Coptic OCR

GCDH GÖTTINGEN CENTRE FOR DIGITAL HUMANITIES

# *Coptic OCR:* Even better models and improvements on user-friendliness

Eliese-Sophia Lincke (Humboldt-Universität zu Berlin)
eslincke@staff.hu-berlin.de

*Digital Coptic 3*, July 12-13, 2020 (online)

## Overview

- more fonts
- *Calamari:* new and better OCR software (as compared to OCRopus)
- *OCR4all*: a graphical user interface (GUI)

Coptic OCR

GCDH GÖTTINGEN CENTRE FOR DIGITAL HUMANITIES

# *Coptic OCR*: Chronological overview since 2016

2016: presentation of OCRopus results at the Coptic Congress, Claremont

2018: *Coptic OCR* becomes an official project

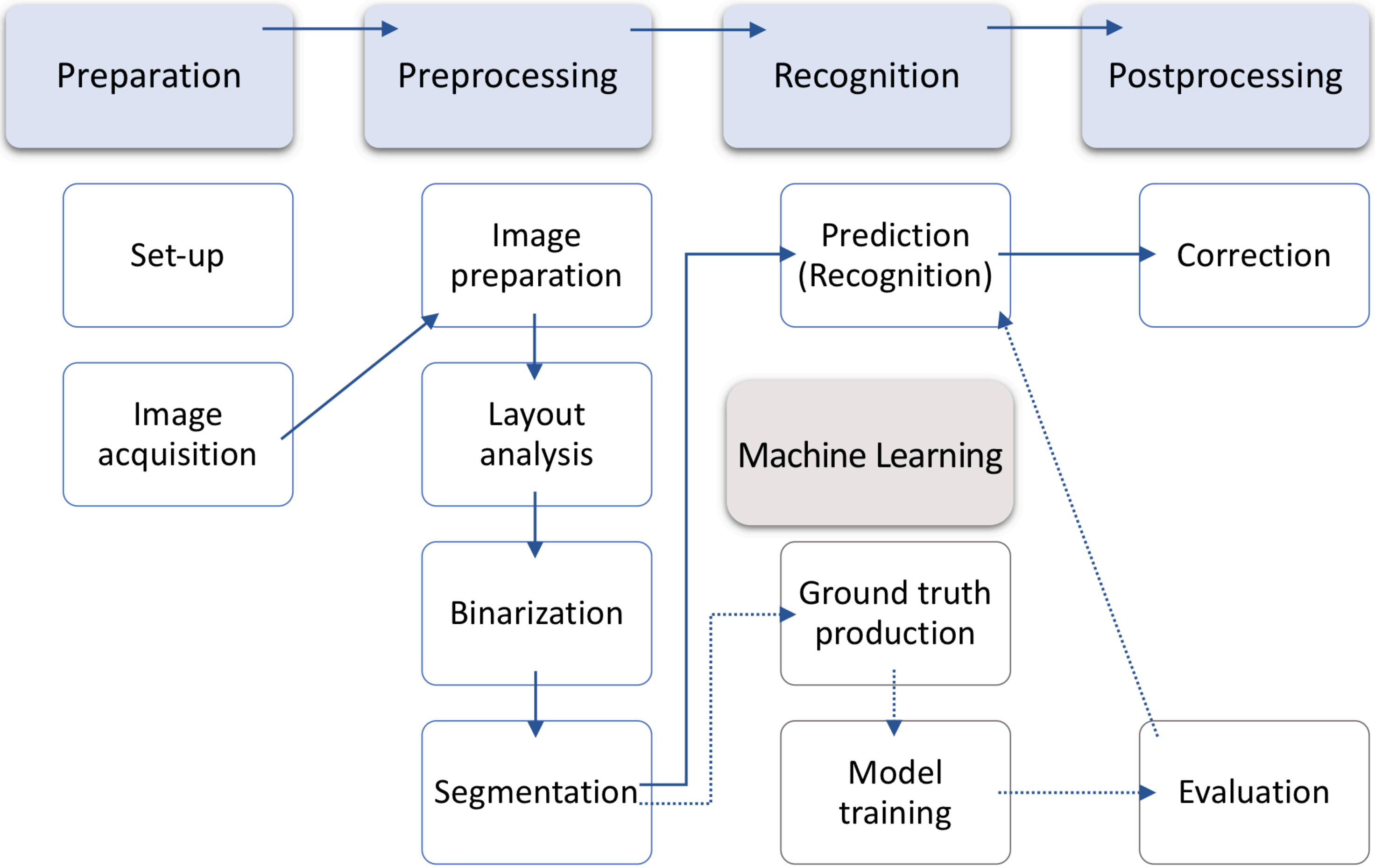
- 6 months PostDoc fellowship at the Göttingen Centre for Digital Humanities, CampusLab *Digitization and Computational Analytics* (DCA)
  PIs: Heike Behlmer, Marco Büchler, Camilla Di Biase-Dyson

2019: publication of the results of the fellowship and new tools

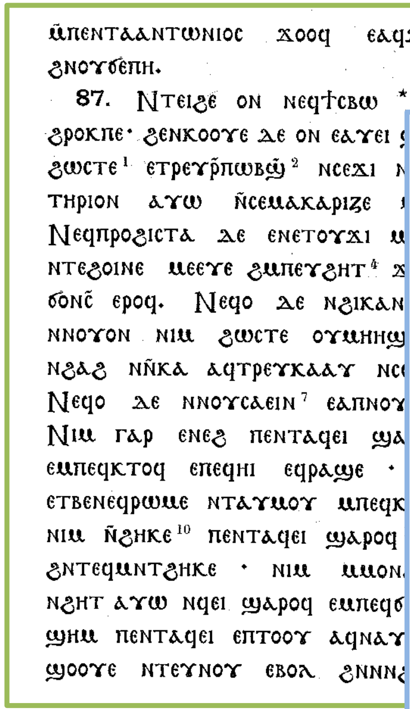
- presentation at the 3rd international conference on *Digital Access to Textual Cultural Heritage* (DATeCH) in Brussels (May)
- publication (open access) in the conference proceedings: Lincke, Bulert & Büchler (2019)
- testing new tool – Calamari
- training data of more fonts, model training (OCRopus, Calamari)
- presentation of new results in the *Berlin Digital Classicist Seminar* (November)

2020: testing new tool – OCR4all

# OCR workflow

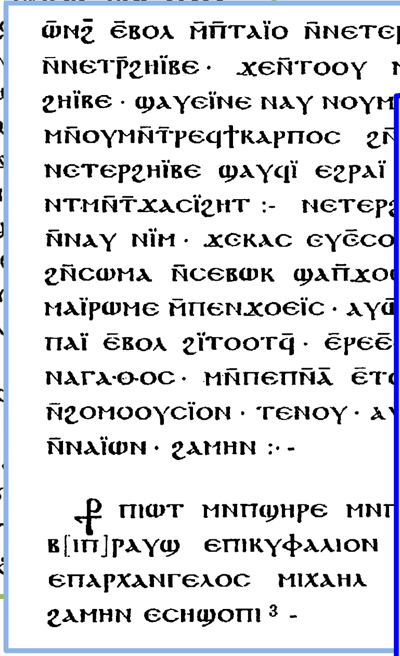


# Font overview



ⲙ̄ⲡⲉⲛⲧⲁⲁⲛⲧⲱⲛⲓⲟⲥ ϫⲟⲟϥ ⲉⲁϥϫⲱⲕ ⲉⲃⲟⲗ ⲉϫⲱϥ
ϩⲛⲟⲩϭⲉⲡⲏ.
87. Ⲛⲧⲉⲓϩⲉ ⲟⲛ ⲛⲉϥϯⲥⲃⲱ *ⲛ̄ⲣⲱⲙⲉ ⲛⲓⲙ ⲉⲧⲣⲉⲩ-
ϩⲣⲟⲕⲡⲉ· ϩⲉⲛⲕⲟⲟⲩⲉ ⲇⲉ ⲟⲛ ⲉⲁⲩⲉⲓ
ϩⲱⲥⲧⲉ[1] ⲉⲧⲣⲉⲩⲣ̄ⲡⲱⲃϣ̄[2] ⲛⲥⲉϫⲓ
ⲧⲏⲣⲓⲟⲛ ⲁⲩⲱ ⲛ̄ⲥⲉⲙⲁⲕⲁⲣⲓⲍⲉ
Ⲛⲉϥⲡⲣⲟϩⲓⲥⲧⲁ ⲇⲉ ⲉⲛⲉⲧⲟⲩϫⲓ
ⲛⲧⲉϩⲟⲓⲛⲉ ⲙⲉⲉⲩⲉ ϩⲙ̄ⲡⲉⲩϩⲏⲧ[4]
ϭⲟⲛⲥ̄ ⲉⲣⲟϥ. Ⲛⲉϥⲟ ⲇⲉ ⲛϩⲓⲕⲁⲛ
ⲛⲛⲟⲩⲟⲛ ⲛⲓⲙ ϩⲱⲥⲧⲉ ⲟⲩⲙⲏⲏϣ
ⲛϩⲁϩ ⲛⲛ̄ⲕⲁ ⲁϥⲧⲣⲉⲩⲕⲁⲁⲩ ⲛⲥ
Ⲛⲉϥⲟ ⲇⲉ ⲛⲛⲟⲩⲥⲁⲉⲓⲛ[7] ⲉⲁⲡⲛⲟⲩ
Ⲛⲓⲙ ⲅⲁⲣ ⲉⲛⲉϩ ⲡⲉⲛⲧⲁϥⲉⲓ ϣⲁ
ⲉⲙⲡⲉϥⲕⲧⲟϥ ⲉⲡⲉϥⲏⲓ ⲉϥⲣⲁϣⲉ ·
ⲉⲧⲃⲉⲛⲉϥⲣⲱⲙⲉ ⲛⲧⲁⲩⲙⲟⲩ ⲙⲡⲉϥⲕ
ⲛⲓⲙ ⲛ̄ϩⲏⲕⲉ[10] ⲡⲉⲛⲧⲁϥⲉⲓ ϣⲁⲣⲟϥ
ϩⲛⲧⲉϥⲙⲛⲧϩⲏⲕⲉ · ⲛⲓⲙ ⲙⲙⲟⲛ.
ⲛϩⲏⲧ ⲁⲩⲱ ⲛϥⲉⲓ ϣⲁⲣⲟϥ ⲉⲙⲡⲉϥϭ
ϣⲏⲙ ⲡⲉⲛⲧⲁϥⲉⲓ ⲉⲡⲧⲟⲟⲩ ⲁϥⲛⲁⲩ
ϣⲟⲟⲩⲉ ⲛⲧⲉⲩⲛⲟⲩ ⲉⲃⲟⲗ ϩⲛⲛϩ

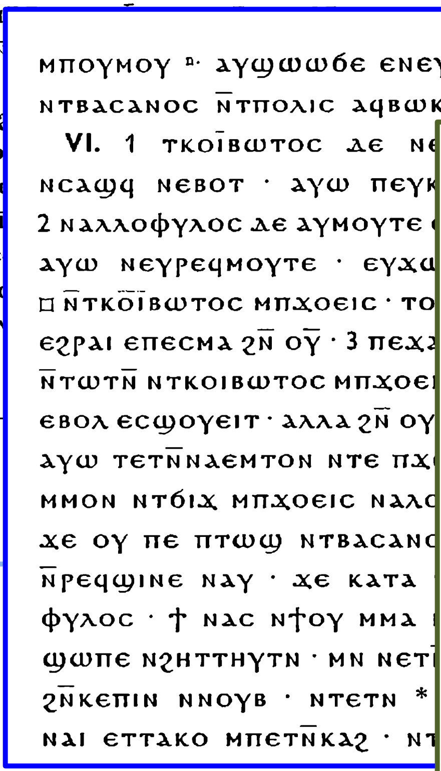
CSSC_1



ⲱⲛϩ̄ ⲉ̄ⲃⲟⲗ ⲙ̄ⲡⲧⲁïⲟ ⲛ̄ⲛⲉⲧⲉⲣϩⲏïⲃⲉ · ⲉϥϫⲱ ⲙ̄ⲙⲟⲥ : ϫⲉⲛⲁïⲁⲧⲟⲩ
ⲛ̄ⲛⲉⲧⲣ̄ϩⲏïⲃⲉ · ϫⲉⲛ̄ⲧⲟⲟⲩ ⲛⲉⲧⲟⲩⲛⲁⲥⲉⲡⲥⲱⲡⲟⲩ[1] :- 3. ⲛⲉⲧⲉⲣ-
ϩⲏïⲃⲉ · ϣⲁⲩⲉïⲛⲉ ⲛⲁⲩ ⲛⲟⲩⲙ
ⲙⲛ̄ⲟⲩⲙⲛ̄ⲧⲣⲉϥϯⲕⲁⲣⲡⲟⲥ ϩⲛ̄
ⲛⲉⲧⲉⲣϩⲏïⲃⲉ ϣⲁⲩϥï ⲉϩⲣⲁï
ⲛⲧⲙⲛ̄ⲧϫⲁⲥïϩⲏⲧ :- ⲛⲉⲧⲉⲣϩ
ⲛ̄ⲛⲁⲩ ⲛïⲙ · ϫⲉⲕⲁⲥ ⲉⲩⲉ̄ⲥⲟ
ϩⲛ̄ⲥⲱⲙⲁ ⲛ̄ⲥⲉⲃⲱⲕ ϣⲁⲡ̄ϫⲟ
ⲙⲁïⲣⲱⲙⲉ ⲙ̄ⲡⲉⲛϫⲟⲉïⲥ · ⲁⲩⲱ
ⲡⲁï ⲉ̄ⲃⲟⲗ ϩïⲧⲟⲟⲧϥ̄ · ⲉ̄ⲣⲉⲉ̄
ⲛⲁⲅⲁⲑⲟⲥ · ⲙⲛ̄ⲡⲉⲡⲛ̄ⲁ̄ ⲉ̄ⲧ
ⲛ̄ϩⲟⲙⲟⲟⲩⲥïⲟⲛ · ⲧⲉⲛⲟⲩ · ⲁ
ⲛ̄ⲛⲁïⲱⲛ · ϩⲁⲙⲏⲛ :· -

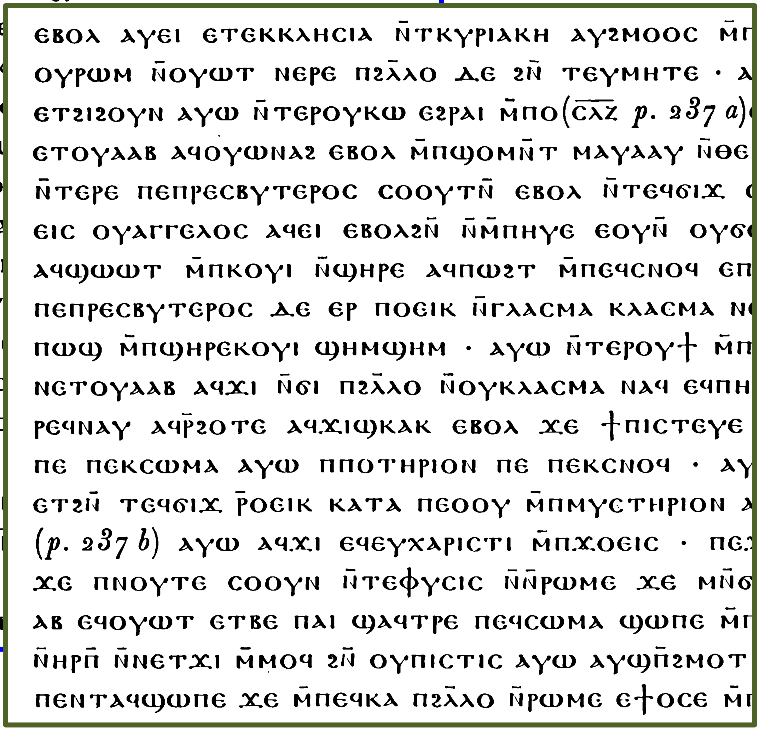
⳨ ⲡⲓⲱⲧ ⲙⲛⲡϣⲏⲣⲉ ⲙⲛⲧ
ⲃ[ⲓⲡ]ⲣⲁⲩϣ ⲉⲡⲓⲕⲩⲫⲁⲗⲓⲟⲛ
ⲉⲡⲁⲣⲭⲁⲛⲅⲉⲗⲟⲥ ⲙⲓⲭⲁⲏⲗ
ϩⲁⲙⲏⲛ ⲉⲥⲏϣⲟⲡⲓ[3] -

CSSC_2



ⲙⲡⲟⲩⲙⲟⲩ ⲛ· ⲁⲩϣⲱⲱϭⲉ ⲉⲛⲉⲩⲙⲁ ⲛⲉⲣⲙⲏ · ⲛ· ⲁⲩⲱ ⲡⲉϩⲣⲟⲟⲩ
ⲛⲧⲃⲁⲥⲁⲛⲟⲥ ⲛ̄ⲧⲡⲟⲗⲓⲥ ⲁϥⲃⲱⲕ ⲉϩⲣⲁⲓ ⲉⲧⲡⲉ .
VI. 1 ⲧⲕⲟⲓ̄ⲃⲱⲧⲟⲥ ⲇⲉ ⲛⲉ
ⲛⲥⲁϣϥ ⲛⲉⲃⲟⲧ · ⲁⲩⲱ ⲡⲉⲩⲕ
2 ⲛⲁⲗⲗⲟⲫⲩⲗⲟⲥ ⲇⲉ ⲁⲩⲙⲟⲩⲧⲉ
ⲁⲩⲱ ⲛⲉⲩⲣⲉϥⲙⲟⲩⲧⲉ · ⲉⲩϫⲱ
□ ⲛ̄ⲧⲕⲟ̄ⲓ̄ⲃⲱⲧⲟⲥ ⲙⲡϫⲟⲉⲓⲥ · ⲧⲟ
ⲉϩⲣⲁⲓ ⲉⲡⲉⲥⲙⲁ ϩⲛ̄ ⲟⲩ̄ · 3 ⲡⲉϫⲁ
ⲛ̄ⲧⲱⲧⲛ̄ ⲛⲧⲕⲟⲓⲃⲱⲧⲟⲥ ⲙⲡϫⲟⲉⲓ
ⲉⲃⲟⲗ ⲉⲥϣⲟⲩⲉⲓⲧ · ⲁⲗⲗⲁ ϩⲛ̄ ⲟⲩ
ⲁⲩⲱ ⲧⲉⲧⲛ̄ⲛⲁⲉⲙⲧⲟⲛ ⲛⲧⲉ ⲡϫ
ⲙⲙⲟⲛ ⲛⲧϭⲓϫ ⲙⲡϫⲟⲉⲓⲥ ⲛⲁⲗ
ϫⲉ ⲟⲩ ⲡⲉ ⲡⲧⲱϣ ⲛⲧⲃⲁⲥⲁⲛ
ⲛ̄ⲣⲉϥϣⲓⲛⲉ ⲛⲁⲩ · ϫⲉ ⲕⲁⲧⲁ
ⲫⲩⲗⲟⲥ · ϯ ⲛⲁⲥ ⲛϯⲟⲩ ⲙⲙⲁ
ϣⲱⲡⲉ ⲛϩⲏⲧⲧⲏⲩⲧⲛ · ⲙⲛ ⲛⲉⲧ
ϩⲛ̄ⲕⲉⲡⲓⲛ ⲛⲛⲟⲩⲃ · ⲛⲧⲉⲧⲛ *
ⲛⲁⲓ ⲉⲧⲧⲁⲕⲟ ⲙⲡⲉⲧⲛ̄ⲕⲁϩ · ⲛⲧ

CSSC_3



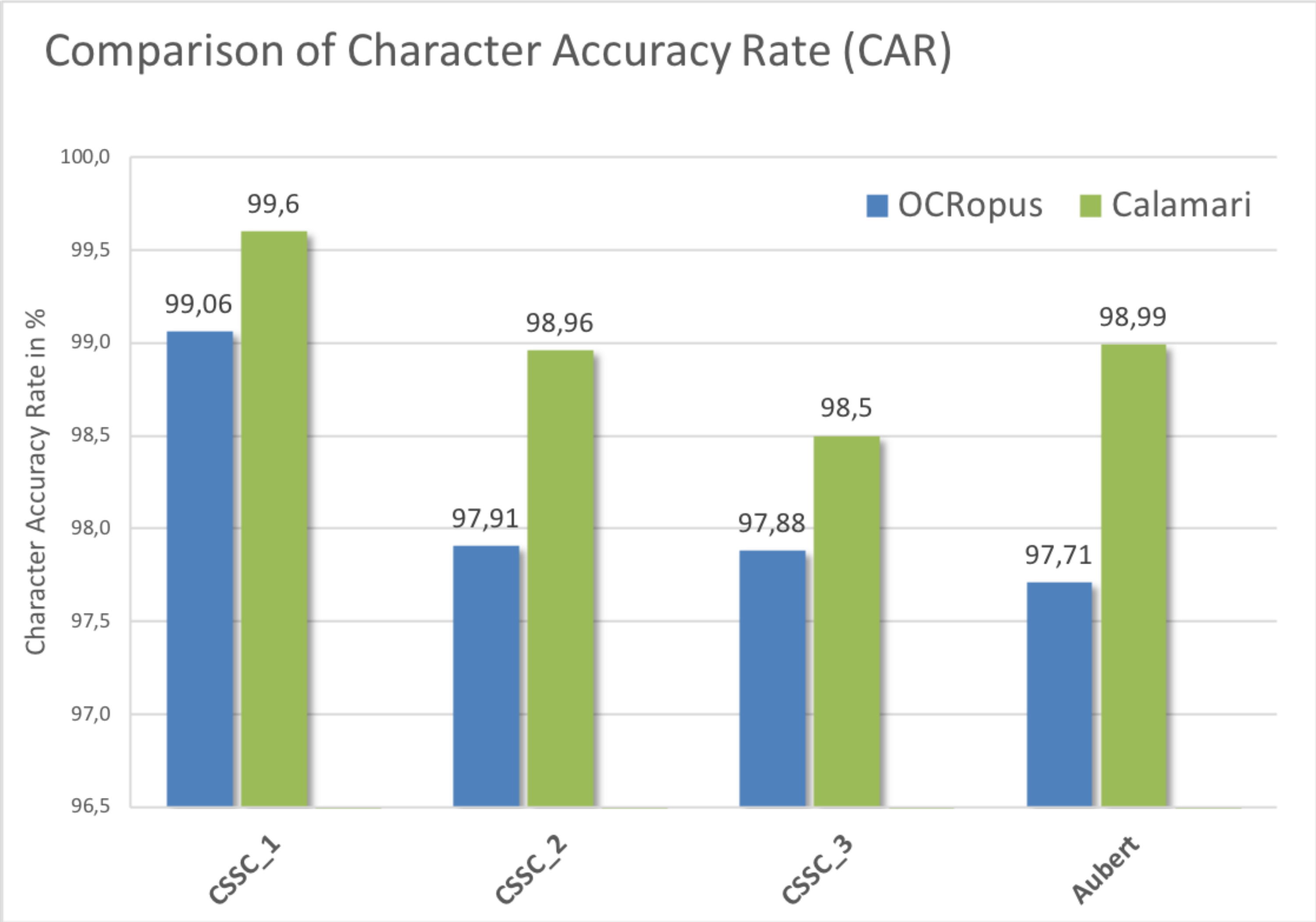
ⲉⲃⲟⲗ ⲁⲩⲉⲓ ⲉⲧⲉⲕⲕⲗⲏⲥⲓⲁ ⲛ̄ⲧⲕⲩⲣⲓⲁⲕⲏ ⲁⲩϩⲙⲟⲟⲥ ⲙ̄ⲡ
ⲟⲩⲣⲱⲙ ⲛ̄ⲟⲩⲱⲧ ⲛⲉⲣⲉ ⲡϩⲗ̄ⲗⲟ ⲇⲉ ϩⲛ̄ ⲧⲉⲩⲙⲏⲧⲉ · ⲁ
ⲉⲧϩⲓϩⲟⲩⲛ ⲁⲩⲱ ⲛ̄ⲧⲉⲣⲟⲩⲕⲱ ⲉϩⲣⲁⲓ ⲙ̄ⲡⲟ(ⲥ̄ⲗ̄ⲍ̄ *p.* 237 *a*)
ⲉⲧⲟⲩⲁⲁⲃ ⲁϥⲟⲩⲱⲛϩ ⲉⲃⲟⲗ ⲙ̄ⲡϣⲟⲙⲛ̄ⲧ ⲙⲁⲩⲁⲁⲩ ⲛ̄ⲑⲉ
ⲛ̄ⲧⲉⲣⲉ ⲡⲉⲡⲣⲉⲥⲃⲩⲧⲉⲣⲟⲥ ⲥⲟⲟⲩⲧⲛ̄ ⲉⲃⲟⲗ ⲛ̄ⲧⲉϥϭⲓϫ
ⲉⲓⲥ ⲟⲩⲁⲅⲅⲉⲗⲟⲥ ⲁϥⲉⲓ ⲉⲃⲟⲗϩⲛ̄ ⲛ̄ⲙ̄ⲡⲏⲩⲉ ⲉⲟⲩⲛ̄ ⲟⲩϭ
ⲁϥϣⲱⲱⲧ ⲙ̄ⲡⲕⲟⲩⲓ ⲛ̄ϣⲏⲣⲉ ⲁϥⲡⲱϩⲧ ⲙ̄ⲡⲉϥⲥⲛⲟϥ ⲉⲡ
ⲡⲉⲡⲣⲉⲥⲃⲩⲧⲉⲣⲟⲥ ⲇⲉ ⲉⲣ ⲡⲟⲉⲓⲕ ⲛ̄ⲅⲗⲁⲥⲙⲁ ⲕⲗⲁⲥⲙⲁ ⲛ
ⲡⲱϣ ⲙ̄ⲡϣⲏⲣⲉⲕⲟⲩⲓ ϣⲏⲙϣⲏⲙ · ⲁⲩⲱ ⲛ̄ⲧⲉⲣⲟⲩϯ ⲙ̄ⲡ
ⲛⲉⲧⲟⲩⲁⲁⲃ ⲁϥϫⲓ ⲛ̄ϭⲓ ⲡϩⲗ̄ⲗⲟ ⲛ̄ⲟⲩⲕⲗⲁⲥⲙⲁ ⲛⲁϥ ⲉϥⲡⲏ
ⲣⲉϥⲛⲁⲩ ⲁϥⲣ̄ϩⲟⲧⲉ ⲁϥϫⲓϣⲕⲁⲕ ⲉⲃⲟⲗ ϫⲉ ϯⲡⲓⲥⲧⲉⲩⲉ
ⲡⲉ ⲡⲉⲕⲥⲱⲙⲁ ⲁⲩⲱ ⲡⲡⲟⲧⲏⲣⲓⲟⲛ ⲡⲉ ⲡⲉⲕⲥⲛⲟϥ · ⲁⲩ
ⲉⲧϩⲛ̄ ⲧⲉϥϭⲓϫ ⲣ̄ⲟⲉⲓⲕ ⲕⲁⲧⲁ ⲡⲉⲟⲟⲩ ⲙ̄ⲡⲙⲩⲥⲧⲏⲣⲓⲟⲛ ⲁ
(*p.* 237 *b*) ⲁⲩⲱ ⲁϥϫⲓ ⲉϥⲉⲩⲭⲁⲣⲓⲥⲧⲓ ⲙ̄ⲡϫⲟⲉⲓⲥ · ⲡⲉϫ
ϫⲉ ⲡⲛⲟⲩⲧⲉ ⲥⲟⲟⲩⲛ ⲛ̄ⲧⲉⲫⲩⲥⲓⲥ ⲛ̄ⲛ̄ⲣⲱⲙⲉ ϫⲉ ⲙⲛ̄ϭ
ⲁⲃ ⲉϥⲟⲩⲱⲧ ⲉⲧⲃⲉ ⲡⲁⲓ ϣⲁϥⲧⲣⲉ ⲡⲉϥⲥⲱⲙⲁ ϣⲱⲡⲉ ⲙ̄ⲡ
ⲛ̄ⲏⲣⲡ̄ ⲛ̄ⲛⲉⲧϫⲓ ⲙ̄ⲙⲟϥ ϩⲛ̄ ⲟⲩⲡⲓⲥⲧⲓⲥ ⲁⲩⲱ ⲁⲩϣⲡ̄ϩⲙⲟⲧ
ⲡⲉⲛⲧⲁϥϣⲱⲡⲉ ϫⲉ ⲙ̄ⲡⲉϥⲕⲁ ⲡϩⲗ̄ⲗⲟ ⲛ̄ⲣⲱⲙⲉ ⲉϯⲟⲥⲉ ⲙ̄ⲡ

Aubert

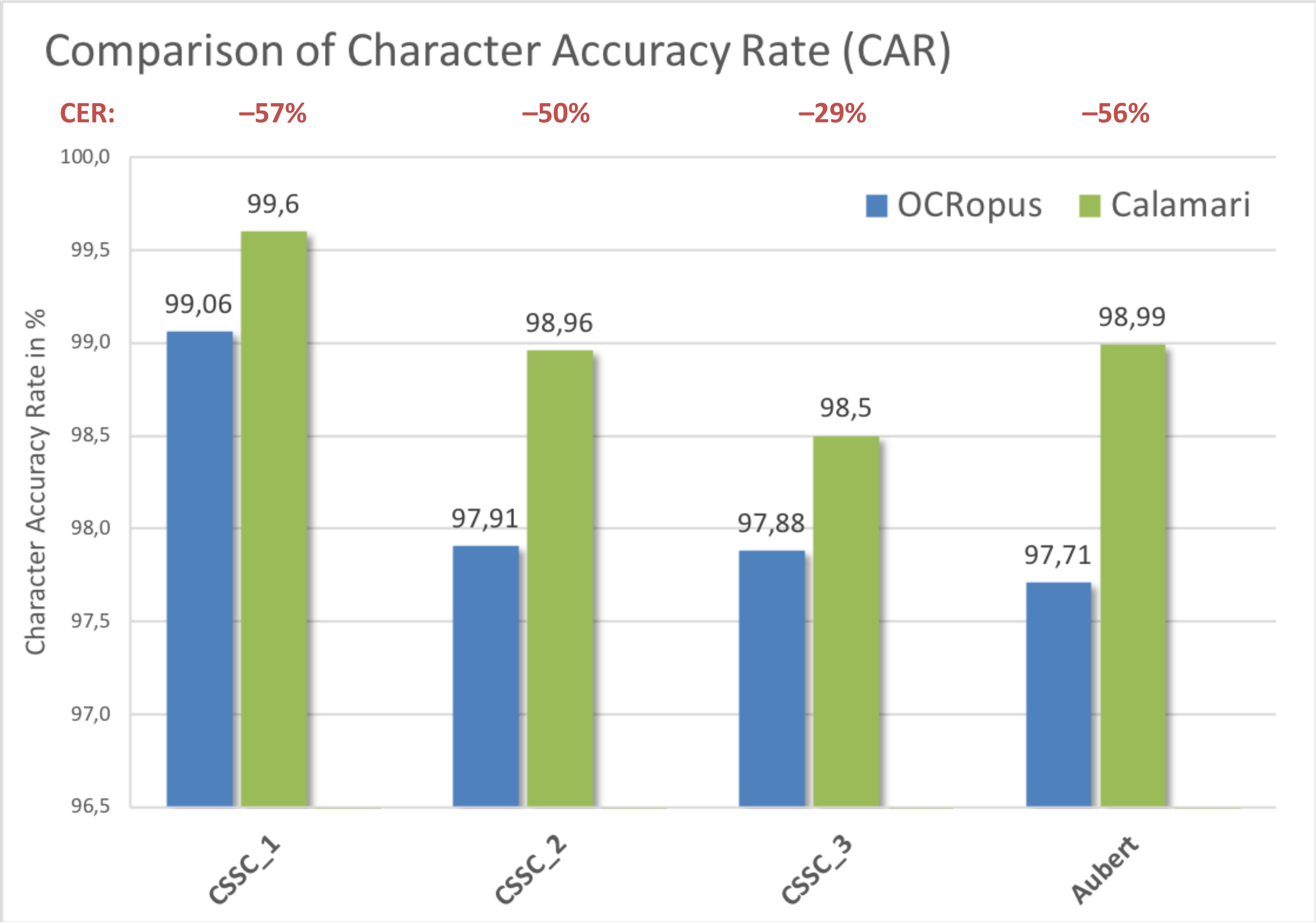
# Criteria for training (and test) data selection

- several printed text editions per font
- inhomogeneous scans
- Sahidic and Bohairic dialects (if available)
- several layouts (one column vs. two columns)
- no normalization, simplification or cleaning
- comprehensive character set for each font
  (lower case, upper case, supralinear strokes, various punctuation marks, footnote signs and other philological markup etc.)

# Best Models

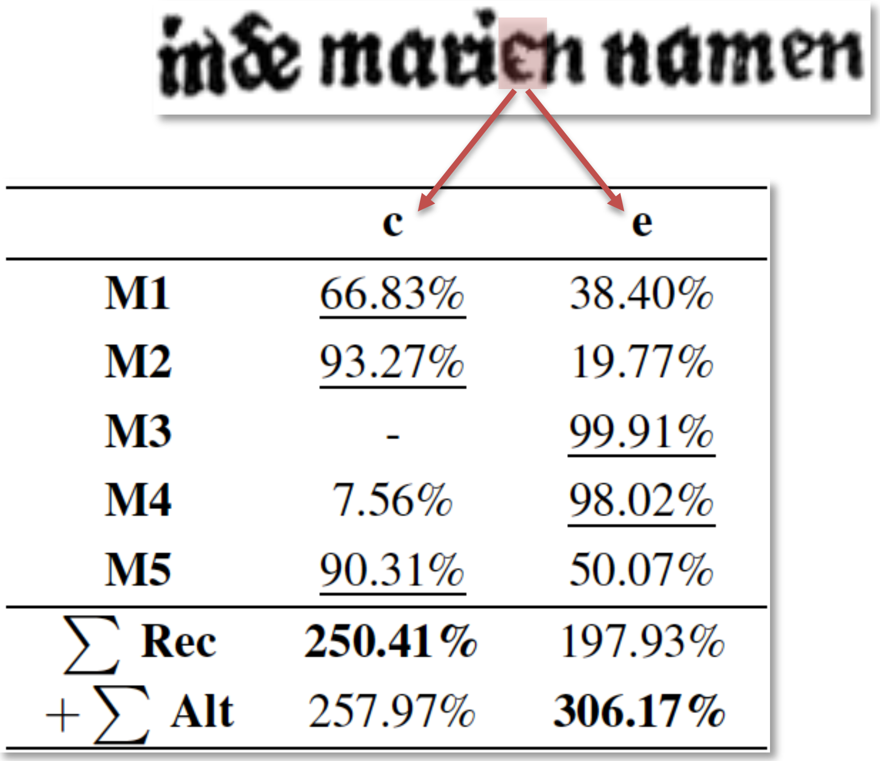

Comparison of Character Accuracy Rate (CAR)
Character Accuracy Rate in %
100,0
99,5
99,0
98,5
98,0
97,5
97,0
96,5
OCRopus
Calamari
99,06
99,6
97,91
98,96
97,88
98,5
97,71
98,99
CSSC_1
CSSC_2
CSSC_3
Aubert

# Best Models


Comparison of Character Accuracy Rate (CAR)
CER:
–57%
–50%
–29%
–56%
OCRopus
Calamari
Character Accuracy Rate in %
100,0
99,5
99,0
98,5
98,0
97,5
97,0
96,5
99,06
99,6
97,91
98,96
97,88
98,5
97,71
98,99
CSSC_1
CSSC_2
CSSC_3
Aubert

# Best Models

***Calamari*** uses *Cross Fold Training* and *Confidence Voting* to improve its predictions
cf. Reul, Springmann, Wick & Puppe (2018: Fig. 1; Table II)


inde marien namen

| | c | e |
|---|---|---|
| **M1** | 66.83% | 38.40% |
| **M2** | 93.27% | 19.77% |
| **M3** | - | 99.91% |
| **M4** | 7.56% | 98.02% |
| **M5** | 90.31% | 50.07% |
| ∑ **Rec** | **250.41%** | 197.93% |
| + ∑ **Alt** | 257.97% | **306.17%** |

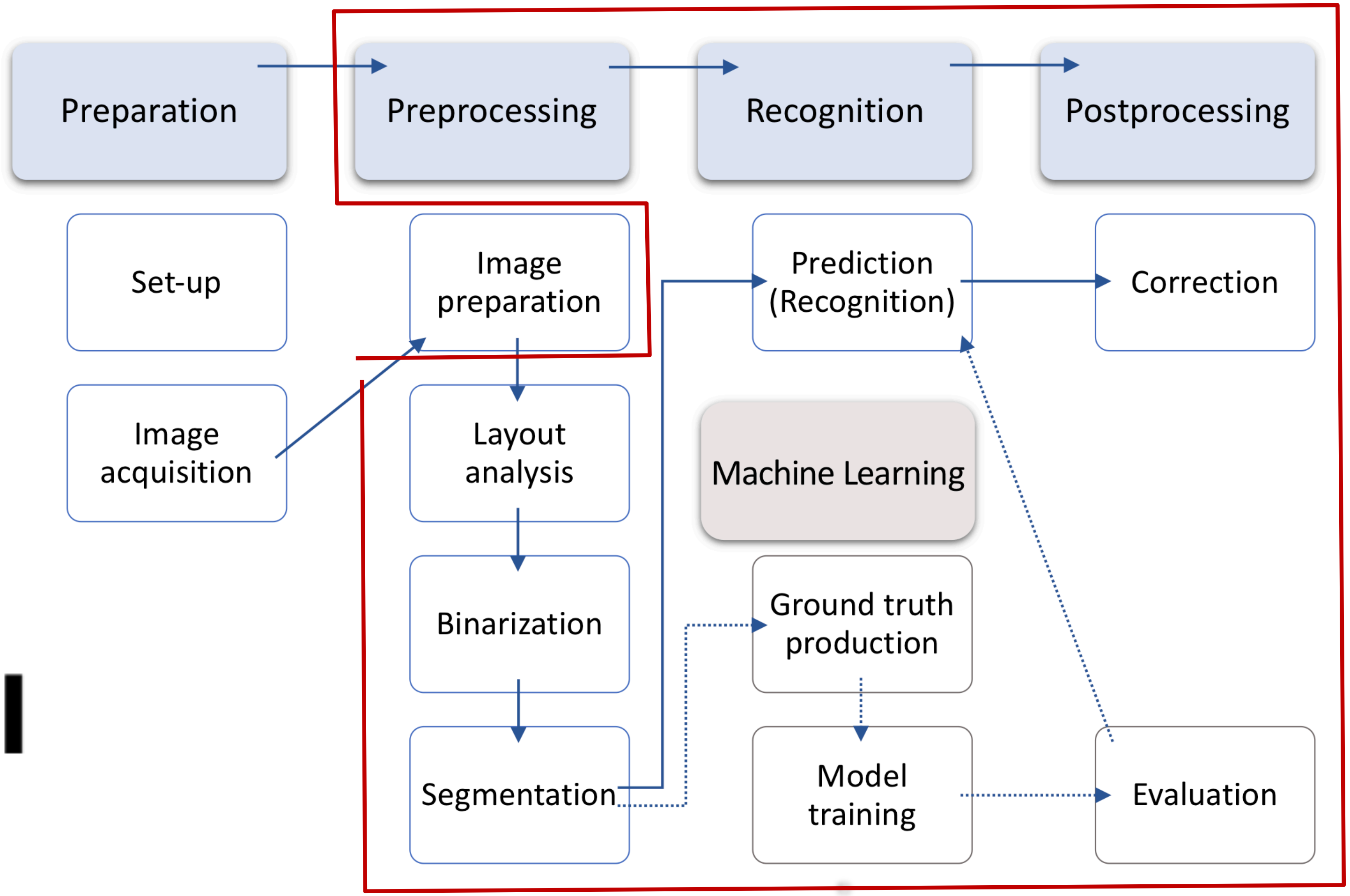
## OCR4all: Workflow overview

cf. Reul et al. (2019)

- Project at the University of Würzburg
- originally designed for historical (German) prints from the Early Modern Period
- can be used for Coptic printed texts too

https://www.uni-wuerzburg.de/en/zpd/ocr4all/

## OCR4all: Workflow overview



https://www.uni-wuerzburg.de/en/zpd/ocr4all/

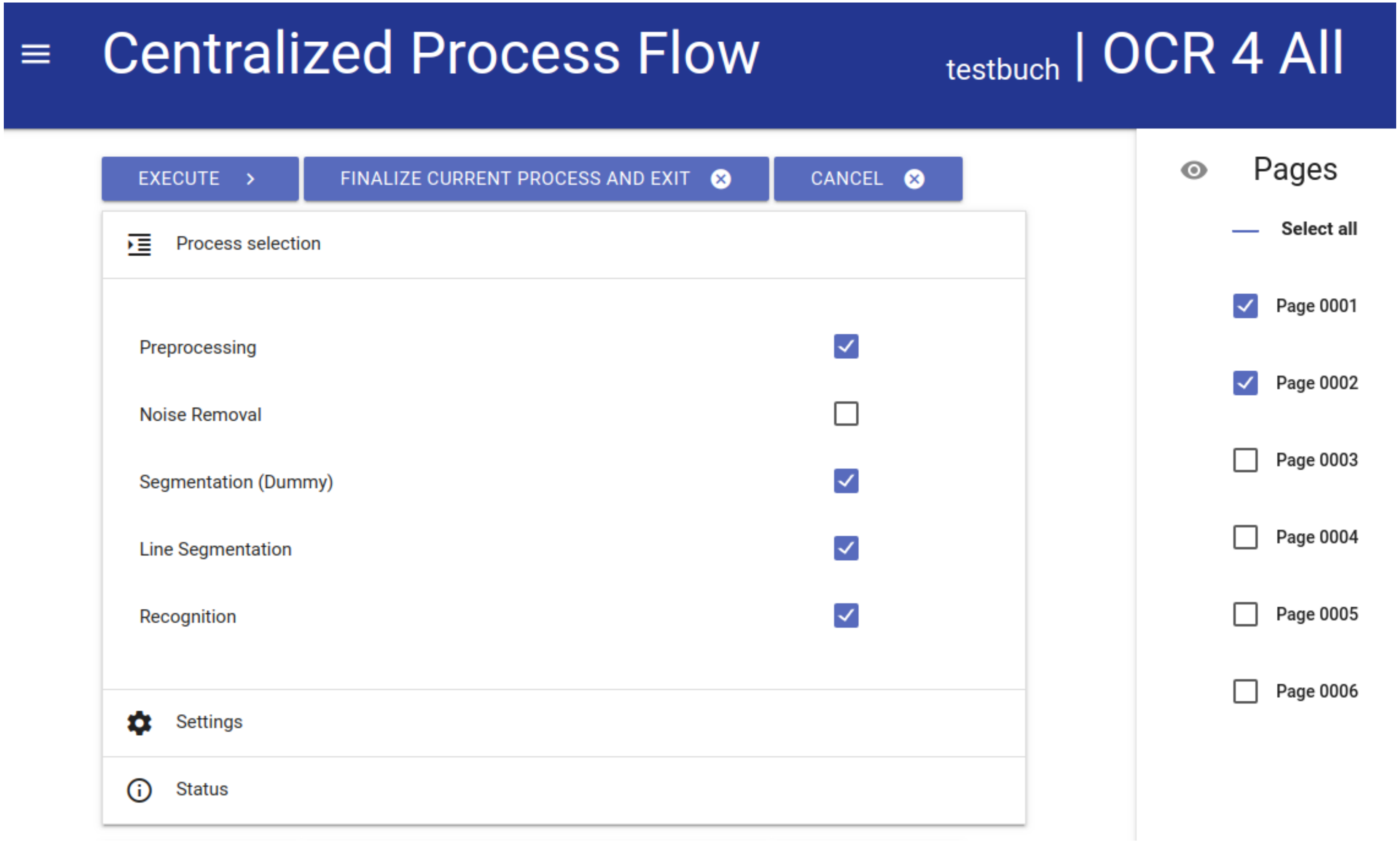
# OCR4all: Workflow overview

cf. Reul et al. (2019)

- Project at the University of Würzburg
- originally designed for historical (German) prints from the Early Modern Period
- can be used for Coptic printed texts too
- **Interface** that runs in a browser (using Docker) or in a virtual machine (VirtualBox)

https://www.uni-wuerzburg.de/en/zpd/ocr4all/

# OCR4all: Workflow overview



Process selection

- Preprocessing ☑
- Noise Removal □
- Segmentation (Dummy) ☑
- Line Segmentation ☑
- Recognition ☑

Settings

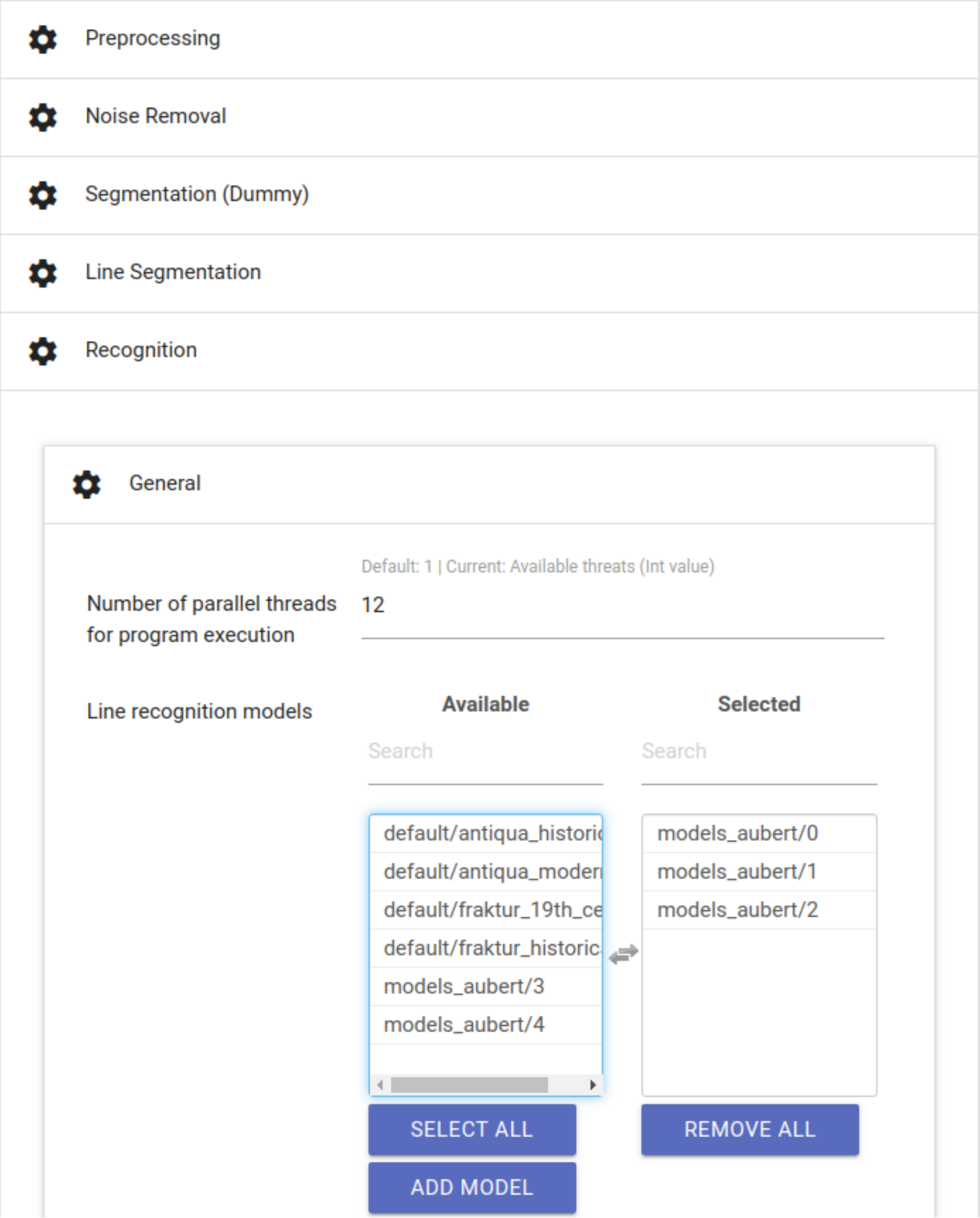
Status

Pages

— Select all

- ☑ Page 0001
- ☑ Page 0002
- □ Page 0003
- □ Page 0004
- □ Page 0005
- □ Page 0006

# OCR4all: Model selection



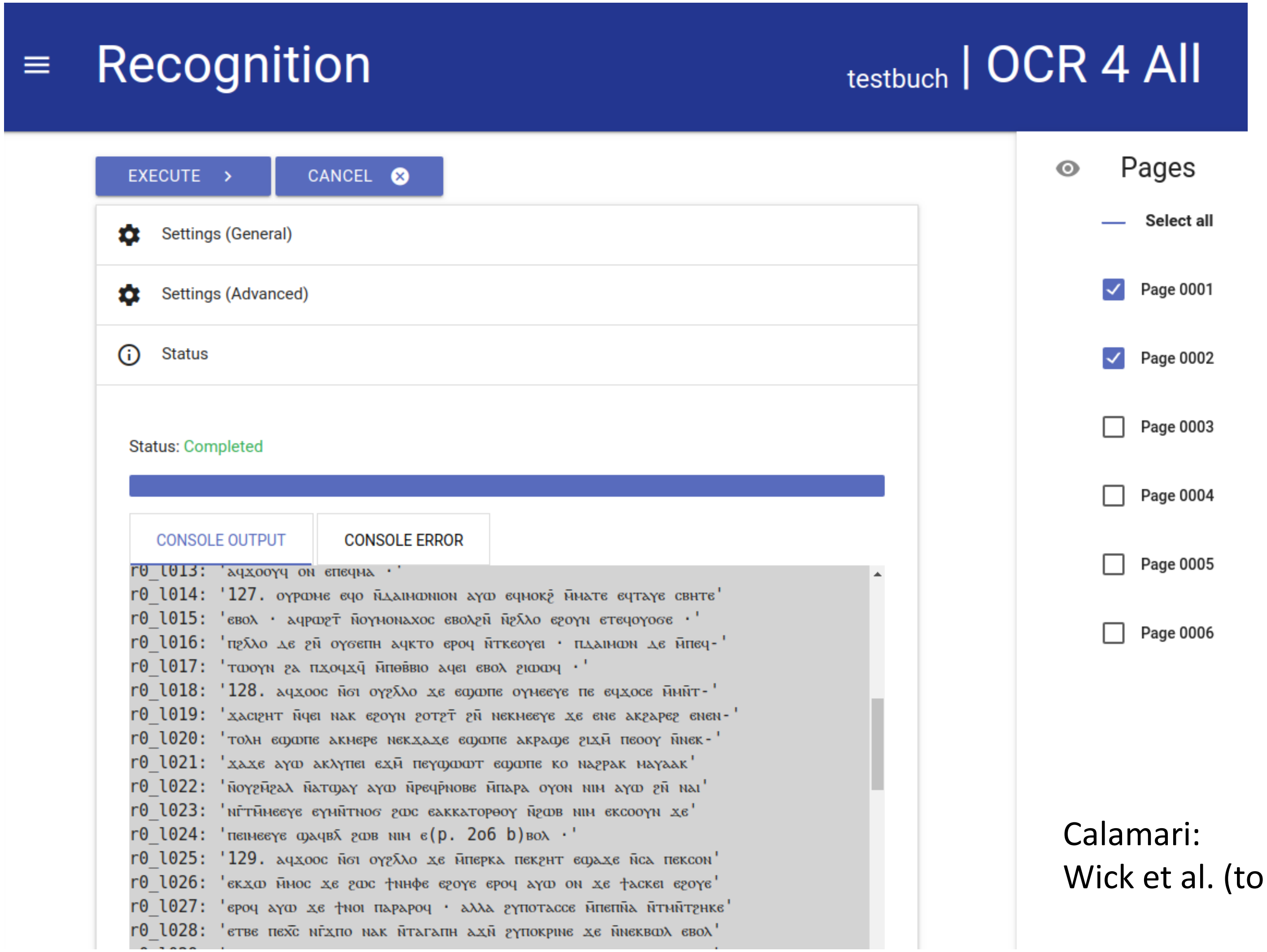
Default: 1 | Current: Available threats (Int value)

Number of parallel threads for program execution: 12

Line recognition models

| Available | Selected |
|---|---|
| default/antiqua_historic | models_aubert/0 |
| default/antiqua_moderi | models_aubert/1 |
| default/fraktur_19th_ce | models_aubert/2 |
| default/fraktur_historic | |
| models_aubert/3 | |
| models_aubert/4 | |

# OCR4all: Recognition (Calamari)

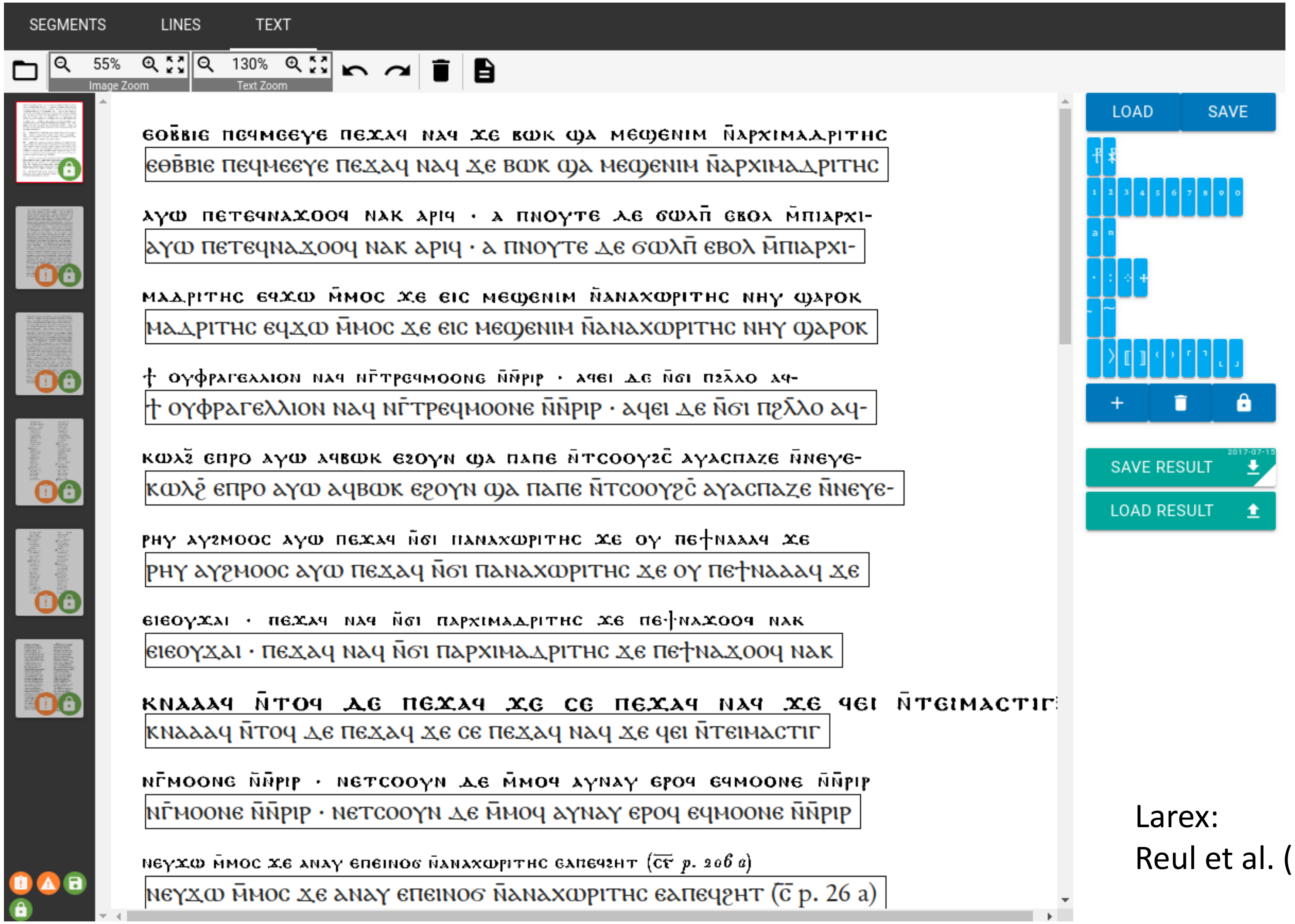


Status: Completed

```
r0_l013: 'ⲁϥϫⲟⲟⲩϥ ⲟⲛ ⲉⲡⲉϥⲙⲁ ·'
r0_l014: '127. ⲟⲩⲣⲱⲙⲉ ⲉϥⲟ ⲛ̄ⲇⲁⲓⲙⲱⲛⲓⲟⲛ ⲁⲩⲱ ⲉϥⲙⲟⲕϩ̄ ⲙ̄ⲙⲁⲧⲉ ⲉϥⲧⲁⲩⲉ ⲥⲃⲏⲧⲉ'
r0_l015: 'ⲉⲃⲟⲗ · ⲁϥⲣⲱϩⲧ̄ ⲛ̄ⲟⲩⲙⲟⲛⲁⲭⲟⲥ ⲉⲃⲟⲗϩⲛ̄ ⲛ̄ϩⲗ̄ⲗⲟ ⲉϩⲟⲩⲛ ⲉⲧⲉϥⲟⲩⲟϭⲉ ·'
r0_l016: 'ⲡϩⲗ̄ⲗⲟ ⲇⲉ ϩⲛ̄ ⲟⲩϭⲉⲡⲏ ⲁϥⲕⲧⲟ ⲉⲣⲟϥ ⲛ̄ⲧⲕⲉⲟⲩⲉⲓ · ⲡⲇⲁⲓⲙⲱⲛ ⲇⲉ ⲙ̄ⲡⲉϥ-'
r0_l017: 'ⲧⲱⲟⲩⲛ ϩⲁ ⲡϫⲟϥϫϥ̄ ⲙ̄ⲡⲑⲃ̄ⲃⲓⲟ ⲁϥⲉⲓ ⲉⲃⲟⲗ ϩⲓⲱⲱϥ ·'
r0_l018: '128. ⲁϥϫⲟⲟⲥ ⲛ̄ϭⲓ ⲟⲩϩⲗ̄ⲗⲟ ϫⲉ ⲉϣⲱⲡⲉ ⲟⲩⲙⲉⲉⲩⲉ ⲡⲉ ⲉϥϫⲟⲥⲉ ⲙ̄ⲙⲛ̄ⲧ-'
r0_l019: 'ϫⲁⲥⲓϩⲏⲧ ⲛ̄ϥⲉⲓ ⲛⲁⲕ ⲉϩⲟⲩⲛ ϩⲟⲧϩⲧ̄ ϩⲛ̄ ⲛⲉⲕⲙⲉⲉⲩⲉ ϫⲉ ⲉⲛⲉ ⲁⲕϩⲁⲣⲉϩ ⲉⲛⲉⲛ-'
r0_l020: 'ⲧⲟⲗⲏ ⲉϣⲱⲡⲉ ⲁⲕⲙⲉⲣⲉ ⲛⲉⲕϫⲁϫⲉ ⲉϣⲱⲡⲉ ⲁⲕⲣⲁϣⲉ ϩⲓϫⲙ̄ ⲡⲉⲟⲟⲩ ⲛ̄ⲛⲉⲕ-'
r0_l021: 'ϫⲁϫⲉ ⲁⲩⲱ ⲁⲕⲗⲩⲡⲉⲓ ⲉϫⲙ̄ ⲡⲉⲩϣⲱⲱⲧ ⲉϣⲱⲡⲉ ⲕⲟ ⲛⲁϩⲣⲁⲕ ⲙⲁⲩⲁⲁⲕ'
r0_l022: 'ⲛ̄ⲟⲩϩⲙ̄ϩⲁⲗ ⲛ̄ⲁⲧϣⲁⲩ ⲁⲩⲱ ⲛ̄ⲣⲉϥⲣ̄ⲛⲟⲃⲉ ⲙ̄ⲡⲁⲣⲁ ⲟⲩⲟⲛ ⲛⲓⲙ ⲁⲩⲱ ϩⲛ̄ ⲛⲁⲓ'
r0_l023: 'ⲛⲅ̄ⲧⲙ̄ⲙⲉⲉⲩⲉ ⲉⲩⲙⲛ̄ⲧⲛⲟϭ ϩⲱⲥ ⲉⲁⲕⲕⲁⲧⲟⲣⲑⲟⲩ ⲛ̄ϩⲱⲃ ⲛⲓⲙ ⲉⲕⲥⲟⲟⲩⲛ ϫⲉ'
r0_l024: 'ⲡⲉⲓⲙⲉⲉⲩⲉ ϣⲁϥⲃⲗ̄ ϩⲱⲃ ⲛⲓⲙ ⲉ(p. 206 b)ⲃⲟⲗ ·'
r0_l025: '129. ⲁϥϫⲟⲟⲥ ⲛ̄ϭⲓ ⲟⲩϩⲗ̄ⲗⲟ ϫⲉ ⲙ̄ⲡⲉⲣⲕⲁ ⲡⲉⲕϩⲏⲧ ⲉϣⲁϫⲉ ⲛ̄ⲥⲁ ⲡⲉⲕⲥⲟⲛ'
r0_l026: 'ⲉⲕϫⲱ ⲙ̄ⲙⲟⲥ ϫⲉ ϩⲱⲥ ϯⲛⲏⲫⲉ ⲉϩⲟⲩⲉ ⲉⲣⲟϥ ⲁⲩⲱ ⲟⲛ ϫⲉ ϯⲁⲥⲕⲉⲓ ⲉϩⲟⲩⲉ'
r0_l027: 'ⲉⲣⲟϥ ⲁⲩⲱ ϫⲉ ϯⲛⲟⲓ ⲡⲁⲣⲁⲣⲟϥ · ⲁⲗⲗⲁ ϩⲩⲡⲟⲧⲁⲥⲥⲉ ⲙ̄ⲡⲉⲡⲛ̄ⲁ ⲛ̄ⲧⲙⲛ̄ⲧϩⲏⲕⲉ'
r0_l028: 'ⲉⲧⲃⲉ ⲡⲉⲭ̄ⲥ̄ ⲛⲅ̄ϫⲡⲟ ⲛⲁⲕ ⲛ̄ⲧⲁⲅⲁⲡⲏ ⲁϫⲛ̄ ϩⲩⲡⲟⲕⲣⲓⲛⲉ ϫⲉ ⲛ̄ⲛⲉⲕⲃⲱⲗ ⲉⲃⲟⲗ'
```

Calamari:
Wick et al. (to appear 2020)

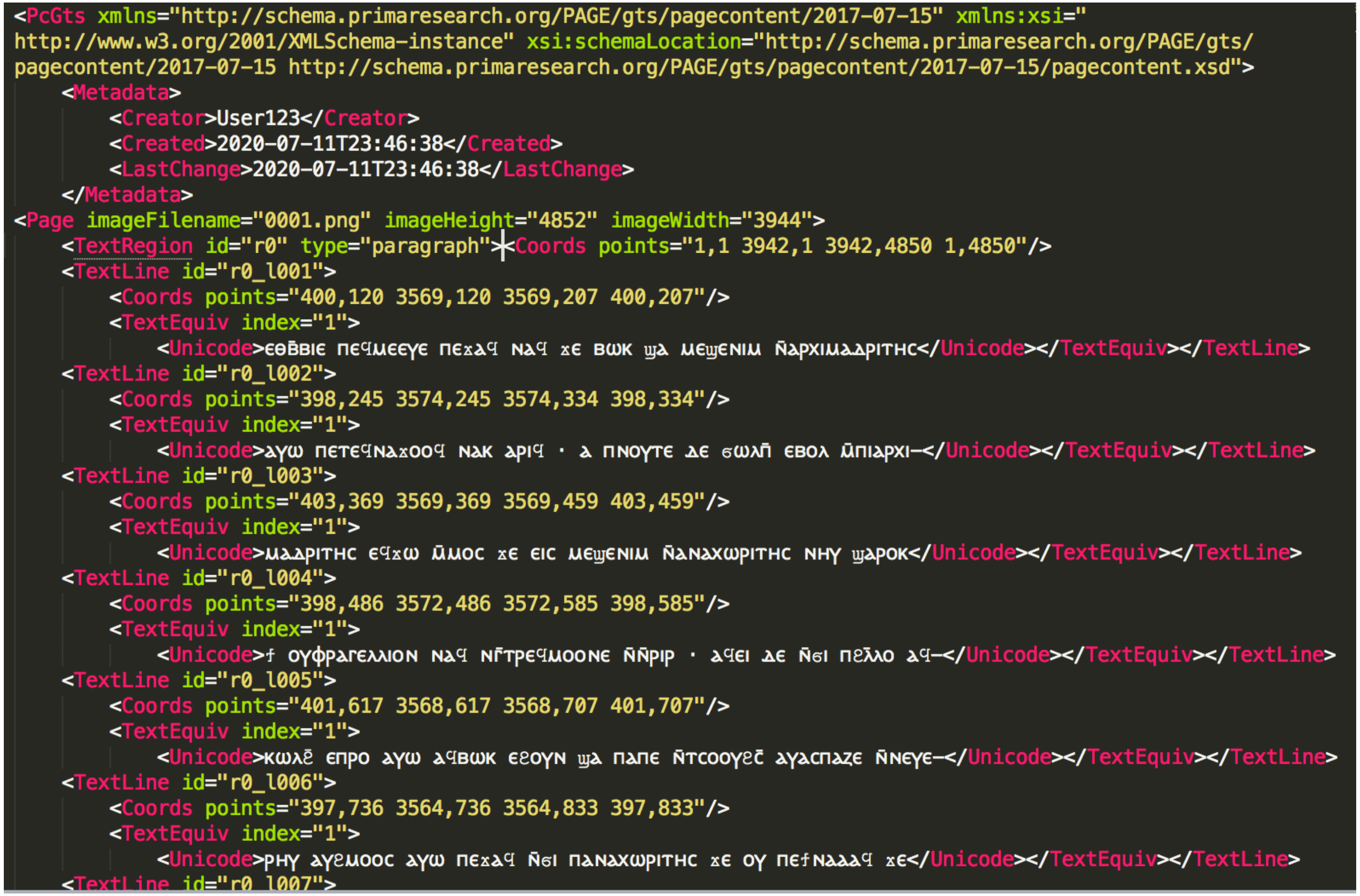
# OCR4all: Postcorrection (Larex)



ⲉⲑⲃ̄ⲃⲓⲉ ⲡⲉϥⲙⲉⲉⲩⲉ ⲡⲉϫⲁϥ ⲛⲁϥ ϫⲉ ⲃⲱⲕ ϣⲁ ⲙⲉϣⲉⲛⲓⲙ ⲛ̄ⲁⲣⲭⲓⲙⲁⲇⲣⲓⲧⲏⲥ
ⲁⲩⲱ ⲡⲉⲧⲉϥⲛⲁϫⲟⲟϥ ⲛⲁⲕ ⲁⲣⲓϥ · ⲁ ⲡⲛⲟⲩⲧⲉ ⲇⲉ ϭⲱⲗⲡ̄ ⲉⲃⲟⲗ ⲙ̄ⲡⲓⲁⲣⲭⲓ-
ⲙⲁⲇⲣⲓⲧⲏⲥ ⲉϥϫⲱ ⲙ̄ⲙⲟⲥ ϫⲉ ⲉⲓⲥ ⲙⲉϣⲉⲛⲓⲙ ⲛ̄ⲁⲛⲁⲭⲱⲣⲓⲧⲏⲥ ⲛⲏⲩ ϣⲁⲣⲟⲕ
† ⲟⲩⲫⲣⲁⲅⲉⲗⲗⲓⲟⲛ ⲛⲁϥ ⲛⲅ̄ⲧⲣⲉϥⲙⲟⲟⲛⲉ ⲛ̄ⲛ̄ⲣⲓⲣ · ⲁϥⲉⲓ ⲇⲉ ⲛ̄ϭⲓ ⲡϩ̄ⲗⲗⲟ ⲁϥ-
ⲕⲱⲗϩ̄ ⲉⲡⲣⲟ ⲁⲩⲱ ⲁϥⲃⲱⲕ ⲉϩⲟⲩⲛ ϣⲁ ⲡⲁⲡⲉ ⲛ̄ⲧⲥⲟⲟⲩϩⲥ̄ ⲁⲩⲁⲥⲡⲁⲍⲉ ⲛ̄ⲛⲉⲩⲉ-
ⲣⲏⲩ ⲁⲩϩⲙⲟⲟⲥ ⲁⲩⲱ ⲡⲉϫⲁϥ ⲛ̄ϭⲓ ⲡⲁⲛⲁⲭⲱⲣⲓⲧⲏⲥ ϫⲉ ⲟⲩ ⲡⲉϯⲛⲁⲁⲁϥ ϫⲉ
ⲉⲓⲉⲟⲩϫⲁⲓ · ⲡⲉϫⲁϥ ⲛⲁϥ ⲛ̄ϭⲓ ⲡⲁⲣⲭⲓⲙⲁⲇⲣⲓⲧⲏⲥ ϫⲉ ⲡⲉϯⲛⲁϫⲟⲟϥ ⲛⲁⲕ
ⲕⲛⲁⲁⲁϥ ⲛ̄ⲧⲟϥ ⲇⲉ ⲡⲉϫⲁϥ ϫⲉ ⲥⲉ ⲡⲉϫⲁϥ ⲛⲁϥ ϫⲉ ϥⲉⲓ ⲛ̄ⲧⲉⲓⲙⲁⲥⲧⲓⲅ
ⲛⲅ̄ⲙⲟⲟⲛⲉ ⲛ̄ⲛ̄ⲣⲓⲣ · ⲛⲉⲧⲥⲟⲟⲩⲛ ⲇⲉ ⲙ̄ⲙⲟϥ ⲁⲩⲛⲁⲩ ⲉⲣⲟϥ ⲉϥⲙⲟⲟⲛⲉ ⲛ̄ⲛ̄ⲣⲓⲣ
ⲛⲉⲩϫⲱ ⲙ̄ⲙⲟⲥ ϫⲉ ⲁⲛⲁⲩ ⲉⲡⲉⲓⲛⲟϭ ⲛ̄ⲁⲛⲁⲭⲱⲣⲓⲧⲏⲥ ⲉⲁⲡⲉϥϩⲏⲧ (ⲥ̄ p. 26 a)

Larex:
Reul et al. (2017)

# OCR4all: Output (xml)



```
<PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2017-07-15" xmlns:xsi="
http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/
pagecontent/2017-07-15 http://schema.primaresearch.org/PAGE/gts/pagecontent/2017-07-15/pagecontent.xsd">
    <Metadata>
        <Creator>User123</Creator>
        <Created>2020-07-11T23:46:38</Created>
        <LastChange>2020-07-11T23:46:38</LastChange>
    </Metadata>
<Page imageFilename="0001.png" imageHeight="4852" imageWidth="3944">
    <TextRegion id="r0" type="paragraph"><Coords points="1,1 3942,1 3942,4850 1,4850"/>
    <TextLine id="r0_l001">
        <Coords points="400,120 3569,120 3569,207 400,207"/>
        <TextEquiv index="1">
            <Unicode>ⲉⲑⲃ̄ⲃⲓⲉ ⲡⲉϥⲙⲉⲉⲩⲉ ⲡⲉϫⲁϥ ⲛⲁϥ ϫⲉ ⲃⲱⲕ ϣⲁ ⲙⲉϣⲉⲛⲓⲙ ⲛ̄ⲁⲣⲭⲓⲙⲁⲁⲣⲓⲧⲏⲥ</Unicode></TextEquiv></TextLine>
    <TextLine id="r0_l002">
        <Coords points="398,245 3574,245 3574,334 398,334"/>
        <TextEquiv index="1">
            <Unicode>ⲁⲩⲱ ⲡⲉⲧⲉϥⲛⲁϫⲟⲟϥ ⲛⲁⲕ ⲁⲣⲓϥ · ⲁ ⲡⲛⲟⲩⲧⲉ ⲇⲉ ϭⲱⲗⲡ̄ ⲉⲃⲟⲗ ⲙ̄ⲡⲓⲁⲣⲭⲓ–</Unicode></TextEquiv></TextLine>
    <TextLine id="r0_l003">
        <Coords points="403,369 3569,369 3569,459 403,459"/>
        <TextEquiv index="1">
            <Unicode>ⲙⲁⲁⲣⲓⲧⲏⲥ ⲉϥϫⲱ ⲙ̄ⲙⲟⲥ ϫⲉ ⲉⲓⲥ ⲙⲉϣⲉⲛⲓⲙ ⲛ̄ⲁⲛⲁⲭⲱⲣⲓⲧⲏⲥ ⲛⲏⲩ ϣⲁⲣⲟⲕ</Unicode></TextEquiv></TextLine>
    <TextLine id="r0_l004">
        <Coords points="398,486 3572,486 3572,585 398,585"/>
        <TextEquiv index="1">
            <Unicode>ϯ ⲟⲩⲫⲁⲣⲁⲅⲉⲗⲗⲓⲟⲛ ⲛⲁϥ ⲛⲅ̄ⲧⲣⲉϥⲙⲟⲟⲛⲉ ⲛ̄ⲛ̄ⲣⲓⲣ · ⲁϥⲉⲓ ⲇⲉ ⲛ̄ϭⲓ ⲡϩⲗⲗⲟ ⲁϥ–</Unicode></TextEquiv></TextLine>
    <TextLine id="r0_l005">
        <Coords points="401,617 3568,617 3568,707 401,707"/>
        <TextEquiv index="1">
            <Unicode>ⲕⲱⲗϩ̄ ⲉⲡⲣⲟ ⲁⲩⲱ ⲁϥⲃⲱⲕ ⲉϩⲟⲩⲛ ϣⲁ ⲡⲁⲡⲉ ⲛ̄ⲧⲥⲟⲟⲩϩⲥ̄ ⲁⲩⲁⲥⲡⲁⲍⲉ ⲛ̄ⲛⲉⲩⲉ–</Unicode></TextEquiv></TextLine>
    <TextLine id="r0_l006">
        <Coords points="397,736 3564,736 3564,833 397,833"/>
        <TextEquiv index="1">
            <Unicode>ⲣⲏⲩ ⲁⲩϩⲙⲟⲟⲥ ⲁⲩⲱ ⲡⲉϫⲁϥ ⲛ̄ϭⲓ ⲡⲁⲛⲁⲭⲱⲣⲓⲧⲏⲥ ϫⲉ ⲟⲩ ⲡⲉϯⲛⲁⲁⲁϥ ϫⲉ</Unicode></TextEquiv></TextLine>
    <TextLine id="r0_l007">
```

# Hopes (Plans?) for the future

- a Digital Coptic repositorium
- an infrastructure in which OCR4all can be run on a (remote) server – no need for local installation

- better documentation for Coptic OCR
- hands-on workshops (like the one planned for this years' Coptic Congress …): offline and online

# Coptic OCR data

Data repository for the Coptic OCR project (“working GitLab repository”)

(At the moment, the repository, is accessible without registration. This may, however, change in the future due to copyright concerns etc. In that case, please, contact us.)

DOI: 21.11101/0000-0007-C9D1-A

https://vcs.etrap.eu/Coptic-OCR/datasets

eslincke@staff.hu-berlin.de

# Thank you …

… for having contributed to and/or supported *Coptic OCR*:

Heike Behlmer
Marco Büchler
Kirill Bulert
Camilla Di Biase-Dyson
Frank Feder
Florian Finck
Jürgen Knauth
So Miyagawa
Tobias Paul
Christian Reul
Malte Rosenau
Caroline Sporleder
Uwe Springmann
Ronnie Vuine

# References

Bourcellier, Laurent. 2006. *Création d'une typographie numérique copte adaptée aux usages éditoriaux*, Diplôme supérieur d'arts appliqués, arts et techniques de communication option création typographique, Livret d'accompagnement, Ecole Estienne, Paris.

Lincke, Eliese-Sophia, Kirill Bulert and Marco Büchler. 2019. Optical Character Recognition for Coptic fonts: A multi-source approach for scholarly editions, in: *DATeCH2019 – Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage,* 87-91.
DOI: 10.1145/3322905.3322931

Reul, Christian, Uwe Springmann, Christoph Wick and Frank Puppe. 2018. Improving OCR Accuracy on Early Printed Books by Utilizing Cross Fold Training and Voting, in: *Proceedings of the 13th IAPR International Workshop on Document Analysis Systems (DAS)*, Vienna, 24-27 April 2018, IEEE: 423-428.
DOI: 10.1109/DAS.2018.30

Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner and Frank Puppe. 2019. OCR4all---An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings, in: *Applied Sciences* 9(22), No. 4853.
DOI: 10.3390/app9224853

Reul, Christian, Uwe Springmann and Frank Puppe. 2017. LAREX: A semi-automatic open-source Tool for Layout Analysis and Region Extraction on Early Printed Books, in: *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage – DATeCH2017*, ACM: 137-142.
DOI: 10.1145/3078081.3078097

Wick, Christoph, Christian Reul and Frank Puppe. to appear (2020). Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition, in: *Digital Humanities Quarterly* 14(2).