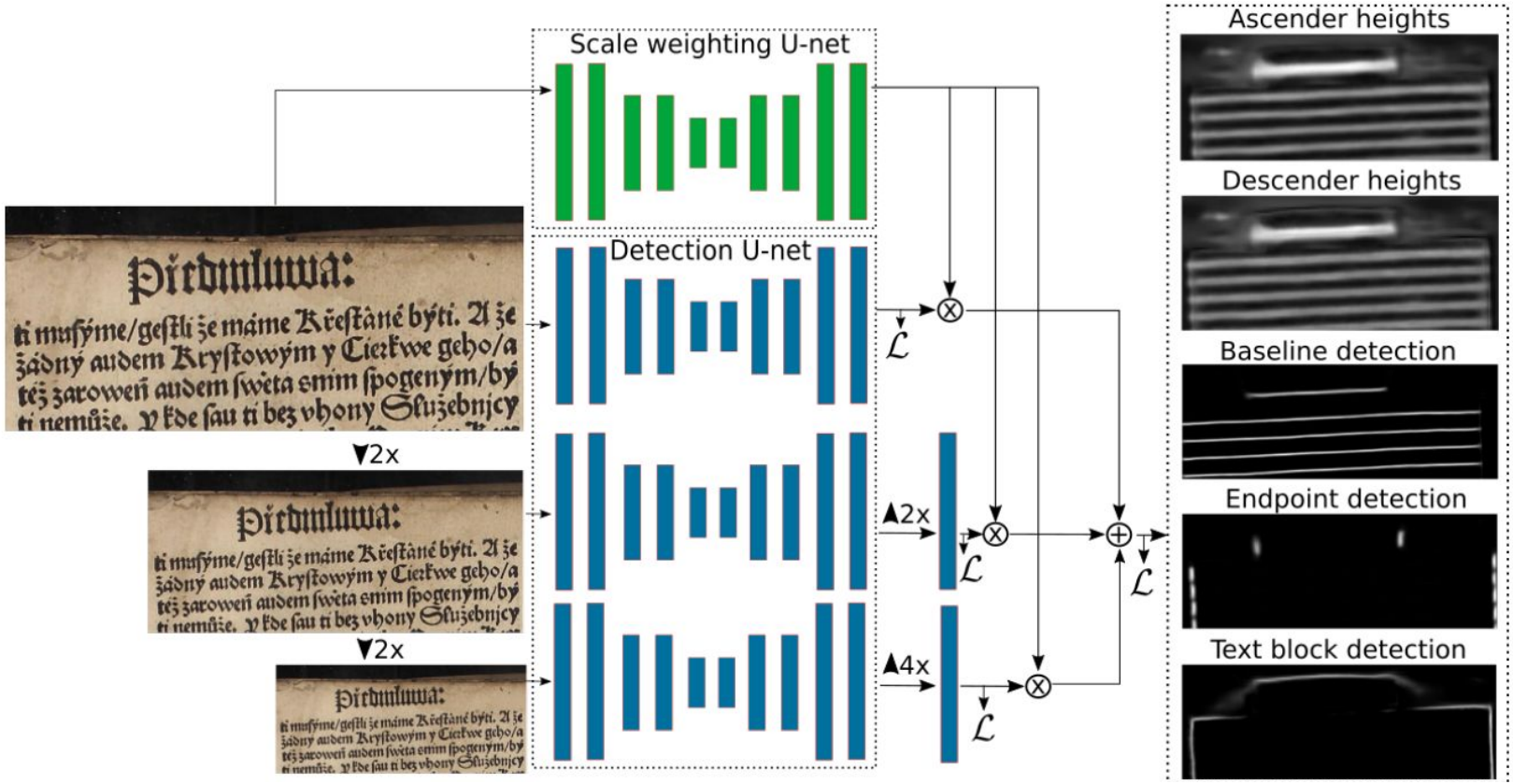
Roku 1906 se stal předsedou spolku přátel školy a podílel se na financování a výstavbě nové školní budovy v Ratibořech nad Pekárkou. Ze své funkce byl odvo-

Web application

Detection of paragraphs and text lines

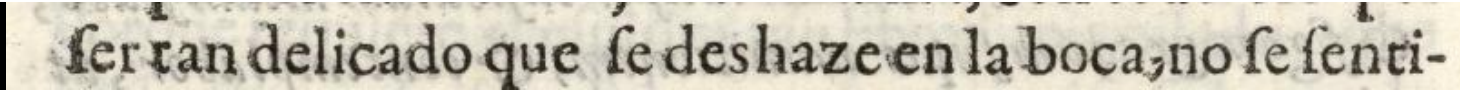
Detetion of paragraphs and text lines is not easy

Text line detection - Fully convolutional networks

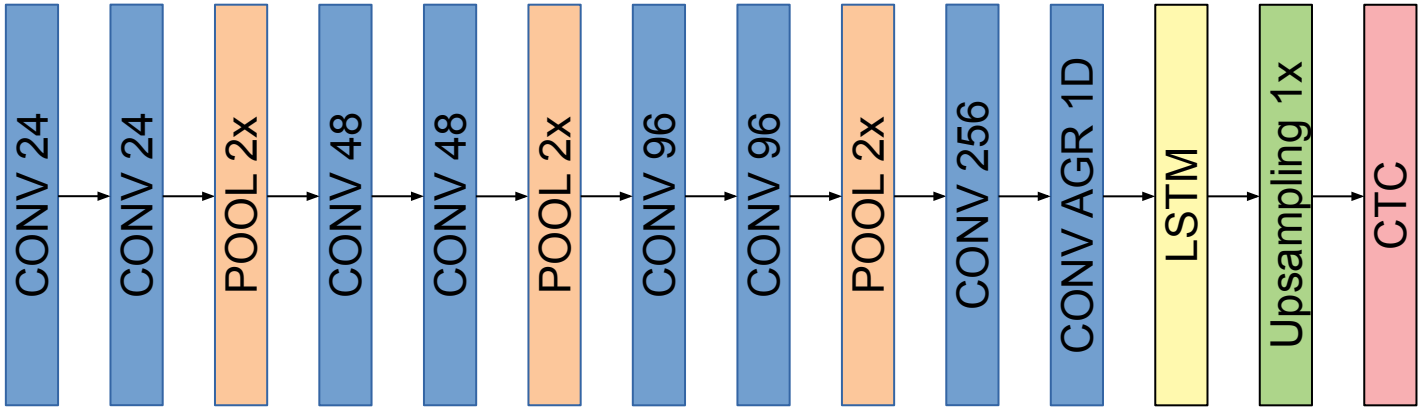


Kodym, O. and Hradiš, M. **Page Layout Analysis System for Unconstrained Historic Documents**. *International Conference on Document Analysis and Recognition (ICDAR)*, 2021.

Text transcription - convolutional/recurrent networks + CTC loss



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



Language models

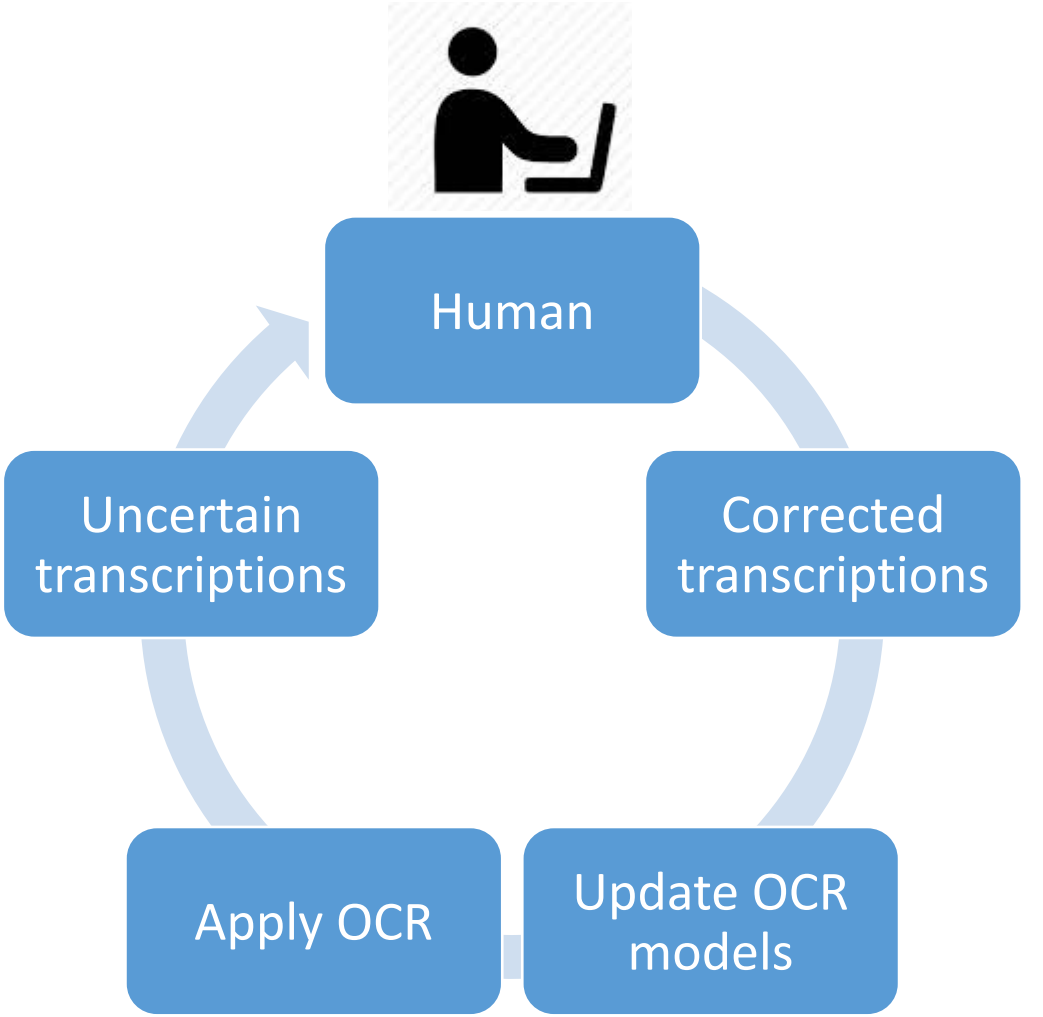
- Recurrent networks working with characters.

Similar to working with words:

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

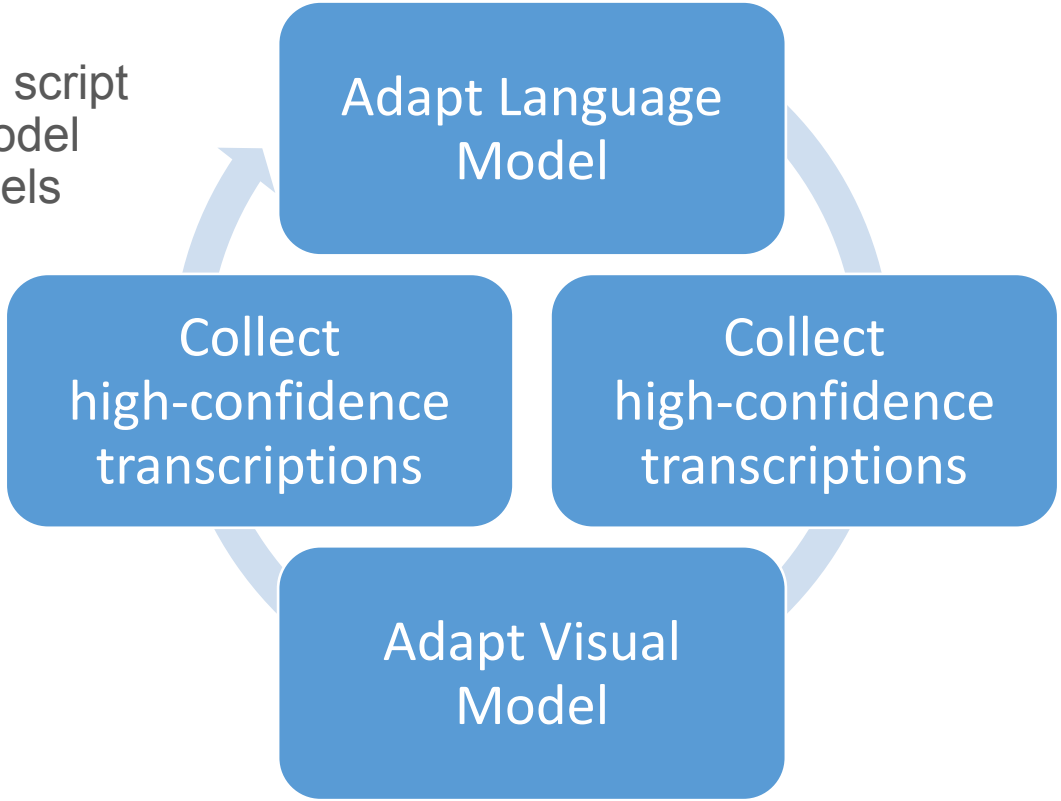
library
cinema
pubs
restaurants
….
work

People are needed

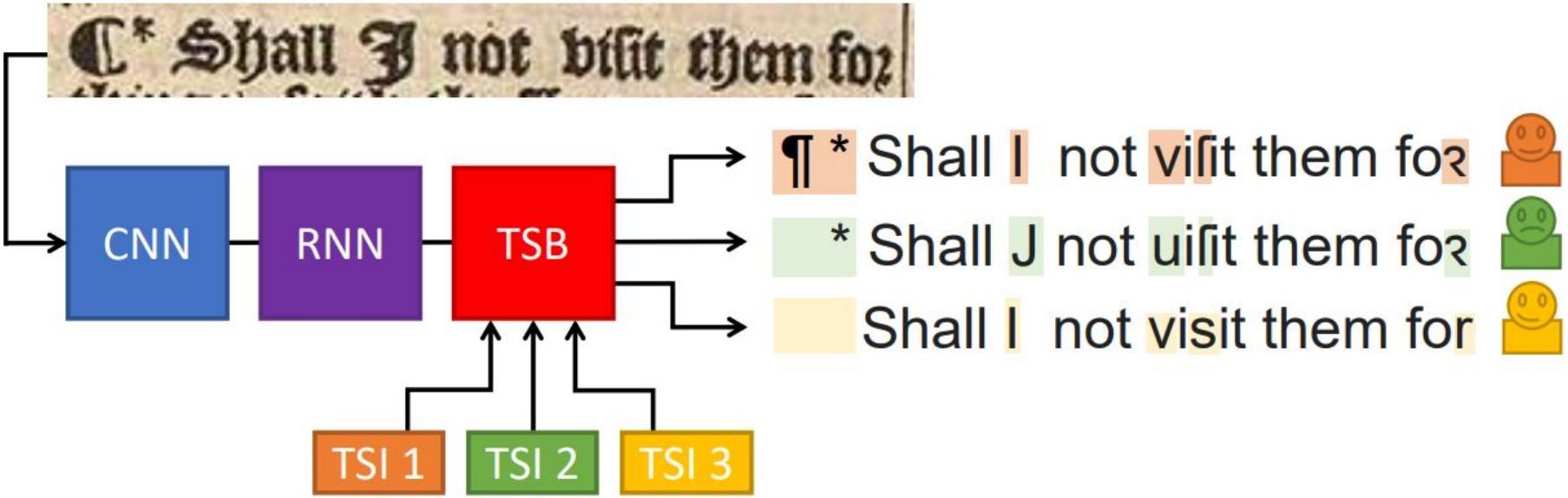


Adapting to a specific document

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



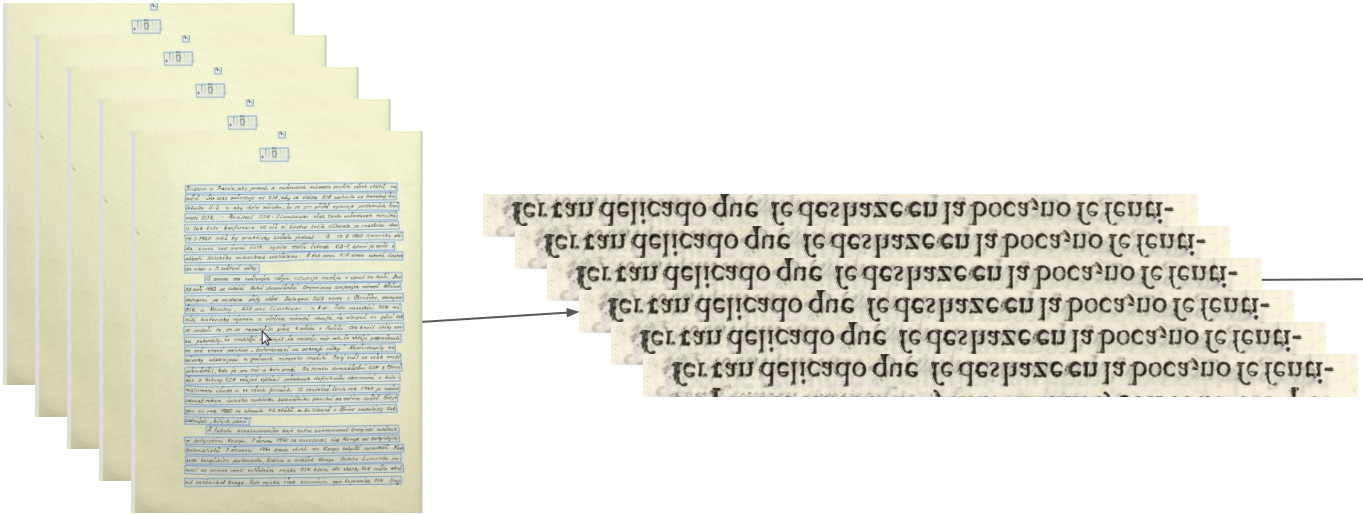
Transcription styles - transcription vs. transliteration



Transcription style selection

Manual

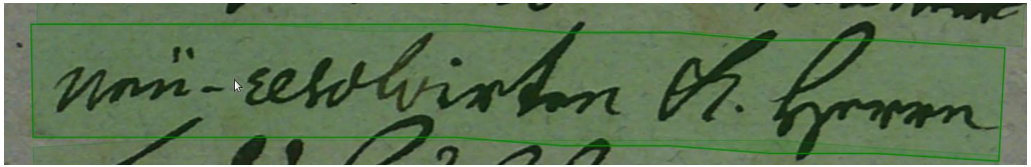
Automatic based on lines already transcribed and corrected



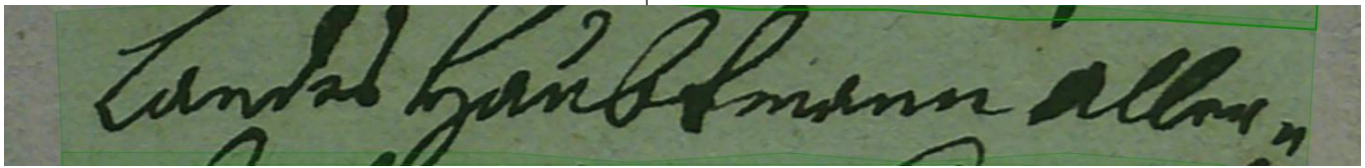
Transcription styles

Immediate transfer of transcription corrections

Corrected line:



neü-resolvirten d. Herrn

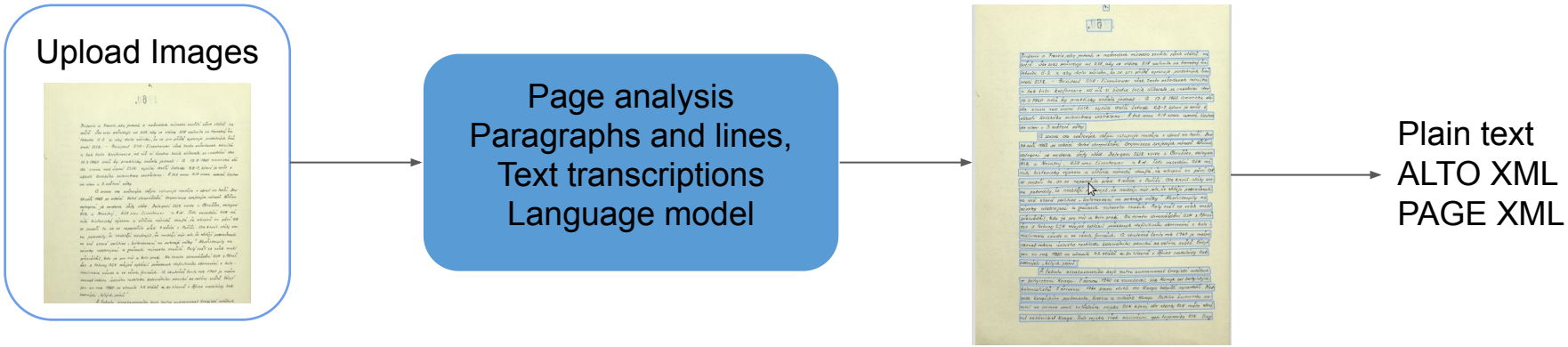


OCR Network

Landes Haubtmann aller„

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



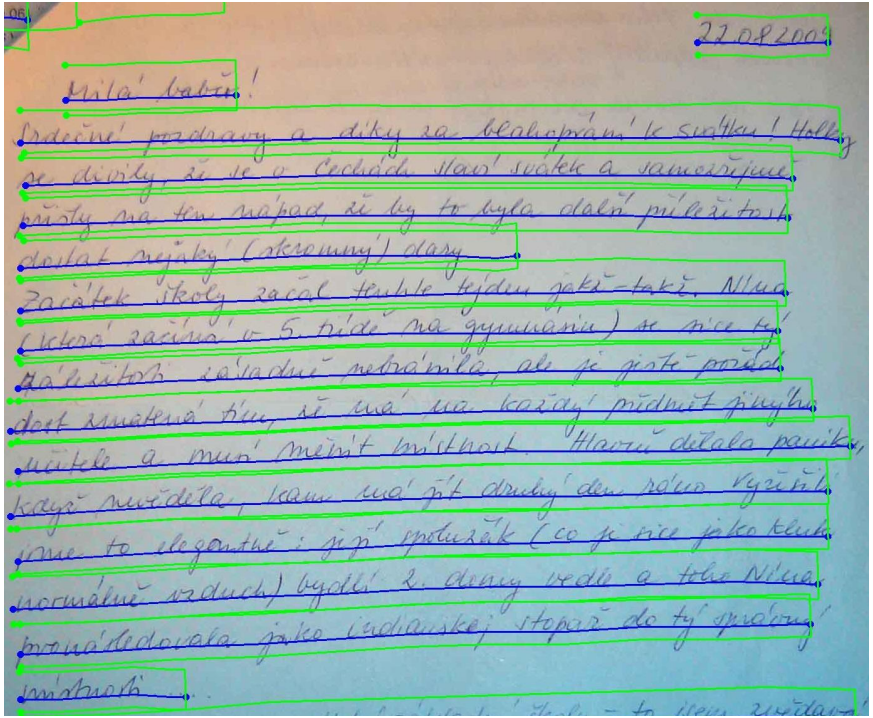
External datasets

- 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



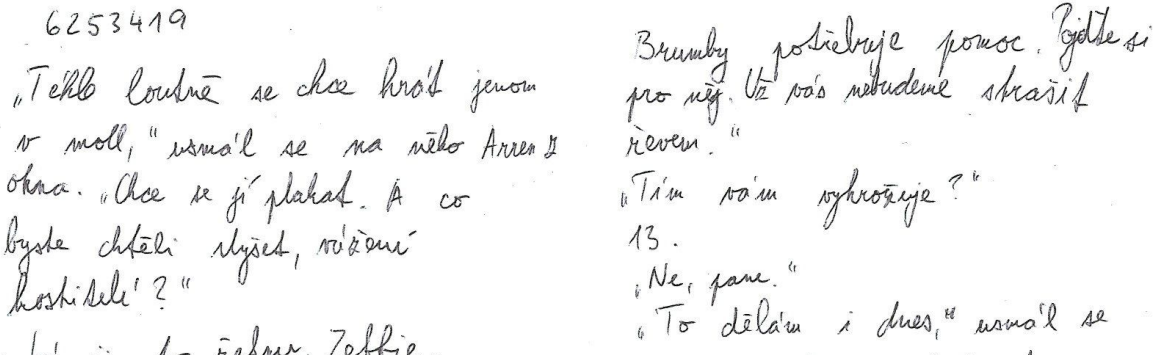
22. 08. 2009
Milá babčo!
Srdečné pozdravy a díky za blahopřání k svátku! Holky se divily, že se v Čechách slaví svátek a samozřejmě přišly na ten nápad, že by to byla další příležitost dostat nějaký (skromný) dary.
Začátek školy začal tenhle tejden jakž-takž. Nína (která začíná v 5. třídě na gymnáziu) 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 tý správný místnosti...

Brno Handwritten Dataset

Přepsáním následujícího textu přispějete k vývoji automatického přepisu českých historických dokumentů. Našim cílem je, aby bylo v budoucnu možné online vyhledávat například v obsahu kronik, matričních a pozemkových knih i historické korespondence. Prosím, 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. Pojďte 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



Document name: WS21: 6. UKÁZKA MZK - KURENT


page=uuid 1df70927-dc31-4150-9df8-9f89501a8edb.jpg


page=uuid 978ce97b-dfe3-4f5d-812c-67e74ef1df73.jpg


page=uuid 8985e604-61f9-4cf7-803d-f998fa306aa4.jpg


page=uuid 88470469-5866-4170-951c-dd1dbe1ecc52.jpg

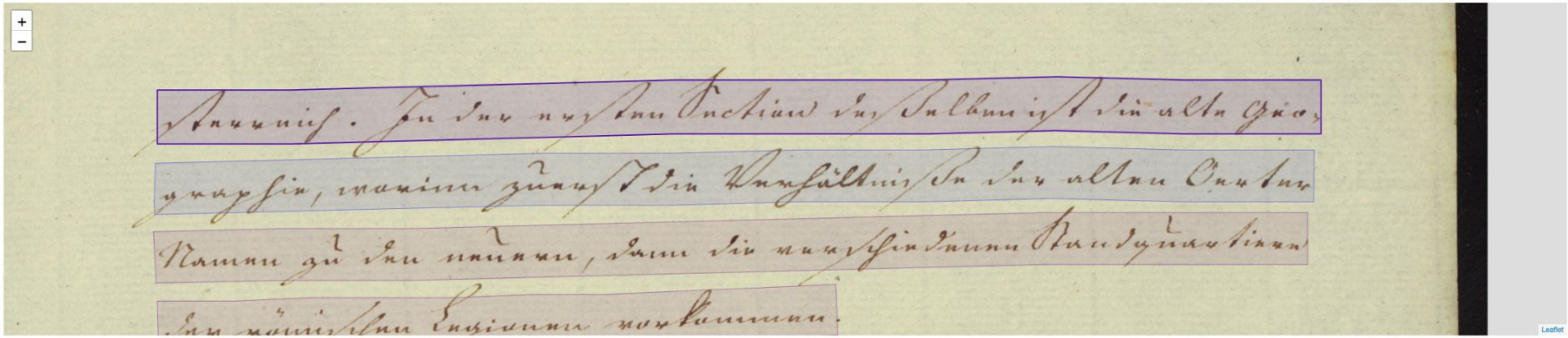

page=uuid b93b1e37-795e-4115-bc28-80dc1a881e08.jpg


page=uuid c2b3615d-366a-4c63-a7b5-67e72373d554.jpg


page=uuid c6c9e18c-9336-402d-9a80-3706f0ce2d6b.jpg

← Back Next →

Export PAGE Export ALTO Export TEXT Export IMAGE Show Suspect Lines



1

sterreich. In der ersten Section deßelben ist die alte Gro-
graphie, worinn zuerst die Verhältniße der alten Oerter
Namen zu den neuern, dann die verschiedenen Standquartiere
der römischen Legionen vorkommen.

Height 81 Pad 100

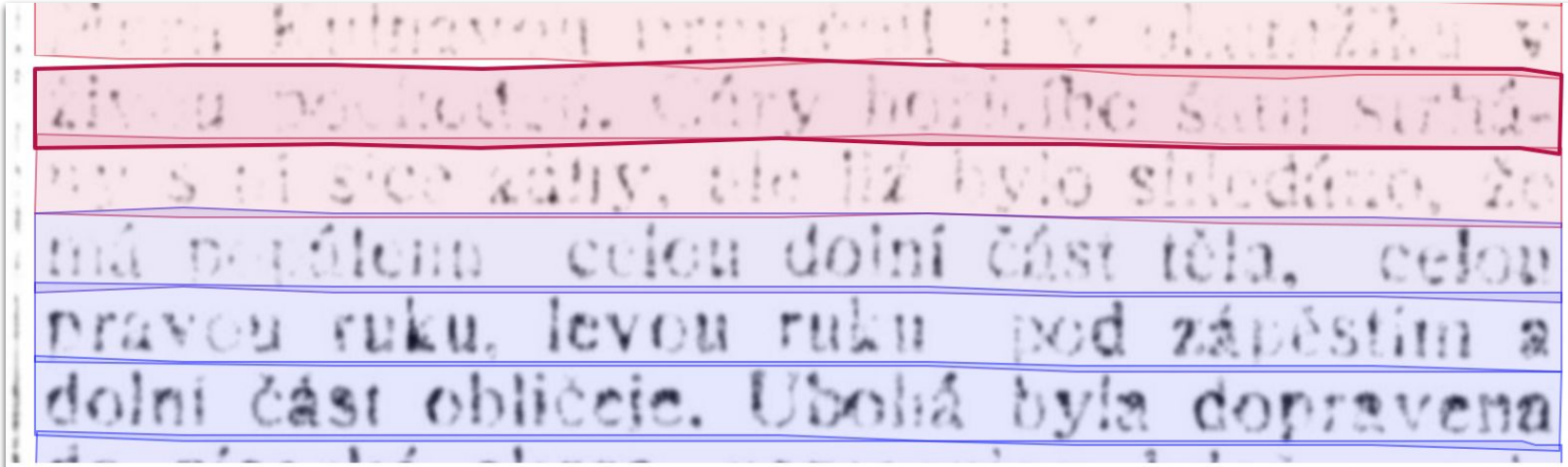
Save Delete line Ignore line Next suspect line →

MORAVSKÁ ZEMSKÁ KNIHOVNA

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

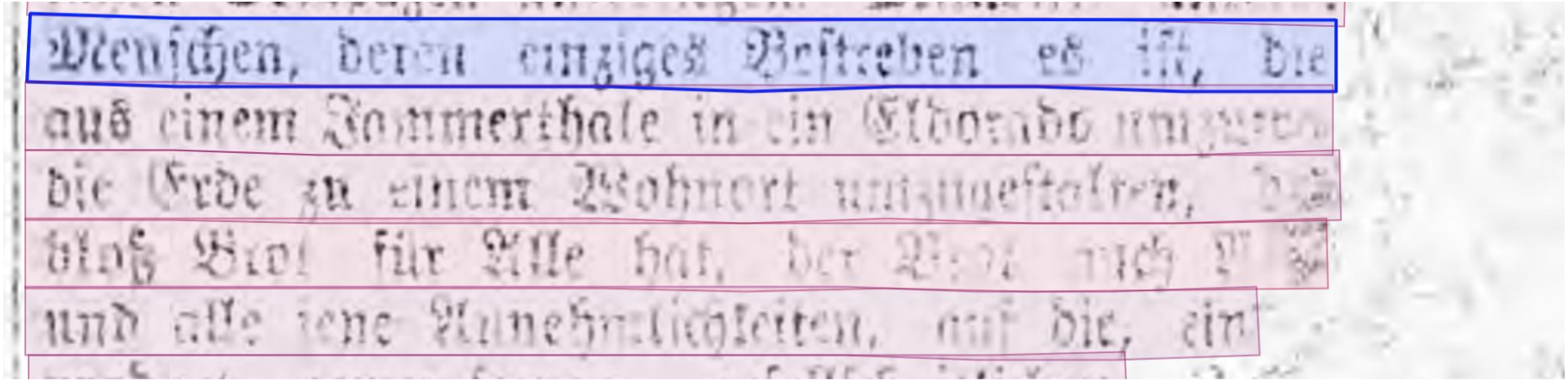
MINISTERSTVO KULTURY

Scans of microfilmed newspaper



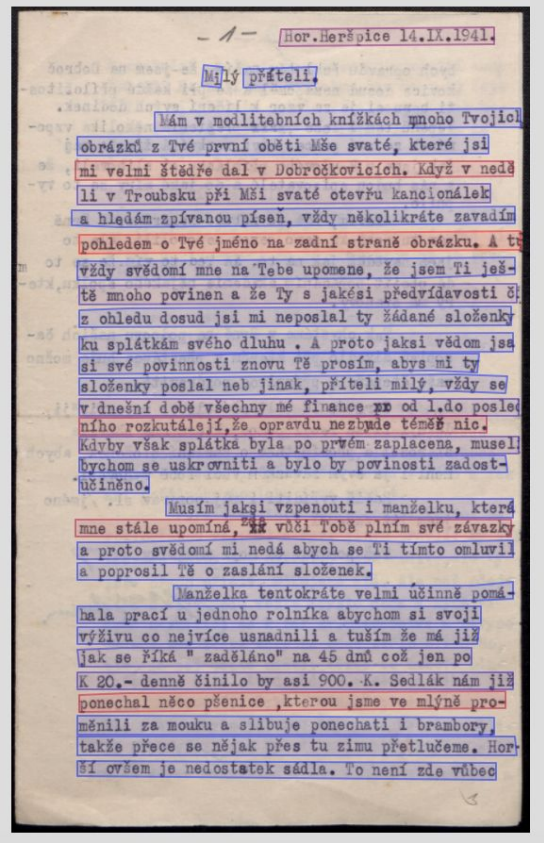
Pipra Kulnavou proměnil ji v okamžiku v
živou pochodeň. Čáry hořícího šatu strhá-
ny s ní sice záhy, ale již bylo shledáno, že
má popálenu celou dolní část těla, celou
pravou ruku, levou ruku pod zápěstím a

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 Annehmlschkeiten, auf die, ein

Typewritten



– 1 – Hor.Heršpice 14.IX.1941.

Milý přátelí,

Mám v modlitebních knížkách mnoho Tvojich obrázků z Tvé první oběti Mše svaté, které jsi mi velmi štědře dal v Dobročkovicích. Když v neděli v Troubsku při Mši svaté otevřu kancionálek a hledám zpívanou píseň, vždy několikráte zavadím pohledem o Tvé jméno na zadní straně obrázku. A ty vždy svědomí mne na Tebe upomene, že jsem Ti ještě mnoho povinen a že Ty s jakési předvídavosti ohledu dosud jsi mi neposlal ty žádané složenky ku splátkám svého dluhu . A proto jaksi vědom jsi své povinnosti znovu Tě prosím, abys mi ty složenky poslal neb jinak, příteli milý, vždy se v dnešní době všechny mé finance xx od 1.do posled ního rozkutálejí,že opravdu nezbude téměř nic. Kdyby však splátka byla po prvém zaplacena, museli bychom se uskrovniti a bylo by povinosti zadostučiněno.

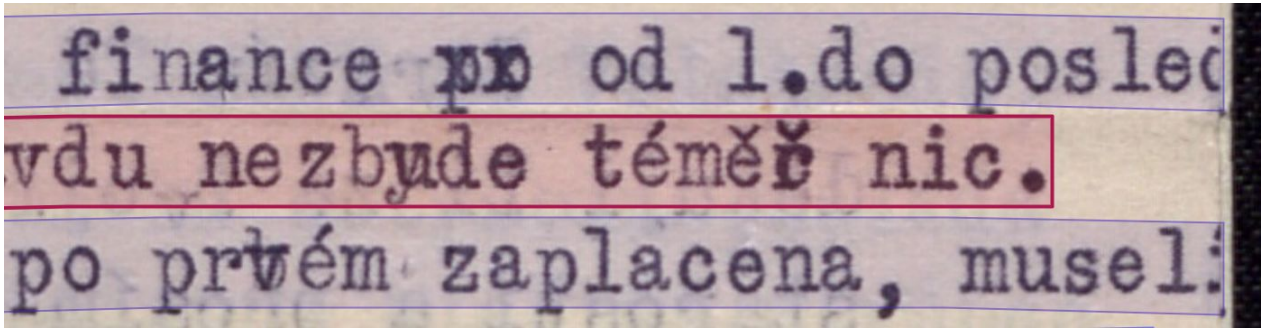
Musím jaksi vzpenouti i manželku, která mne stále upomíná, že vůči Tobě plním své závazky a proto svědomí mi nedá abych se Ti tímto omluvil a poprosil Tě o zaslání složenek.

Manželka tentokráte velmi účinně pomáhala prací u jednoho rolníka abychom si svoji výživu co nejvíce usnadnili a tuším že má již jak se říká " zaděláno" na 45 dnů což jen po K 20.- denně činilo by asi 900.-K. Sedlák nám již ponechal něco pšenice ,kterou jsme ve mlýně proměnili za mouku a slibuje ponechati i brambory, takže přece se nějak přes tu zimu přetlučeme. Horší ovšem je nedostatek sádla. To není zde vůbec

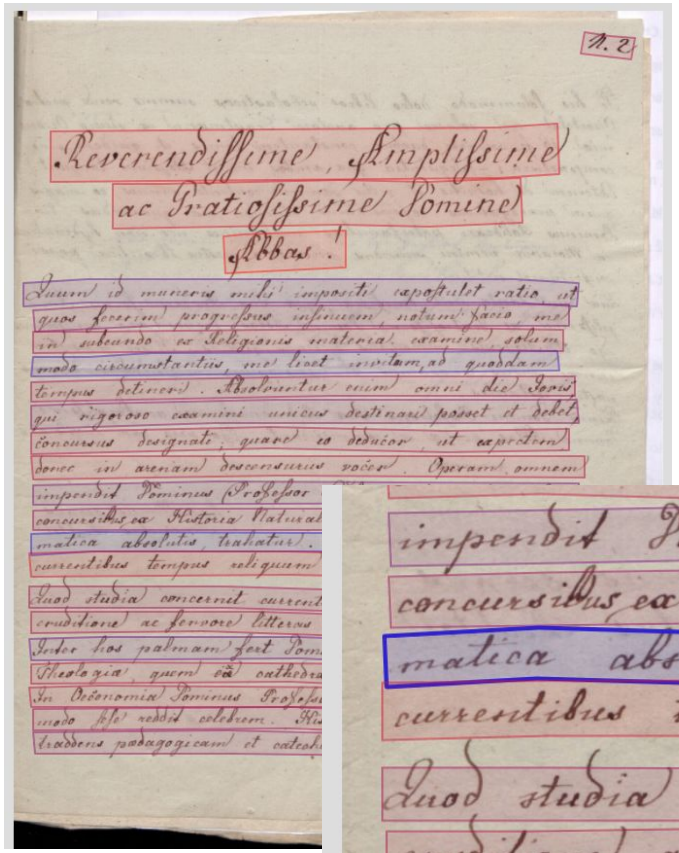
v dnešní době všechny mé finance od 1. do posled

ního rozkutálejí,že opravdu nezbude téměř nic.

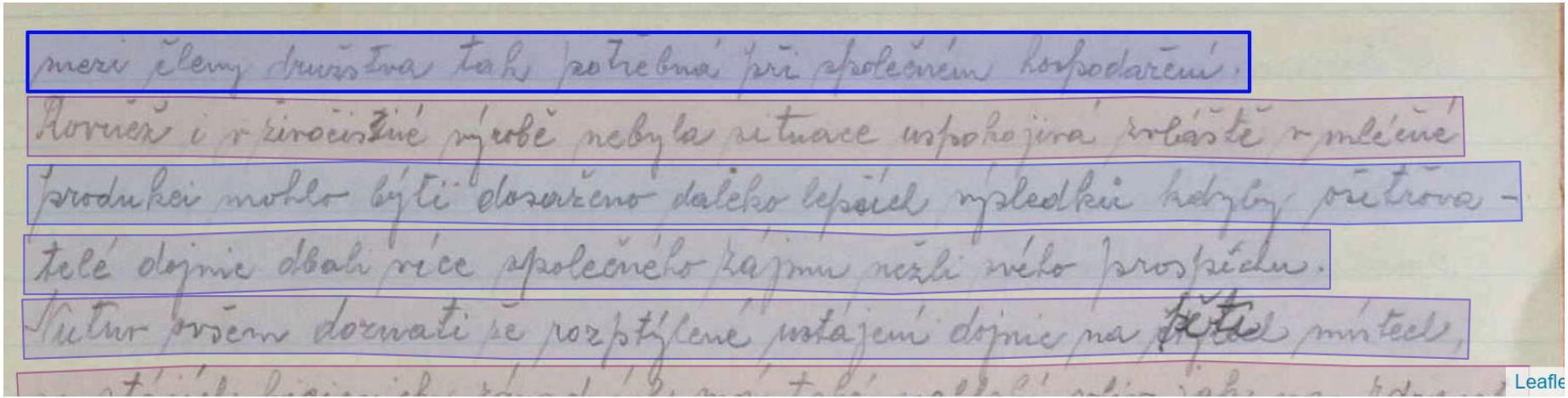
Kdyby však splátka byla po prvém zaplacena, museli



Manuscripts



impendit Pominus Proſeſsor Löcher ne mora amplius,
concursilus ex Historia Naturali, Diplomatica, et Nemis,
matica absolutis, trahatur Interea omne a studis
currentibus tempus reliquum materia revolvenda conseoro.
Quod, studia concernit currentia, nactus sum dučès ingenio



mezi členy družstva tak potřebná při společném hospodaření.

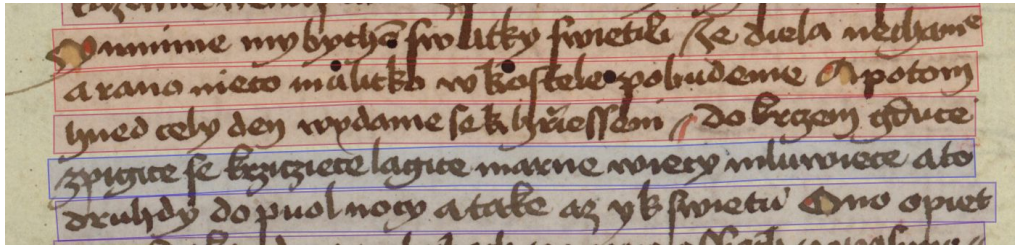
Rovněž i v živočišné výrobě nebyla situace uspokojivá zvláště v mléčné

produkci mohlo býti dosaženo daleko lepších výsledků kdyby ošetřova-

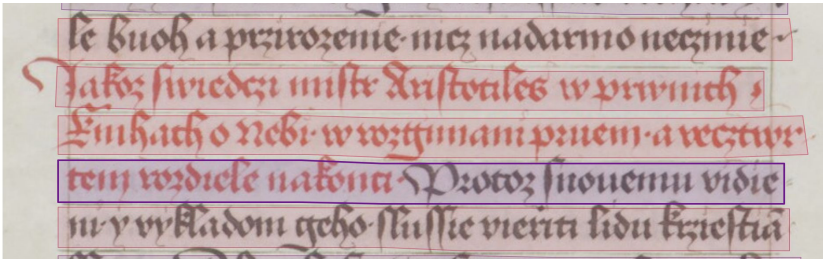
telé dojnic dbali více společného zájmu nežli svého prospěchu.

Nutno ovšem doznati že rozptýlené ustájení dojnic nach místech,

Czech medieval manuscripts

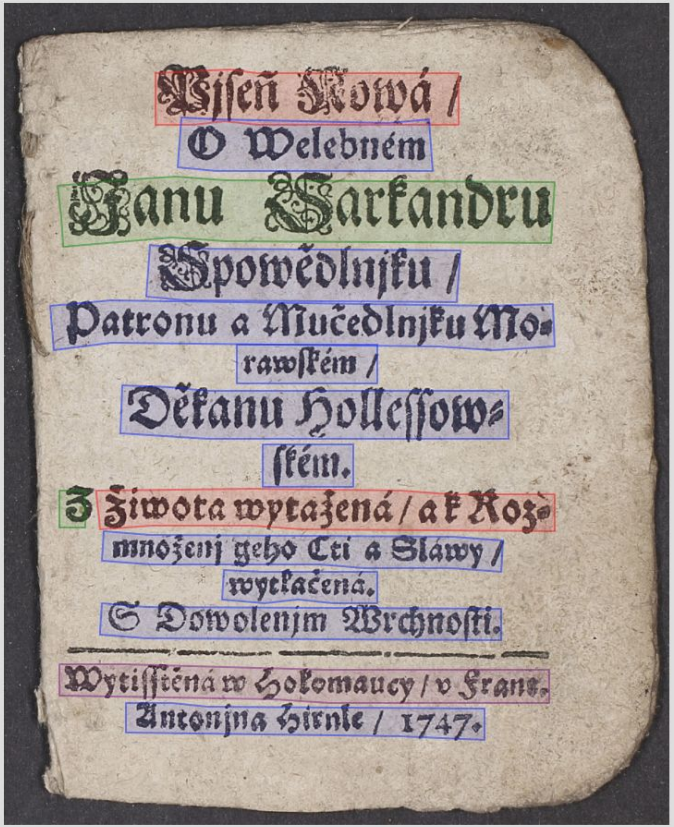


I mnime mybycha ſwatky ſwietili ſe diela nedani
a rano nieto malitko w koſtee pobudeme A potom
hned cely den wydame ſek hriieſſem Do krczem gduce
zpigice ſe krzicziece agite marne wiery mluwiece a to
druhdy do puo nocy a take az ys ſwietu Ono opiet

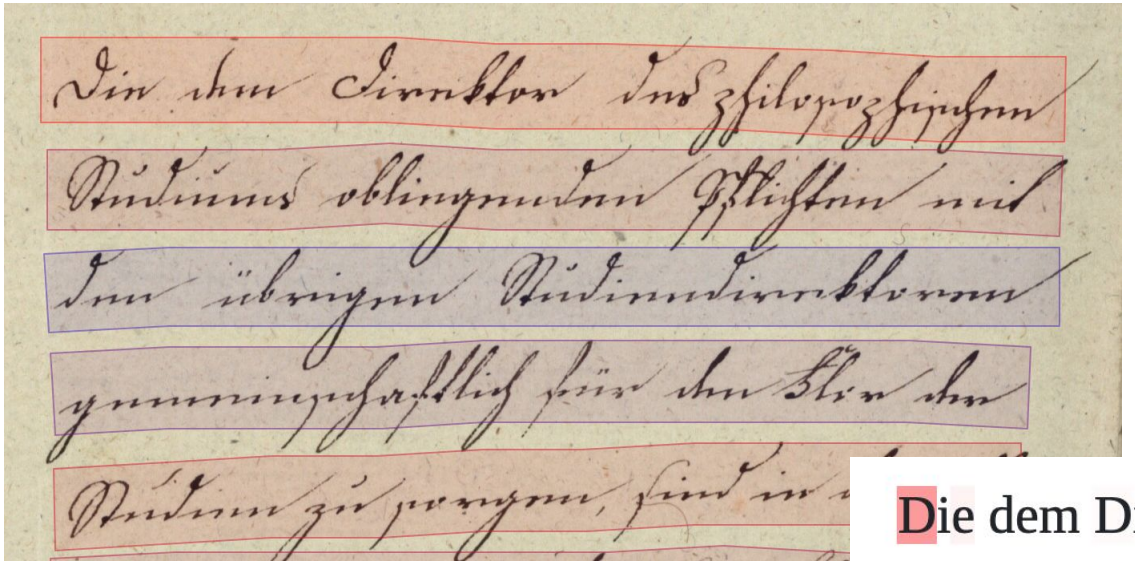


le buoh a przirozenie·niez nadarmo necziie-
Jakoz ſwiedczi miſtr Ariſtocileś w prwnich
Eiihach o Nebi w rozgiani pruem·a vecztyr
tem rozdiele nakona Protoz ſnouemu vidie
ni·y vykladom geho ſluſſie vieriti lidu krzieſtia

Broadside Ballads

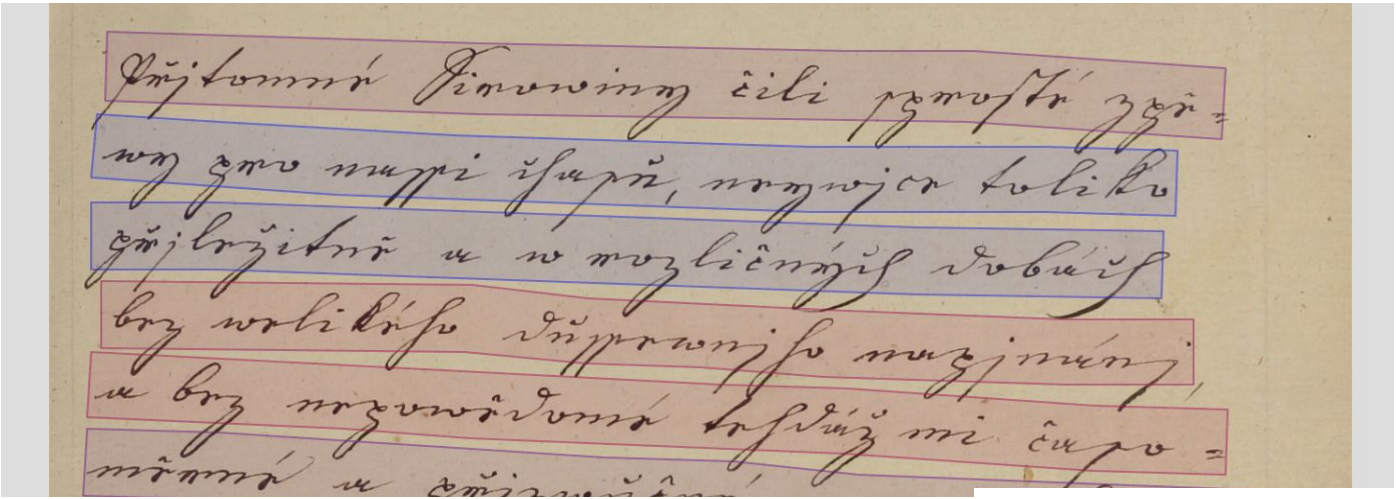

Pjſeň Nowá/
O Welebném
Janu Sarkandru
Spowědlnjku/
Patronu a Mučedlnjku Mo=
rawſkém/
Děkanu Holleſſow=
ſkém.
Z iwota wytažená/ a k Roz=
množenj geho Cti a Sláwy/
wytľačená.
S Dowolenjm Wrchnoſti.
Z
Wytiſſtěná w Hoľomaucy/ v Frane.
Antonjna Hirnle/ 1747.

Kurrent Script (German)



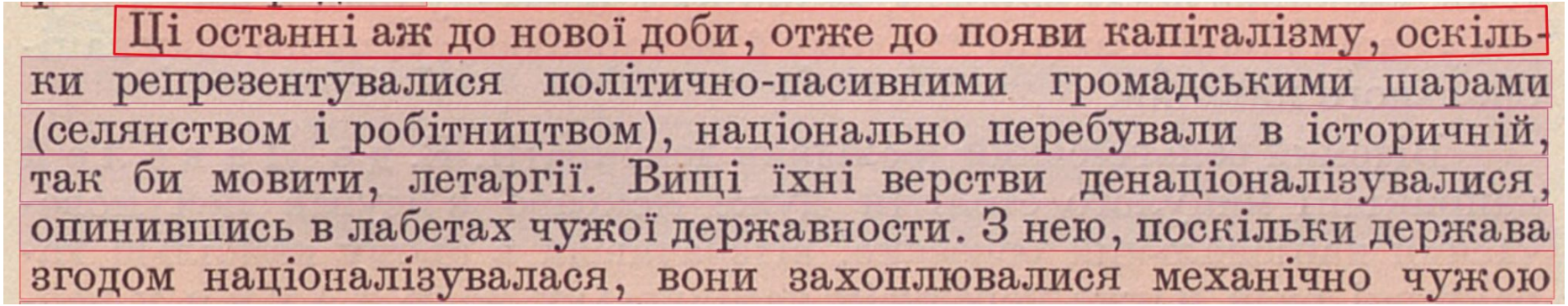
Die dem Direktor, des philoſophiſchen
Studiums obliegenden Pflichten mit
den übrigen Studiendirektoren
gemeinſchaftlich für den Fler der
Studien zu ſorgen, ſind in der all„

Kurrent Script (Czech)



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

Cyrillic



Ці останні аж до нової доби, отже до появи капіталізму, оскіль-
ки репрезентувалися політично-пасивними громадськими шарами
(селянством і робітництвом), національно перебували в історичній,
так би мовити, летаргії. Вищі їхні верстви денаціоналізувалися,
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/