

# OCR for Neo-Latin

Uwe Springmann

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

This work is licensed under a Creative Commons Attribution 4.0 International License



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

2021-04-14

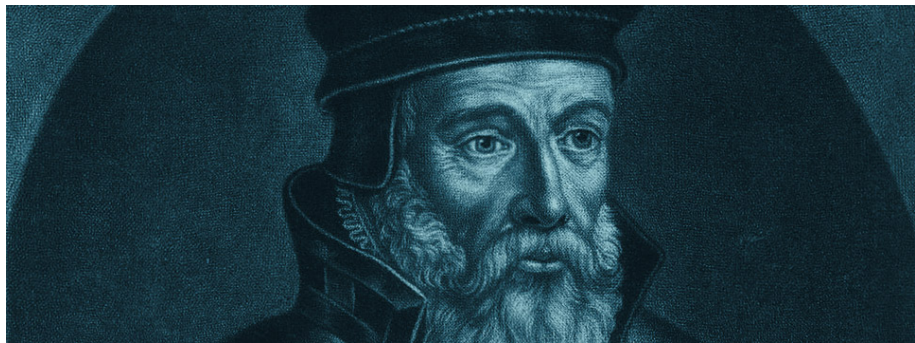
# Goals of this talk

- 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](#)

OCR4all - Project Overview - Mozilla Firefox

http://localhost:1476/ocr4all/

Project Overview 1564-ElementaRhetoricae-Camerarius | OCR4all

LOAD PROJECT ▶ CANCEL PROJECT ADJUSTMENT ⚙ EXPORT GT ⌵

Settings

Status

Overview

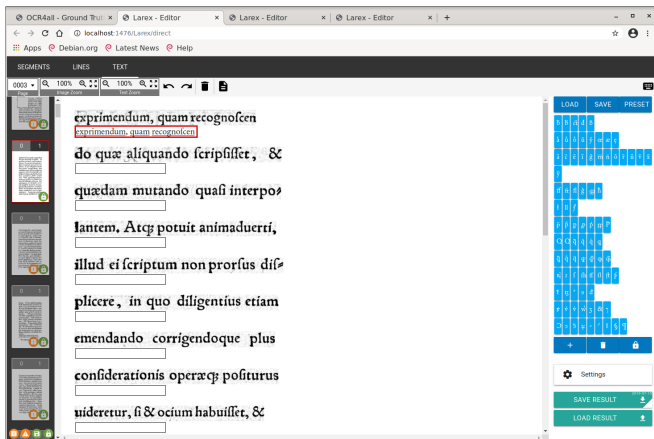
Show 10 Search:

entries

Page Identifier	Preprocessing	Noise Removal	Segmentation	Line Segmentation	Recognition	Ground Truth
0001	X	X	X	X	X	X
0002	X	X	X	X	X	X
0003	X	X	X	X	X	X
0004	X	X	X	X	X	X
0005	X	X	X	X	X	X
0006	X	X	X	X	X	X
0007	X	X	X	X	X	X
0008	X	X	X	X	X	X

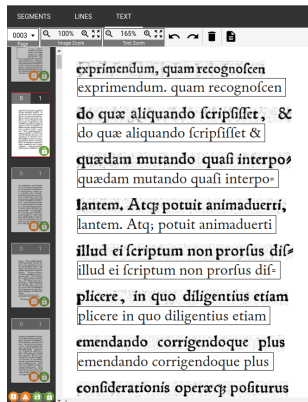
# Segment pages with LAREX

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](#))



0003 100% 165%

expriumdum, quam recognofcen  
expriumdum. quam recognofcen

do quæ aliquando fcripffiet, &  
do quæ aliquando fcripffiet &

quædam mutando quali interpo  
quædam mutando quali interpo

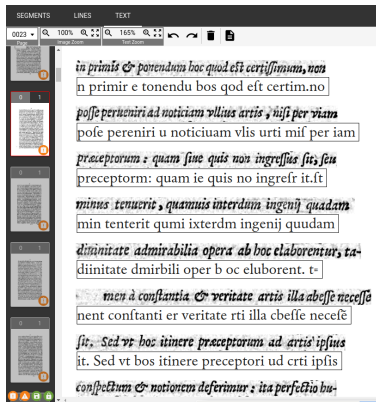
lantem. Atq; potuit animaduerti,  
lantem. Atq; potuit animaduerti

illud ei fcriptum non prorfus dif  
illud ei fcriptum non prorfus dif

plicere, in quo diligentius etiã  
plicere in quo diligentius etiã

emendando corrigendoque plus  
emendando corrigendoque plus

confiderationis operæq; politurus



0023 100% 165%

*in primis & potendum hoc quod est certiffimum, non  
n primir e tonendu bos qod eft certim.no*

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

*preceptorum: quam ſine quis non ingreſſus ſit, ſeu  
preceptorm: quam ie quis no ingrefr it.ft*

*minus tenuerit, quamuis interdum ingenij quadam  
min tenerit qumi interdm ingenij quudam*

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

*men à conſtantia & veritate arti illa abeſſe neceſſe  
nent conſtanti er veritate rti illa cbeſſe neceſe*

*ſit. Sed vt hoc itinere preceptorum ad artis ipſius  
it. Sed vt bos itinere preceptoru ud crti ipſis*

*conſpectum & notionem deferimur: ita perfectio hu-*

# 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*  
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

*praeceptorum: quam siue quis non ingressus sit, seu*  
praeceptorum: 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 à constantia & veritate artis illa abesse necesse

*sit. Sed vt hoc itinere praeceptorum ad artis ipsius*  
sit. Sed vt hoc itinere praeceptorum ad artis ipsius

*conspetum & notionem deferimur: ita perfectio hu-*  
conspetum & notionem 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

*praeceptorum: quam siue quis non ingressus sit, seu*  
praeceptorum: 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 à constantia & veritate artis illa abesse necesse

*sit. Sed vt hoc itinere praeceptorum ad artis ipsius*  
sit. Sed vt hoc itinere praeceptorum ad artis ipsius

*conspetum & notionem deferimur: ita perfectio hu-*  
conspetum & 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%

ἐκ αὐτῶν βέλονται τούτω, οἱ αὐτοὶ ἔργου σοὶ οἰκείως  
ouk an belointroutoto, oi autoi esti soi oikeiōs

ἐκ ἔχουσιν, ὅτι κεκοσμηκῶς τὰ περὶ αὐτῆς εἰης.  
ouk echousin, oti kekosmēkōs ta peri autēs eīēs.

ἡμεῖς ἢ ταύτην παρ’ ἑλασσον περὶ τὴν ὑπόθεσιν  
hēmeīs ē taūtēn par’ elasson peri tēn hypōthesin

ἀξιώματα ἔχομεν, ἢ πλεόνων σοὶ ὑπόρρουσιν ὄντες  
axiōmata echomen, ē pleōnōn soi hypōrrousin ōntes

δοκῶμεν. τὸ ἢ ἡμετέρας γνώμης ἀρχαίσιον ἢ  
dokōmen. to ē tēs hēmeteras gnōmēs archaīsiōn ē

τῶν ἀνθρώπων ὑποψία, οἱ τοῖς περὶ πομπήιον  
tōn anthrōpōn hypōpsia, oi toīs peri pomphēion

καρτερεῖς ἀν’ ἐτώσιον, ἀφανίζει. καθάπερ δὲ  
kartereīs an’ etōsion, apfanizei. kathāper dē

τρι καὶ πολλῶν πρὶν ἀποκεχωρηκέναι σὲ ἐνθενδε,  
tri kai pollōn prīn apokechōrhēkenai se enthende,

ἐκ αὐτῶν βέλονται τούτω, οἱ αὐτοὶ ἔργου σοὶ οἰκείως  
ouk an bēlōintr tōtōtō, oi autoi outoi soi oikeiōs

ἐκ ἔχουσιν, ὅτι κεκοσμηκῶς τὰ περὶ αὐτῆς εἰης.  
ouk echousin, oti kekosmēkōs ta peri autēs eīēs.

ἡμεῖς ἢ ταύτην παρ’ ἑλασσον περὶ τὴν ὑπόθεσιν  
hēmeīs dē taūtēn pou elasson peri tēn hypōthesin

ἀξιώματα ἔχομεν, ἢ πλεόνων σοὶ ὑπόρρουσιν ὄντες  
axiōmata echomen, ē pleōnōn soi hypōrrousin ōntes

δοκῶμεν. τὸ δὲ τῆς ἡμετέρας γνώμης ἀρχαίσιον ἢ  
dokōmen. to dē tēs hēmeteras gnōmēs archaīsiōn ē

τῶν ἀνθρώπων ὑποψία, οἱ τοῖς περὶ πομπήιον  
tōn anthrōpōn hypōpsia, oi toīs peri pomphēion

καρτερεῖς ἀν’ ἐτώσιον, ἀφανίζει. καθάπερ δὲ  
kartereīs an’ etōsion, apfanizei. kathāper dē

τρι καὶ πολλῶν πρὶν ἀποκεχωρηκέναι σὲ ἐνθενδε,  
toi kai pollōn prīn apokechōrhēkenai se enthende,

# 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 Ocrop by Tom Breuel

Ground Truth for early printings:

[GT4HistOCR](#)

Thank you for your attention!

Uwe Springmann

↻ digital humanist ↻

Email: `firstname [ A T ] lastname.net`