OCR for Neo-Latin

Uwe Springmann

Centrum für Informations- und Sprachverarbeitung (CIS)
Ludwig-Maximilians-Universität München

Online Conference
"Digital Humanities and Neo-Latin Studies",
14-16 April 2021
https://dnls.hypotheses.org

2021-04-14

Goals of this talk

1. How you can quickly convert scanned page images of early printings into text by **optica characterum recognitio (OCR)**
   - lower the barrier to use OCR (no longer necessary to use a Linux computer)
   - show that model training can be quickly done on top of existing models
2. How to adapt existing recognition models to your needs:
   - e.g., cursive printing
   - e.g., ligature-ridden neo-ancient Greek

Example: Camerarius' Elementa Rhetoricae

Elementa Rhetoricae, sive capita exercitationum studii puerilis et stili, ad comparandum utriusque linguae facultatem

Joachimus Camerarius the elder (1500 - 1574), one of the most eminent humanists in the 16th century (see "Opera Camerarii" project)

A good OCR result depends on a good scan

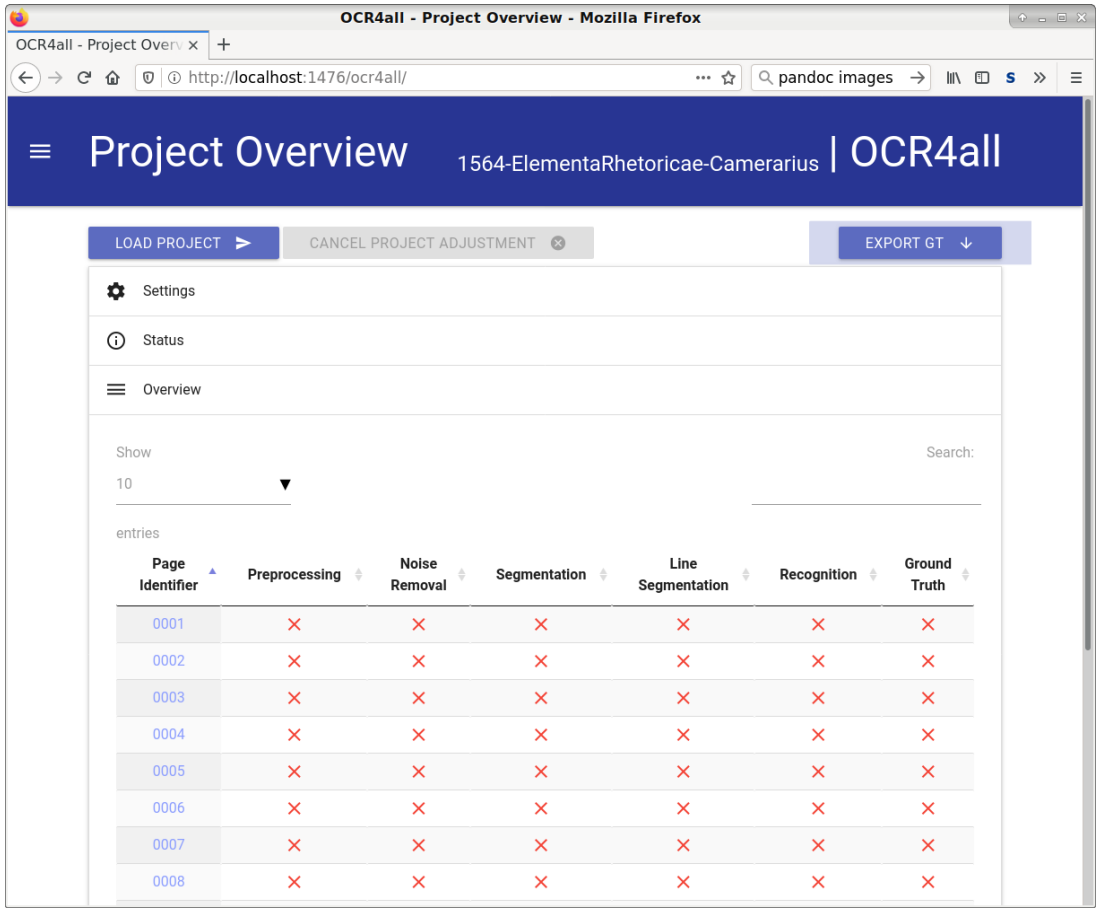
1. search Europeana for "camerarius elementa rhetoricae"
2. best scan: 1564 edition, Czech National Library
3. download images
   - in highest available resolution
   - use your favorite download manager

Preprocess with Scantailor

deskew, get rid of noisy edges:

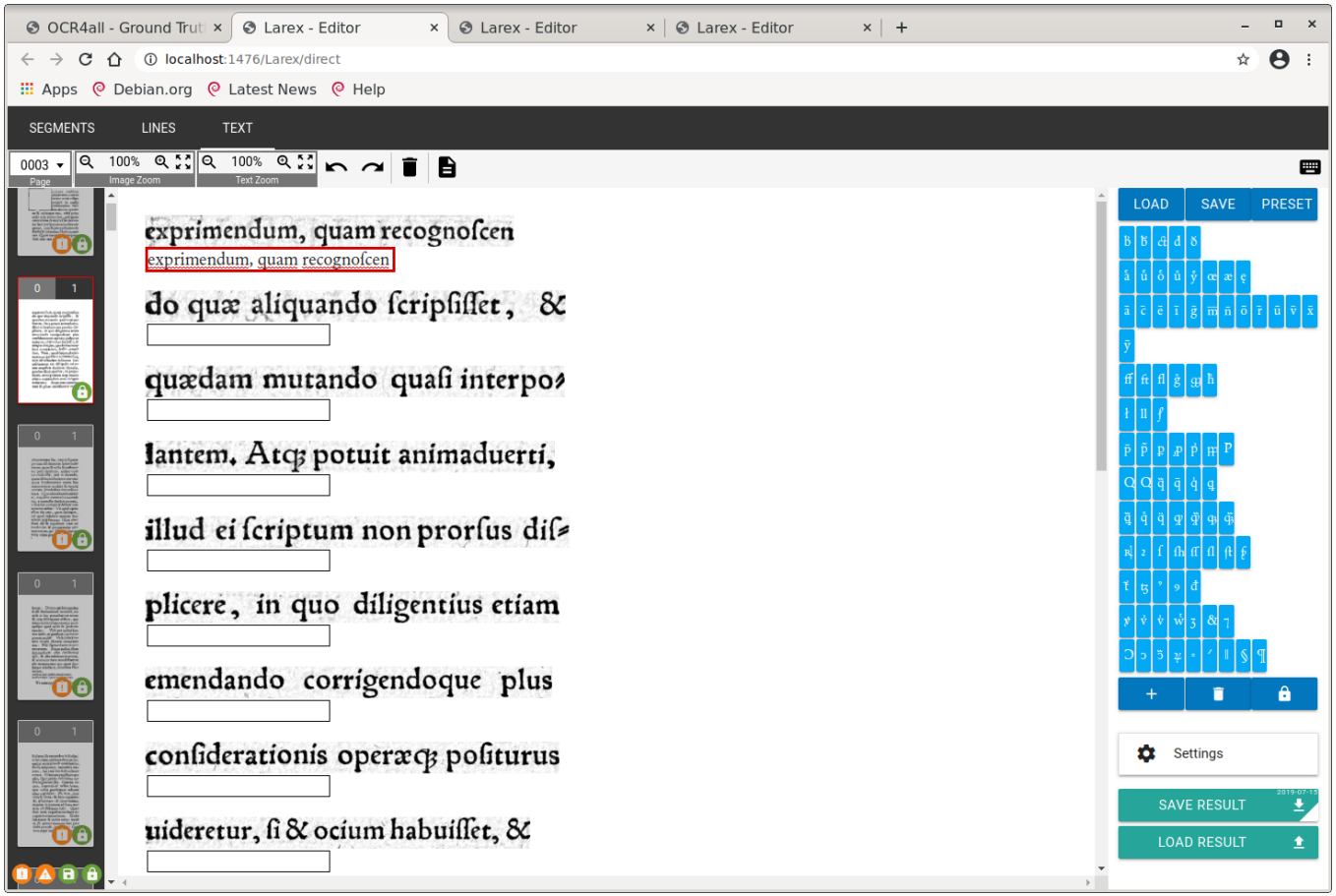
Process with OCR4all

- available for Windows, Linux, Mac
- minimal technical knowledge required
- all of the following steps can be done within OCR4all



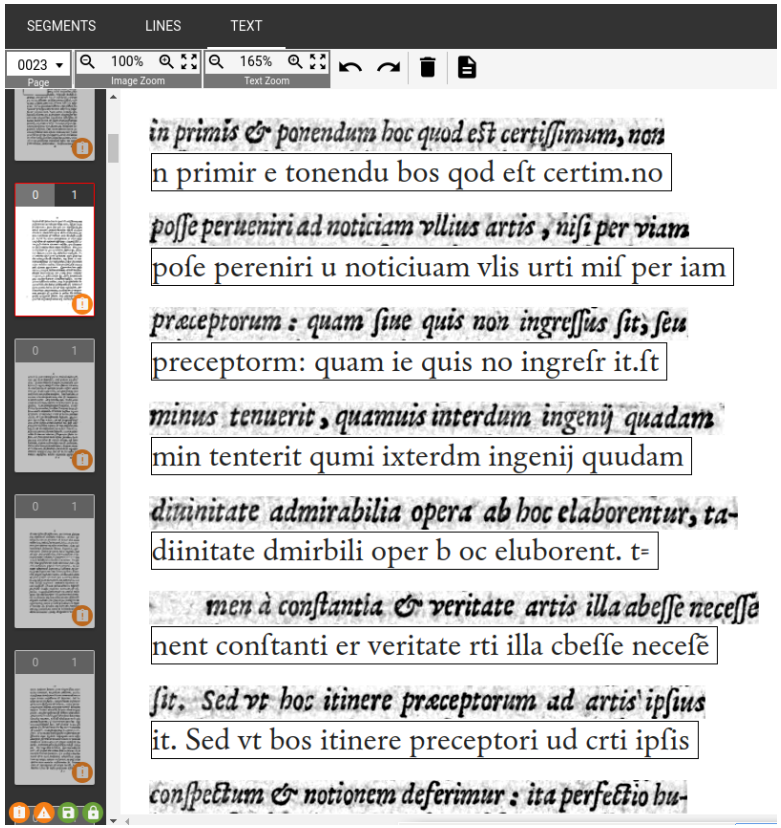
Segment pages with LAREX

binarization + grayscaling, region and line segmentation with LAREX



Recognition with existing (default) model

recognition with default/antiqua_historical (based largely on the subcorpora Early Modern Latin and Kallimachos of GT4HistOCR)



Train a book-specific model

Prepare some pages of ground truth by correcting the OCR, or transcribing from scratch (do it iteratively) and train on top of existing model, then recognise again! Remaining character error rate around 1%.

in primis & ponendum hoc quod eſt certiſſimum, non
n primir e tonendu bos qod eſt certim.no

poſſe perueniri ad noticiam vllius artis, niſi per viam
poſe pereniri u noticiuam vlis urti miſ per iam

præceptorum: quam ſiue quis non ingreſſus ſit, ſeu
preceptorm: quam ie quis no ingreſr it.ſt

minus tenuerit, quamuis interdum ingenij quadam
min tenterit qumi ixterdm ingenij quudam

diuinitate admirabilia opera ab hoc elaborentur, ta-
diinitate dmirbili oper b oc eluborent. t=

men à conſtantia & veritate artis illa abeſſe neceſſe
nent conſtanti er veritate rti illa cbeſſe neceſẽ

ſit. Sed vt hoc itinere præceptorum ad artis ipſius
it. Sed vt bos itinere preceptori ud crti ipſis

conſpectum & notionem deferimur: ita perfectio hu-
conſpectum e notioem deferimur : ita perfectio hu.

in primis & ponendum hoc quod eſt certiſſimum, non
in primis & ponendum hoc quod est certiſſimum, non

poſſe perueniri ad noticiam vllius artis, niſi per viam
poſſe perueniri ad noticiam vllius artis, niſi per viam

præceptorum: quam ſiue quis non ingreſſus ſit, ſeu
præceptorum: quam ſiue quis non ingreſſus ſit, ſeu

minus tenuerit, quamuis interdum ingenij quadam
minus tenuerit, quamuis interdum ingenij quadam

diuinitate admirabilia opera ab hoc elaborentur, ta-
diuinitate admirabilia opera ab hoc elaborentur, ta-

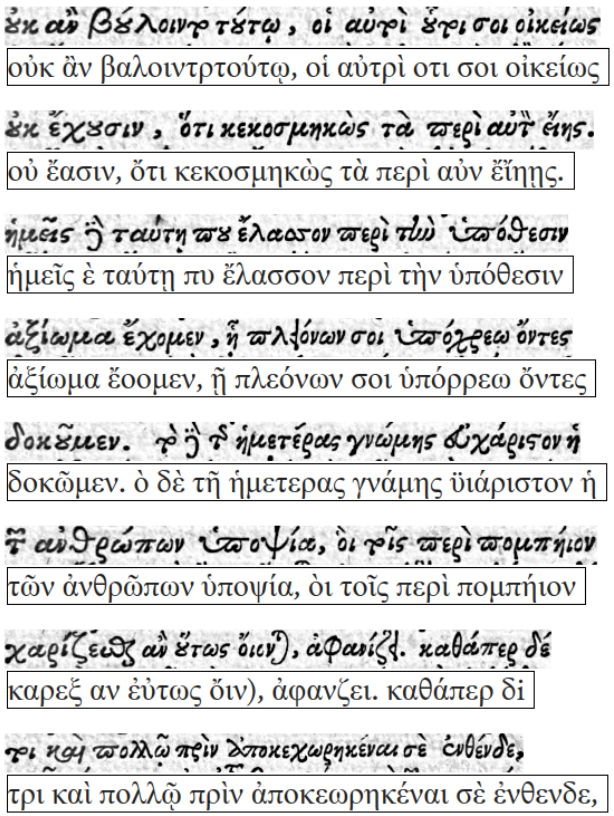
men à conſtantia & veritate artis illa abeſſe neceſſe
men conſtantia & veritate artis illa abeſſe neceſſe

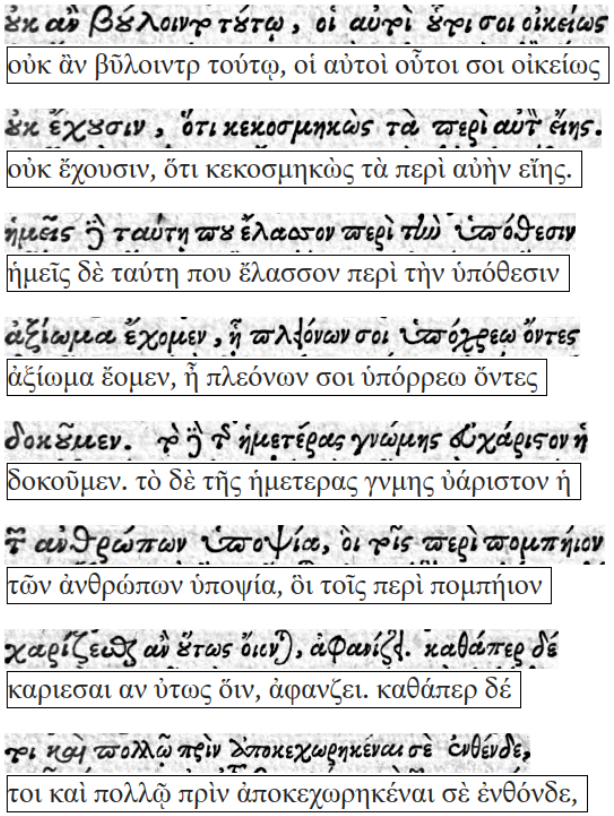
ſit. Sed vt hoc itinere præceptorum ad artis ipſius
ſit. Sed vt hos itinere præceptorum ad artis ipſius

conſpectum & notionem deferimur: ita perfectio hu-
conſpectum & notionem deferimur: ita perfectio hu-

Do the same for Greek

left: existing model trained on another greek printing and synthetic material;
right: trained with 2 pages (267-268) of ground truth and applied to p. 269;
remaining character error rate about 6%





What do we do with OCR generated texts?

- postprocessing
  - manual correction (if needed)
  - spelling normalization
- named entity recognition
- search
- prepare an edition
- . . .
- see the rest of this conference!

References

Interactive tool for scanned pages:
Scantailor-Advanced

User-friendly OCR environment with GUI:
OCR4all - *Christian Reul et al.*

Trainable state-of-the-art OCR engine (included in OCR4all):
Calamari–OCR - *Christoph Wick*, based on Ocropy by Tom Breuel

Ground Truth for early printings:
GT4HistOCR

Thank you for your attention!

Uwe Springmann
digital humanist
Email: firstname [A T] lastname.net