Utilization of Common OCR Tools for Typeset Coptic Texts

So Miyagawa Kirill Bulert Marco Büchler

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN
BILDUNG UND RELIGION SFB 1136
eTRAP
KELLIA
NATIONAL ENDOWMENT FOR THE HUMANITIES
DFG

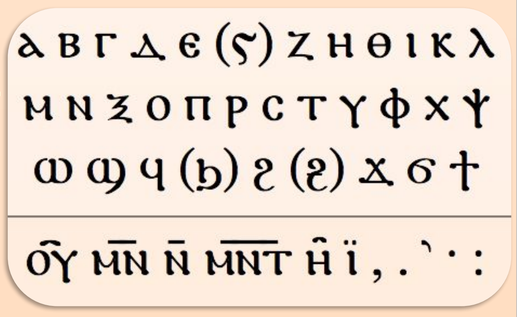
Coptic

- The final stage of the Ancient Egyptian language used in Egypt from ca. the third century
- A new writing system based on the Greek alphabet and several letters from Demotic (a language stage and writing system used in Egypt from ~ 700 BCE)
- A language transmitted in several regional forms (dialects) with a large production of manuscripts in Sahidic Coptic, the dialect at the basis of our OCR work

Coptic alphabet

- Ca. 30 letters.
- Several diacritics such as tremas, circumflexes, supralinear strokes etc.
- Several punctuation marks such as dots, commas, and colons
- Editorial marks in editions

Upper: the Coptic alphabet
Lower: diacritics and punctuation



Why is Coptic OCR needed?

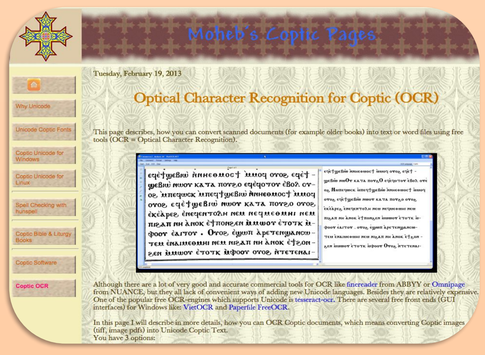
- OCR for Coptic is not well-developed.
- Almost all the Coptic texts in past publications were not OCRed.
- OCR for Coptic is needed by many DH projects in Coptic.
- There is a small amount of human power in Coptology compared with the large amount of unOCRed Coptic editions.

Coptic DH projects (selected)

- SFB 1136 (Göttingen)
  - ➢ Creates a text corpus of selected monastic works in Coptic
- Digital Edition of the Coptic Old Testament (Göttingen)
  - ➢ Creates a digital edition of the Coptic translation of the Old Testament
- Coptic SCRIPTORIUM (Georgetown/Pacific)
  - ➢ Creates a linguistically annotated Coptic corpus

Existing Coptic OCR

- Tesseract (developed by Ray Smith) for Coptic, trained by Moheb Mekhaiel (http://www.moheb.de/ocr.html)

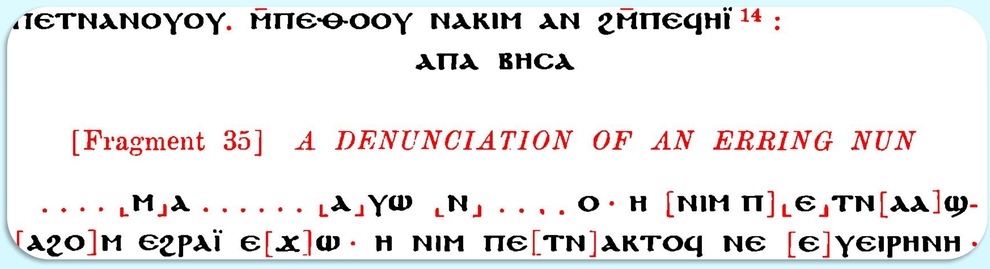


New method: Ocropy

- Python-based OCR package
- Using recurrent neural networks
- Originally developed by Thomas Breuel
- Available at https://github.com/tmbdev/ocropy
- Trained for Coptic by our group and our collaborator Eliese-Sophia Lincke (Berlin)

Coptic diacritics, punctuation, and editorial signs

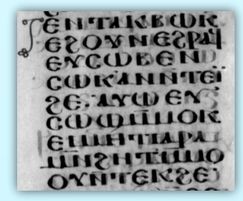
- Create problems with Tesseract
- Better processed by Ocropy


ⲙⲉⲧⲛⲁⲛⲟⲩⲟⲩ. ⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲁⲕⲓⲙ ⲁⲛ ϩⲙ̄ⲡⲉϥⲏⲓ [14] :
ⲁⲡⲁ ⲃⲏⲥⲁ

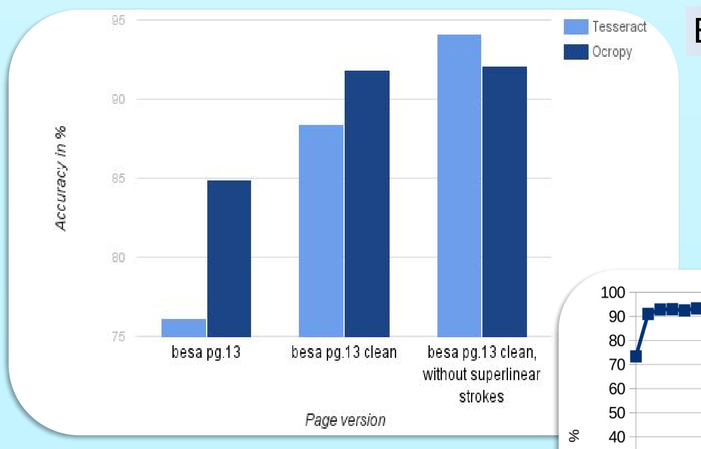
[Fragment 35] *A DENUNCIATION OF AN ERRING NUN*

. . . . ⸢ⲙ⸣ⲁ . . . . . ⸢ⲁ⸣ⲩⲱ ⸢ⲛ⸣ . . . . ⲟ · ⲏ [ⲛⲓⲙ ⲡ]⸢ⲉ⸣ⲧⲛ[ⲁⲁ]ⲯ-
[ⲁϩⲟ]ⲙ ⲉϩⲣⲁⲓ ⲉ[ⲭ]ⲱ · ⲏ ⲛⲓⲙ ⲡⲉ[ⲧⲛ]ⲁⲕⲧⲟϥ ⲛⲉ [ⲉ]ⲩⲉⲓⲣⲏⲛⲏ ·

Comparison between Tesseract and Ocropy

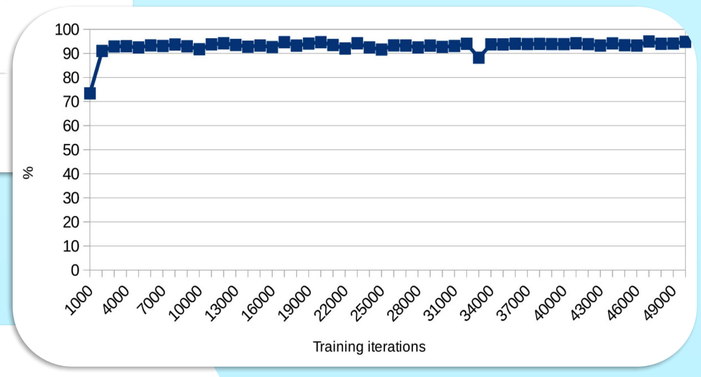


- Our test case: 1956 edition of the works of Besa, a fifth-century abbot in Upper Egypt, written in the Sahidic dialect of Coptic
- Mekhaiel's model: aimed at the Bohairic dialect (different type face, but containing all the diacritics and letters of Sahidic)
  1. With diacritics, punctuation, and editorial signs
  2. Without these

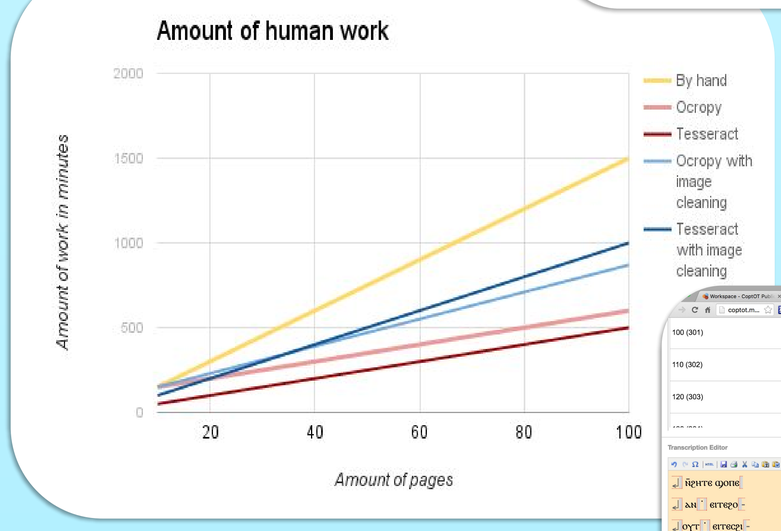


Evaluated by Ocrevalutf8

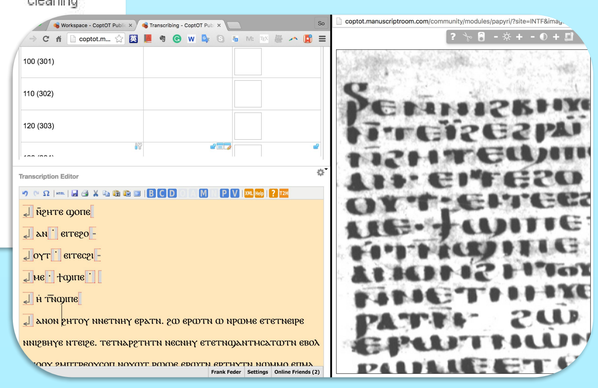
Accuracy of Ocropy



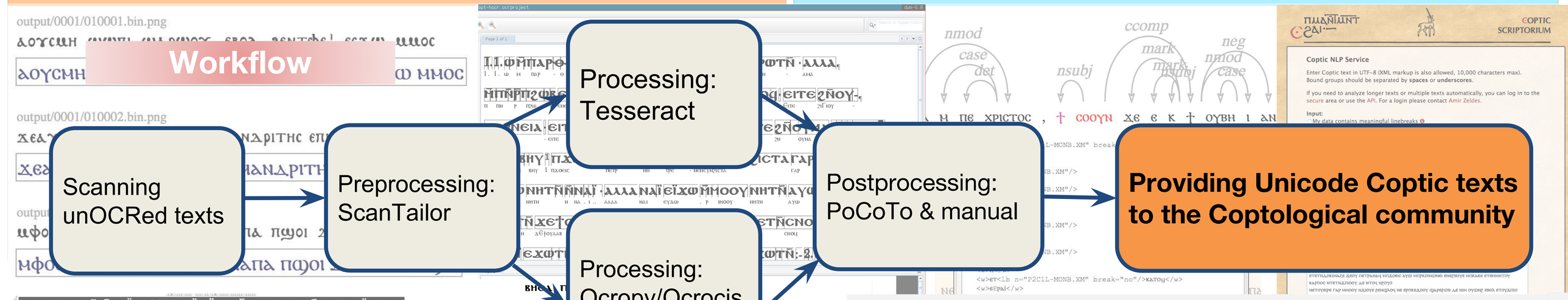
Toward less human power



Providing Unicode texts to Coptic DH projects



Workflow



Acknowledgements: Special thanks to Heike Behlmer, Uwe Springmann and Florian Fink. Also, thank you for the comments, corrections, and suggestions, Jane Akshar, Renate Dekker, Julien Delhez, Steven Feuer, George Ghaly, Marwan Kilani, Rebecca Krawiec, and Nicola Denzey Lewis.

Selected bibliography

- J. He, Q. D. M. Do, A. C. Downton, and J. H. Kim, "A comparison of binarization methods for historical archive documents," in Eighth International Conference on Document Analysis and Recognition (ICDAR'05), 2005, p. 538–542 Vol. 1.
- M. R. Gupta, N. P. Jacobson, and E. K. Garcia, "{OCR} binarization and image pre-processing for searching historical documents," Pattern Recognit., vol. 40, no. 2, pp. 389–397, 2007.
- V. Kluzner, A. Tzadok, Y. Shimony, E. Walach, and A. Antonacopoulos, "Word-Based Adaptiv OCR for Historical Books," in 2009 10th International Conference on Document Analysis and Recognition, 2009, pp. 501–505.
- U. Springmann, D. Najock, H. Morgenroth, H. Schmid, A. Gotscharek, and F. Fink, "OCR of Historical Printings of Latin Texts: Problems, Prospects, Progress," in Proceedings of the Firs International Conference on Digital Access to Textual Cultural Heritage, 2014, pp. 71–75.
- T. M. Breuel, "The OCRopus open source OCR system," Proc. SPIE 6815, Doc. Recognit. Retr. XV, 2008.