Optical Character Recognition with a Neural Network Model for Coptic

Kirill Bulert So Miyagawa Marco Buechler

December 8, 2017

DH2017 Montreal, Canada

Virtual Short Paper

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

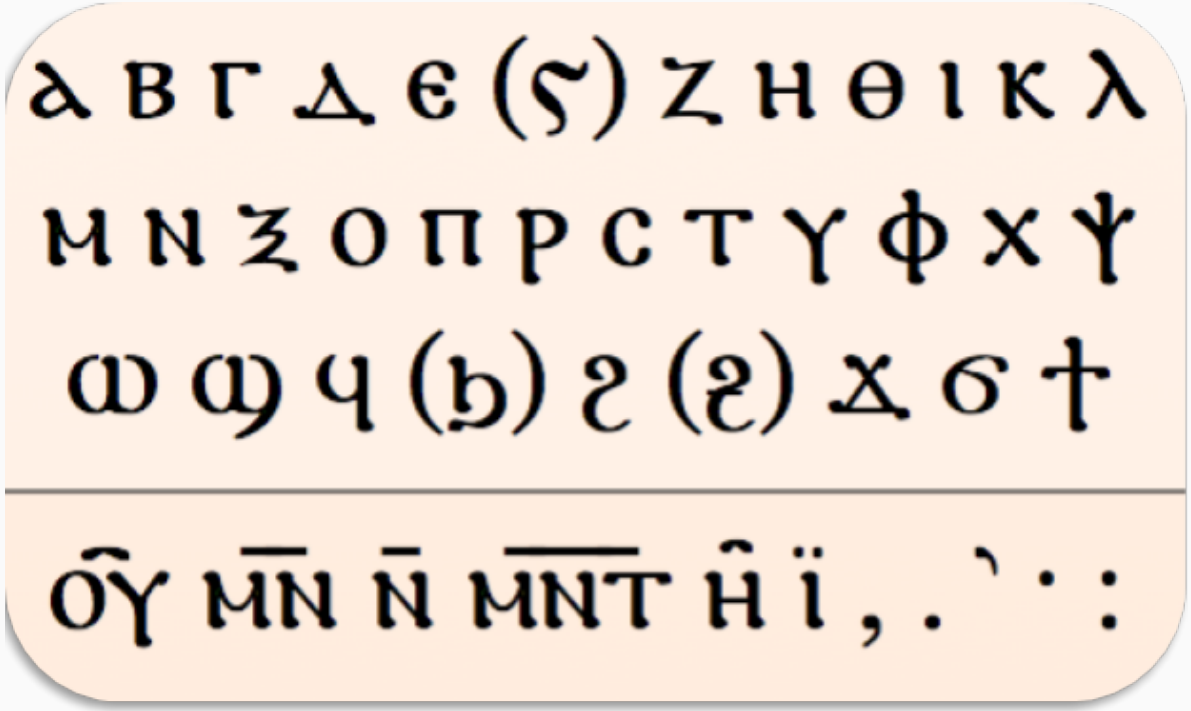
Coptic

About Coptic

- The final stage of the Ancient Egyptian language (third century)
- Multiple dialects (most important: Bohairic, Sahidic)
- Many manuscripts in Sahidic Coptic

Alphabet

- Simple: Ca. 30 unique characters
- No upper/lower case in historic texts
- Diacritics add some complexity, but not much
- Based on Greek alphabet



Coptic projects

- SFB 1136 (Goettingen)
  https://www.uni-goettingen.de/de/sfb-1136/521113.html
- Digital Edition of the Coptic old testament (Goettingen)
  http://coptot.manuscriptroom.com/
- Coptic SCRIPTORIUM (Georgetown/Pacific)
  http://copticscriptorium.org/

OCR Systems

Tesseract vs Ocropy, free OCR frameworks

Tesseract

- Font based learning
- Single characters are decomposed
- Decomposed parts are matched against given text

Input

- Fonts

Problem

- Few Coptic fonts available, not all historic variations covered

Ocropy/Ocropus

- Neural nets with online learning
- Model accuracy proportional to ground truth size
- But, size of ground truth is limited

Input

- Ground truth and corresponding images

Problem

- Limited data for learning

Tesseract vs Ocropy, free OCR frameworks

Tesseract

- Font based learning
- Single characters are decomposed
- Decomposed parts are matched against given text

Input

- Fonts

Problem

- Few Coptic fonts available, not all historic variations covered

Ocropy/Ocropus

- Neural nets with online learning
- Model accuracy proportional to ground truth size
- But, size of ground truth is limited

Input

- Ground truth and corresponding images

Problem

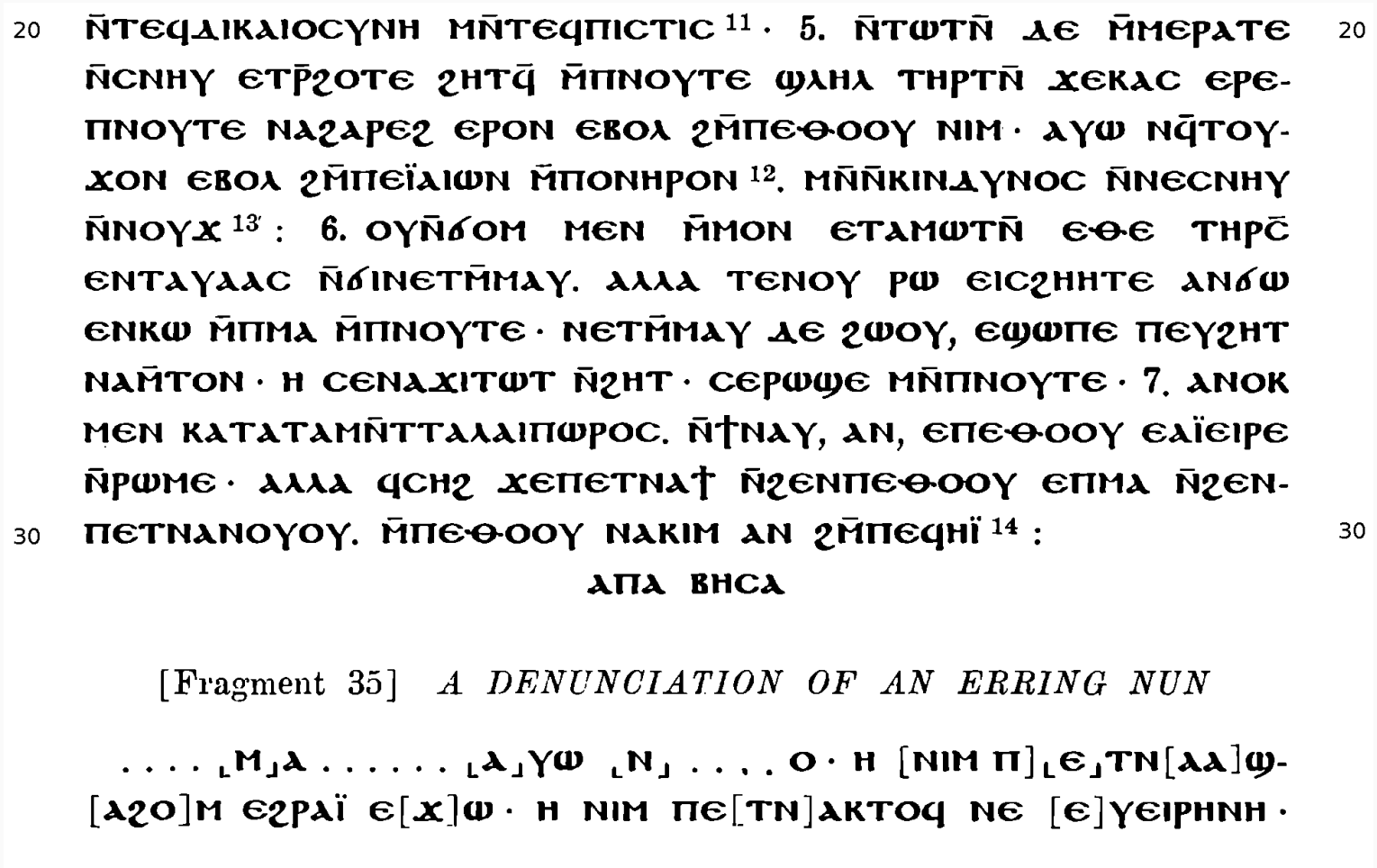
- Limited data for learning

Pre-Processing

GIGO - principle

- Training: Data $\rightarrow$ training $\rightarrow$ model
- Garbage in, garbage out
- Cleaner images improve results tremendously
- Dust and stains can be cleaned algorithmically, but ...
- What if text itself is ~~noise~~ the problem?

Original Page



ⲛ̄ⲧⲉϥⲇⲓⲕⲁⲓⲟⲥⲩⲛⲏ ⲙ̄ⲛ̄ⲧⲉϥⲡⲓⲥⲧⲓⲥ [11] · 5. ⲛ̄ⲧⲱⲧⲛ̄ ⲇⲉ ⲙ̄ⲙⲉⲣⲁⲧⲉ
ⲛ̄ⲥⲛⲏⲩ ⲉⲧⲣ̄ϩⲟⲧⲉ ϩⲛⲧ̄ⲕ̄ ⲙ̄ⲡⲛⲟⲩⲧⲉ ⲱⲗⲏⲗ ⲧⲏⲣⲧⲛ̄ ϫⲉⲕⲁⲥ ⲉⲣⲉ-
ⲡⲛⲟⲩⲧⲉ ⲛⲁϩⲁⲣⲉϩ ⲉⲣⲟⲛ ⲉⲃⲟⲗ ϩⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲓⲙ · ⲁⲩⲱ ⲛ̄ⲧⲟⲩ-
ϫⲟⲛ ⲉⲃⲟⲗ ϩⲙ̄ⲡⲉⲓⲁⲓⲱⲛ ⲙ̄ⲡⲟⲛⲏⲣⲟⲛ [12]. ⲙ̄ⲛ̄ⲛ̄ⲕⲓⲛⲇⲩⲛⲟⲥ ⲛ̄ⲛⲉⲥⲛⲏⲩ
ⲛ̄ⲛⲟⲩϫ [13] : 6. ⲟⲩⲛ̄ⲇⲟⲙ ⲙⲉⲛ ⲙ̄ⲙⲟⲛ ⲉⲧⲁⲙⲱⲧⲛ̄ ⲉⲑⲉ ⲧⲏⲣ̄
ⲉⲛⲧⲁⲩⲁⲁⲥ ⲛ̄ⲇⲓⲛⲉⲧⲙ̄ⲙⲁⲩ. ⲁⲗⲗⲁ ⲧⲉⲛⲟⲩ ⲣⲱ ⲉⲓⲥϩⲏⲧⲉ ⲁⲛⲇⲱ
ⲉⲛⲕⲱ ⲙ̄ⲡⲙⲁ ⲙ̄ⲡⲛⲟⲩⲧⲉ · ⲛⲉⲧⲙ̄ⲙⲁⲩ ⲇⲉ ϩⲱⲟⲩ, ⲉⲱⲱⲡⲉ ⲡⲉⲩϩⲏⲧ
ⲛⲁⲙ̄ⲧⲟⲛ · ⲏ ⲥⲉⲛⲁϫⲓⲧⲱⲧ ⲛ̄ϩⲏⲧ · ⲥⲉⲣⲱⲱⲉ ⲙ̄ⲛ̄ⲡⲛⲟⲩⲧⲉ · 7. ⲁⲛⲟⲕ
ⲙⲉⲛ ⲕⲁⲧⲁⲧⲁⲙ̄ⲛ̄ⲧⲁⲗⲁⲓⲡⲱⲣⲟⲥ. ⲛ̄ⲧⲛⲁⲩ, ⲁⲛ, ⲉⲡⲉⲑⲟⲟⲩ ⲉⲁⲓⲉⲓⲣⲉ
ⲛ̄ⲣⲱⲙⲉ · ⲁⲗⲗⲁ ⲫⲥϩ ϫⲉⲡⲉⲧⲛⲁ† ⲛ̄ϩⲉⲛⲡⲉⲑⲟⲟⲩ ⲉⲡⲙⲁ ⲛ̄ϩⲉⲛ-
ⲡⲉⲧⲛⲁⲛⲟⲩⲟⲩ. ⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲁⲕⲓⲙ ⲁⲛ ϩⲙ̄ⲡⲉⲫⲏⲓ [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[ϫ]ⲱ · ⲏ ⲛⲓⲙ ⲡⲉ[ⲧⲛ]ⲁⲕⲧⲟϥ ⲛⲉ [ⲉ]ⲩⲉⲓⲣⲏⲛⲏ ·

- Flawless, almost ...

Original Page

20 ⲛ̄ⲧⲉⲫⲇⲓⲕⲁⲓⲟⲥⲩⲛⲏ ⲙ̄ⲛ̄ⲧⲉⲫⲡⲓⲥⲧⲓⲥ [11] · 5. ⲛ̄ⲧⲱⲧⲛ̄ ⲇⲉ ⲙ̄ⲙⲉⲣⲁⲧⲉ 20
ⲛ̄ⲥⲛⲏⲩ ⲉⲧⲣ̄ϩⲟⲧⲉ ϩⲏⲧⲫ̄ ⲙ̄ⲡⲛⲟⲩⲧⲉ ⲱⲗⲏⲗ ⲧⲏⲣⲧⲛ̄ ⲭⲉⲕⲁⲥ ⲉⲣⲉ-
ⲡⲛⲟⲩⲧⲉ ⲛⲁϩⲁⲣⲉϩ ⲉⲣⲟⲛ ⲉⲃⲟⲗ ϩⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲓⲙ · ⲁⲩⲱ ⲛⲫ̄ⲧⲟⲩ-
ⲭⲟⲛ ⲉⲃⲟⲗ ϩⲙ̄ⲡⲉⲓⲁⲓⲱⲛ ⲙ̄ⲡⲟⲛⲏⲣⲟⲛ [12]. ⲙ̄ⲛ̄ⲛ̄ⲕⲓⲛⲇⲩⲛⲟⲥ ⲛ̄ⲛⲉⲥⲛⲏⲩ
ⲛ̄ⲛⲟⲩⲭ [13] : 6. ⲟⲩⲛ̄ⲅⲟⲙ ⲙⲉⲛ ⲙ̄ⲙⲟⲛ ⲉⲧⲁⲙⲱⲧⲛ̄ ⲉⲑⲉ ⲧⲏⲣ̄ⲥ̄
ⲉⲛⲧⲁⲩⲁⲁⲥ ⲛ̄ⲅⲓⲛⲉⲧⲙ̄ⲙⲁⲩ. ⲁⲗⲗⲁ ⲧⲉⲛⲟⲩ ⲣⲱ ⲉⲓⲥϩⲏⲏⲧⲉ ⲁⲛⲅⲱ
ⲉⲛⲕⲱ ⲙ̄ⲡⲙⲁ ⲙ̄ⲡⲛⲟⲩⲧⲉ · ⲛⲉⲧⲙ̄ⲙⲁⲩ ⲇⲉ ϩⲱⲟⲩ, ⲉⲱⲱⲡⲉ ⲡⲉⲩϩⲏⲧ
ⲛⲁⲙ̄ⲧⲟⲛ · ⲏ ⲥⲉⲛⲁⲭⲓⲧⲱⲧ ⲛ̄ϩⲏⲧ · ⲥⲉⲣⲱⲱⲉ ⲙ̄ⲛ̄ⲡⲛⲟⲩⲧⲉ · 7. ⲁⲛⲟⲕ
ⲙⲉⲛ ⲕⲁⲧⲁⲧⲁⲙⲛ̄ⲧⲧⲁⲗⲁⲓⲡⲱⲣⲟⲥ. ⲛ̄ⲧⲛⲁⲩ, ⲁⲛ, ⲉⲡⲉⲑⲟⲟⲩ ⲉⲁⲓⲉⲓⲣⲉ
ⲛ̄ⲣⲱⲙⲉ · ⲁⲗⲗⲁ ⲫⲥϩ ϩⲉⲡⲉⲧⲛⲁⲧ ⲛ̄ϩⲉⲛⲡⲉⲑⲟⲟⲩ ⲉⲡⲙⲁ ⲛ̄ϩⲉⲛ-
30 ⲡⲉⲧⲛⲁⲛⲟⲩⲟⲩ. ⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲁⲕⲓⲙ ⲁⲛ ϩⲙ̄ⲡⲉⲫⲏⲓ [14] : 30

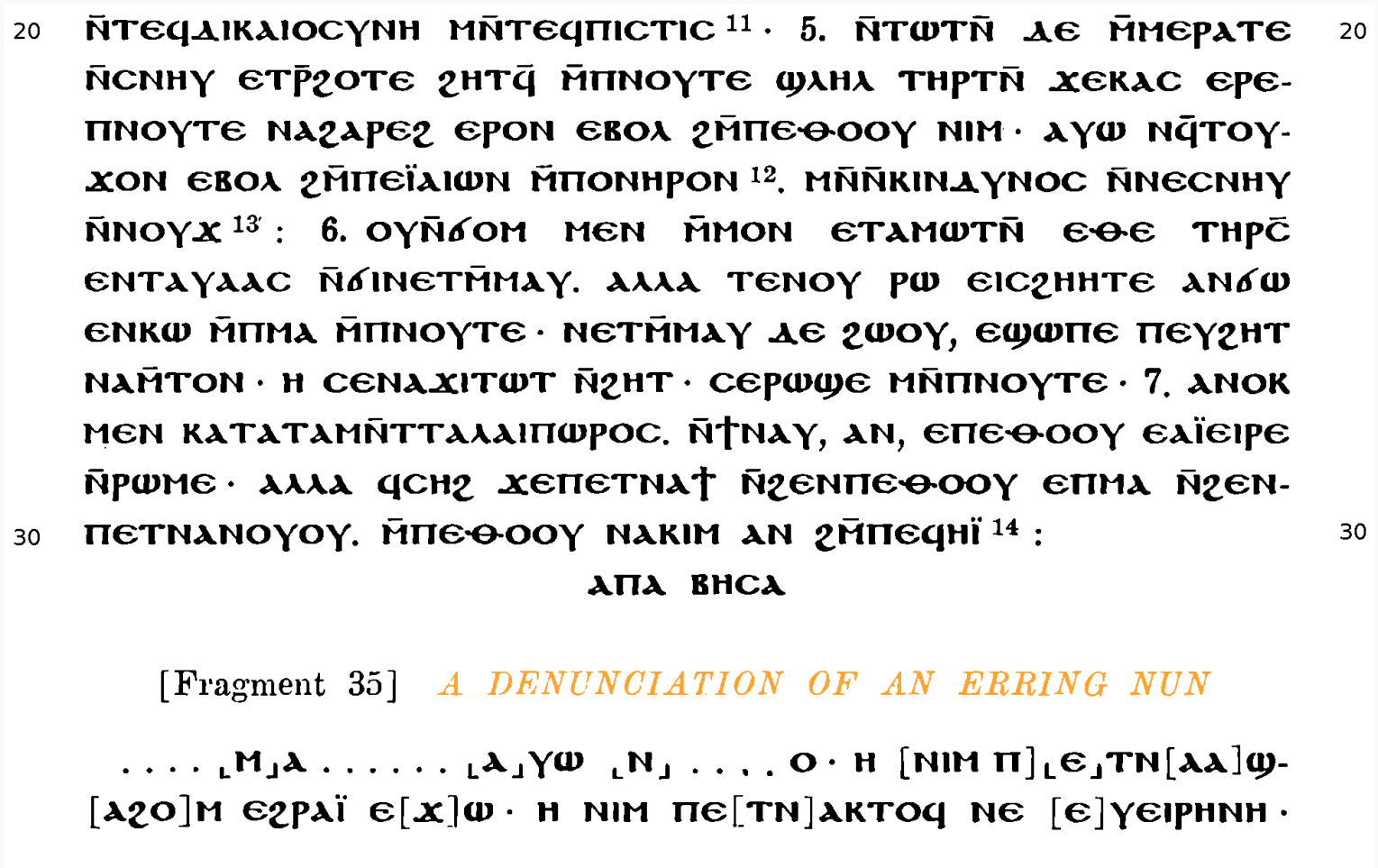
ⲁⲡⲁ ⲃⲏⲥⲁ

[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]ⲁⲕⲧⲟⲫ ⲛⲉ [ⲉ]ⲩⲉⲓⲣⲏⲛⲏ ·

- Easily removable with ScanTailor

Foreign language



20 ⲛ̄ⲧⲉϥⲇⲓⲕⲁⲓⲟⲥⲩⲛⲏ ⲙ̄ⲛ̄ⲧⲉϥⲡⲓⲥⲧⲓⲥ [11] · 5. ⲛ̄ⲧⲱⲧⲛ̄ ⲇⲉ ⲙ̄ⲙⲉⲣⲁⲧⲉ 20
ⲛ̄ⲥⲛⲏⲩ ⲉⲧⲣ̄ϩⲟⲧⲉ ϩⲛⲧⲫ̄ ⲙ̄ⲡⲛⲟⲩⲧⲉ ⲱⲗⲏⲗ ⲧⲏⲣⲧⲛ̄ ⲭⲉⲕⲁⲥ ⲉⲣⲉ-
ⲡⲛⲟⲩⲧⲉ ⲛⲁϩⲁⲣⲉϩ ⲉⲣⲟⲛ ⲉⲃⲟⲗ ϩⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲓⲙ · ⲁⲩⲱ ⲛⲫ̄ⲧⲟⲩ-
ⲭⲟⲛ ⲉⲃⲟⲗ ϩⲙ̄ⲡⲉⲓⲁⲓⲱⲛ ⲙ̄ⲡⲟⲛⲏⲣⲟⲛ [12]. ⲙ̄ⲛ̄ⲛ̄ⲕⲓⲛⲇⲩⲛⲟⲥ ⲛ̄ⲛⲉⲥⲛⲏⲩ
ⲛ̄ⲛⲟⲩⲭ [13] : 6. ⲟⲩⲛ̄ⲇⲟⲙ ⲙⲉⲛ ⲙ̄ⲙⲟⲛ ⲉⲧⲁⲙⲱⲧⲛ̄ ⲉⲑⲉ ⲧⲏⲣ̄ⲥ̄
ⲉⲛⲧⲁⲩⲁⲁⲥ ⲛ̄ⲇⲓⲛⲉⲧⲙ̄ⲙⲁⲩ. ⲁⲗⲗⲁ ⲧⲉⲛⲟⲩ ⲣⲱ ⲉⲓⲥϩⲏⲧⲉ ⲁⲛⲇⲱ
ⲉⲛⲕⲱ ⲙ̄ⲡⲙⲁ ⲙ̄ⲡⲛⲟⲩⲧⲉ · ⲛⲉⲧⲙ̄ⲙⲁⲩ ⲇⲉ ϩⲱⲟⲩ, ⲉⲫⲱⲡⲉ ⲡⲉⲩϩⲏⲧ
ⲛⲁⲙ̄ⲧⲟⲛ · ⲏ ⲥⲉⲛⲁⲭⲓⲧⲱⲧ ⲛ̄ϩⲏⲧ · ⲥⲉⲣⲱⲫⲉ ⲙ̄ⲛ̄ⲡⲛⲟⲩⲧⲉ · 7. ⲁⲛⲟⲕ
ⲙⲉⲛ ⲕⲁⲧⲁⲧⲁⲙⲛ̄ⲧⲧⲁⲗⲁⲓⲡⲱⲣⲟⲥ. ⲛ̄ⲧⲛⲁⲩ, ⲁⲛ, ⲉⲡⲉⲑⲟⲟⲩ ⲉⲁⲓⲉⲓⲣⲉ
ⲛ̄ⲣⲱⲙⲉ · ⲁⲗⲗⲁ ⲫⲥϩ ϩⲉⲡⲉⲧⲛⲁϯ ⲛ̄ϩⲉⲛⲡⲉⲑⲟⲟⲩ ⲉⲡⲙⲁ ⲛ̄ϩⲉⲛ-
30 ⲡⲉⲧⲛⲁⲛⲟⲩⲟⲩ. ⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲁⲕⲓⲙ ⲁⲛ ϩⲙ̄ⲡⲉⲫⲏⲓ [14] : 30

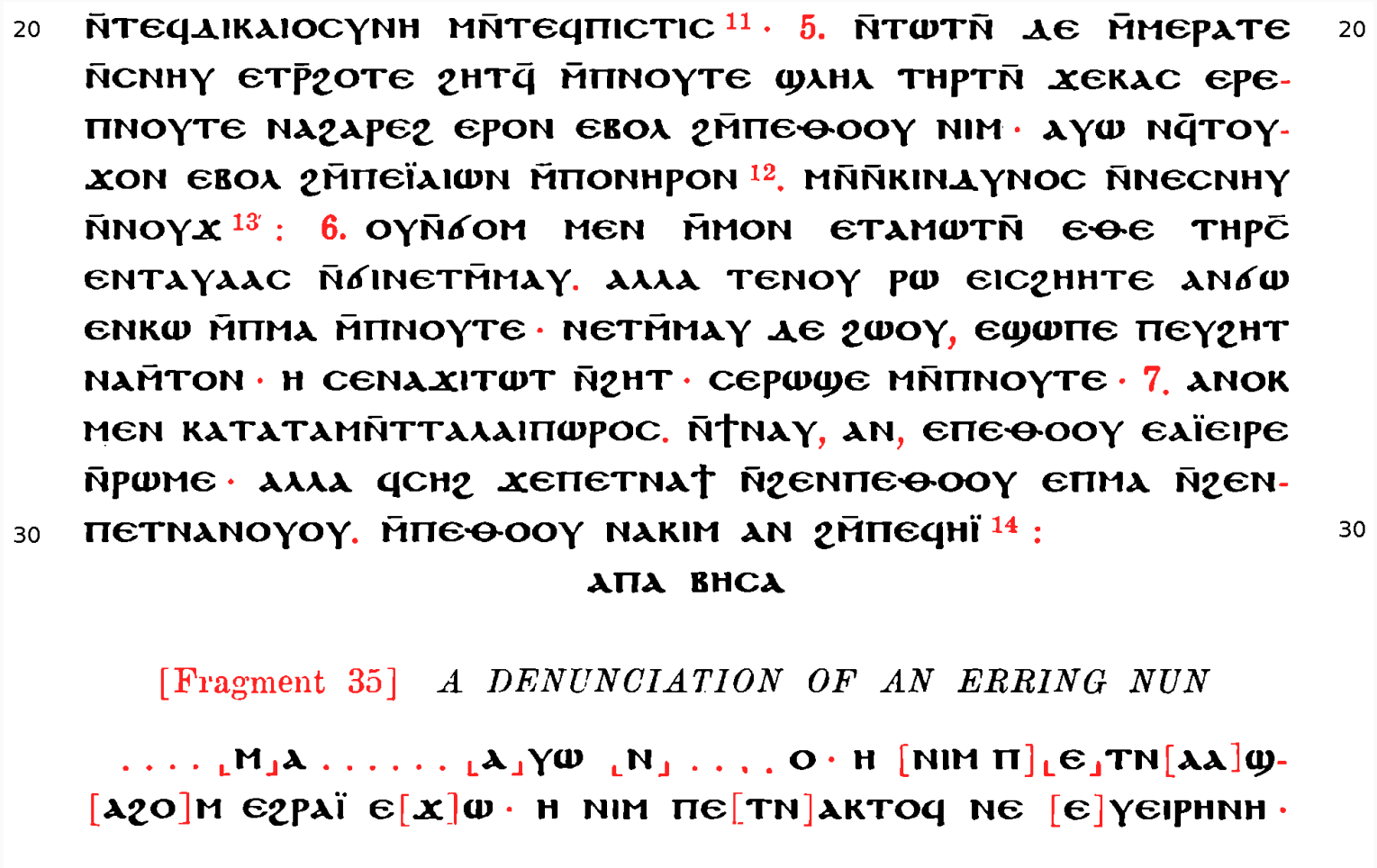
ⲁⲡⲁ ⲃⲏⲥⲁ

[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]ⲁⲕⲧⲟⲫ ⲛⲉ [ⲉ]ⲩⲉⲓⲣⲏⲛⲏ ·

- No multilingual OCR models for Coptic

Annotations


20 ⲛ̄ⲧⲉϥⲇⲓⲕⲁⲓⲟⲥⲩⲛⲏ ⲙ̄ⲛ̄ⲧⲉϥⲡⲓⲥⲧⲓⲥ [11] · 5. ⲛ̄ⲧⲱⲧⲛ̄ ⲇⲉ ⲙ̄ⲙⲉⲣⲁⲧⲉ 20
ⲛ̄ⲥⲛⲏⲩ ⲉⲧⲣ̄ϩⲟⲧⲉ ϩⲛⲧⲫ̄ ⲙ̄ⲡⲛⲟⲩⲧⲉ ⲱⲗⲏⲗ ⲧⲏⲣⲧⲛ̄ ⲭⲉⲕⲁⲥ ⲉⲣⲉ-
ⲡⲛⲟⲩⲧⲉ ⲛⲁϩⲁⲣⲉϩ ⲉⲣⲟⲛ ⲉⲃⲟⲗ ϩⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲓⲙ · ⲁⲩⲱ ⲛⲫ̄ⲧⲟⲩ-
ⲭⲟⲛ ⲉⲃⲟⲗ ϩⲙ̄ⲡⲉⲓⲁⲓⲱⲛ ⲙ̄ⲡⲟⲛⲏⲣⲟⲛ [12]. ⲙ̄ⲛ̄ⲛ̄ⲕⲓⲛⲇⲩⲛⲟⲥ ⲛ̄ⲛⲉⲥⲛⲏⲩ
ⲛ̄ⲛⲟⲩⲭ [13] : 6. ⲟⲩⲛ̄ⲅⲟⲙ ⲙⲉⲛ ⲙ̄ⲙⲟⲛ ⲉⲧⲁⲙⲱⲧⲛ̄ ⲉⲑⲉ ⲧⲏⲣ̄ⲥ̄
ⲉⲛⲧⲁⲩⲁⲁⲥ ⲛ̄ⲅⲓⲛⲉⲧⲙ̄ⲙⲁⲩ. ⲁⲗⲗⲁ ⲧⲉⲛⲟⲩ ⲣⲱ ⲉⲓⲥϩⲏⲧⲉ ⲁⲛⲅⲱ
ⲉⲛⲕⲱ ⲙ̄ⲡⲙⲁ ⲙ̄ⲡⲛⲟⲩⲧⲉ · ⲛⲉⲧⲙ̄ⲙⲁⲩ ⲇⲉ ϩⲱⲟⲩ, ⲉⲱⲱⲡⲉ ⲡⲉⲩϩⲏⲧ
ⲛⲁⲙ̄ⲧⲟⲛ · ⲏ ⲥⲉⲛⲁⲭⲓⲧⲱⲧ ⲛ̄ϩⲏⲧ · ⲥⲉⲣⲱⲱⲉ ⲙ̄ⲛ̄ⲡⲛⲟⲩⲧⲉ · 7. ⲁⲛⲟⲕ
ⲙⲉⲛ ⲕⲁⲧⲁⲧⲁⲙⲛ̄ⲧⲧⲁⲗⲁⲓⲡⲱⲣⲟⲥ. ⲛ̄ⲧⲛⲁⲩ, ⲁⲛ, ⲉⲡⲉⲑⲟⲟⲩ ⲉⲁⲩⲉⲓⲣⲉ
ⲛ̄ⲣⲱⲙⲉ · ⲁⲗⲗⲁ ⲫⲥϩϩ ⲭⲉⲡⲉⲧⲛⲁ† ⲛ̄ϩⲉⲛⲡⲉⲑⲟⲟⲩ ⲉⲡⲙⲁ ⲛ̄ϩⲉⲛ-
30 ⲡⲉⲧⲛⲁⲛⲟⲩⲟⲩ. ⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲁⲕⲓⲙ ⲁⲛ ϩⲙ̄ⲡⲉⲫⲏⲓ [14] : 30

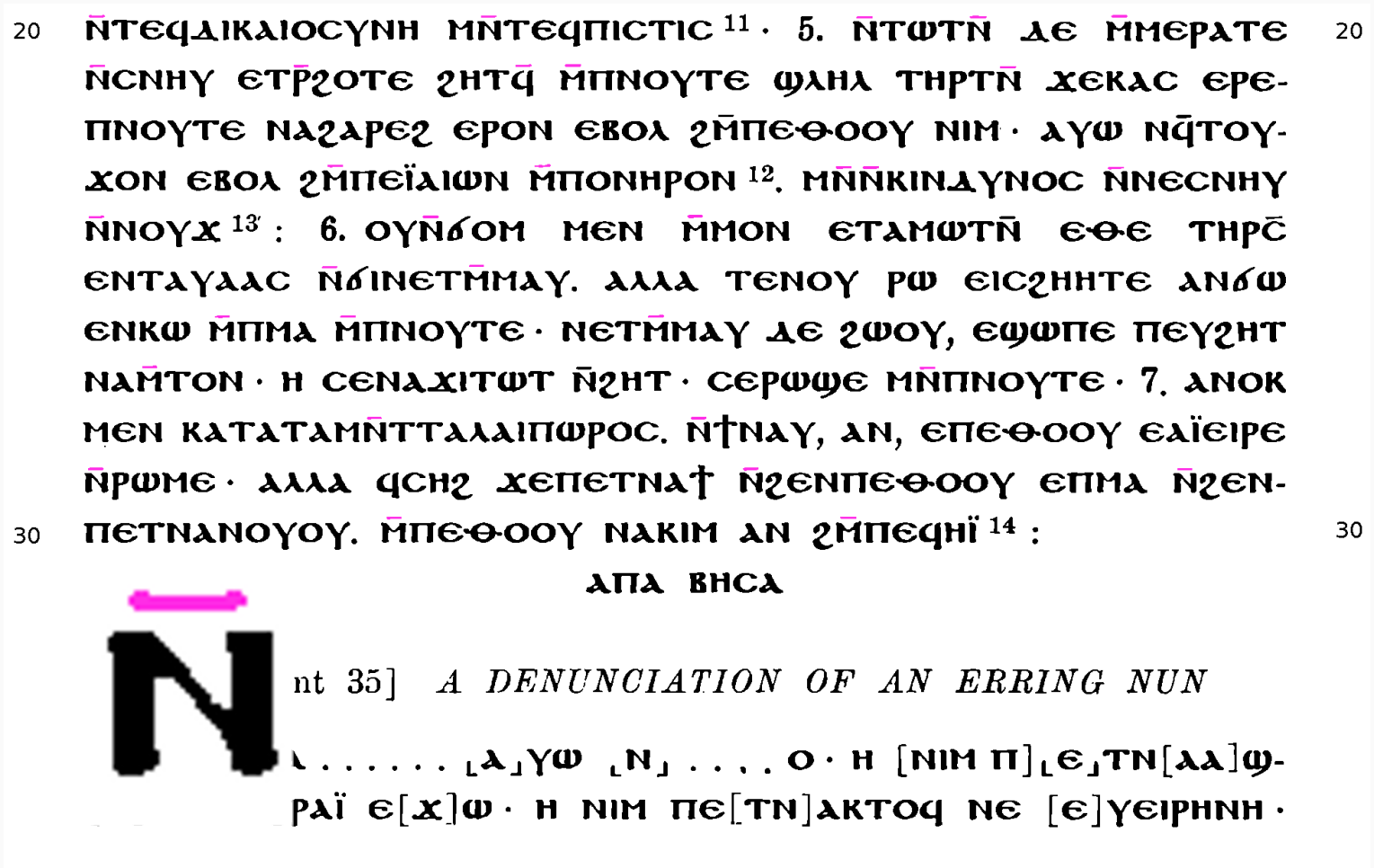
ⲁⲡⲁ ⲃⲏⲥⲁ

[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]ⲁⲕⲧⲟⲫ ⲛⲉ [ⲉ]ⲩⲉⲓⲣⲏⲛⲏ ·

- Special characters might not be part of any model (⸢ ⸣)
- Not all annotations wanted

Language specific variations


20 ⲛ̄ⲧⲉϥⲇⲓⲕⲁⲓⲟⲥⲩⲛⲏ ⲙ̄ⲛ̄ⲧⲉϥⲡⲓⲥⲧⲓⲥ [11] · 5. ⲛ̄ⲧⲱⲧⲛ̄ ⲇⲉ ⲙ̄ⲙⲉⲣⲁⲧⲉ 20
ⲛ̄ⲥⲛⲏⲩ ⲉⲧⲣ̄ϩⲟⲧⲉ ϩⲏⲧϥ ⲙ̄ⲡⲛⲟⲩⲧⲉ ⲱⲗⲏⲗ ⲧⲏⲣⲧⲛ̄ ⲭⲉⲕⲁⲥ ⲉⲣⲉ-
ⲡⲛⲟⲩⲧⲉ ⲛⲁϩⲁⲣⲉϩ ⲉⲣⲟⲛ ⲉⲃⲟⲗ ϩⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲓⲙ · ⲁⲩⲱ ⲛϥ̄ⲧⲟⲩ-
ⲭⲟⲛ ⲉⲃⲟⲗ ϩⲙ̄ⲡⲉⲓⲁⲓⲱⲛ ⲙ̄ⲡⲟⲛⲏⲣⲟⲛ [12]. ⲙ̄ⲛ̄ⲛⲕⲓⲛⲇⲩⲛⲟⲥ ⲛ̄ⲛⲉⲥⲛⲏⲩ
ⲛ̄ⲛⲟⲩⲭ [13] : 6. ⲟⲩⲛ̄ⲇⲟⲙ ⲙⲉⲛ ⲙ̄ⲙⲟⲛ ⲉⲧⲁⲙⲱⲧⲛ̄ ⲉⲑⲉ ⲧⲏⲣ̄ⲥ̄
ⲉⲛⲧⲁⲩⲁⲁⲥ ⲛ̄ⲇⲓⲛⲉⲧⲙ̄ⲙⲁⲩ. ⲁⲗⲗⲁ ⲧⲉⲛⲟⲩ ⲣⲱ ⲉⲓⲥϩⲏⲧⲉ ⲁⲛⲇⲱ
ⲉⲛⲕⲱ ⲙ̄ⲡⲙⲁ ⲙ̄ⲡⲛⲟⲩⲧⲉ · ⲛⲉⲧⲙ̄ⲙⲁⲩ ⲇⲉ ϩⲱⲟⲩ, ⲉⲱⲱⲡⲉ ⲡⲉⲩϩⲏⲧ
ⲛⲁⲙ̄ⲧⲟⲛ · ⲏ ⲥⲉⲛⲁⲭⲓⲧⲱⲧ ⲛ̄ϩⲏⲧ · ⲥⲉⲣⲱⲱⲉ ⲙ̄ⲛ̄ⲡⲛⲟⲩⲧⲉ · 7. ⲁⲛⲟⲕ
ⲙⲉⲛ ⲕⲁⲧⲁⲧⲁⲙⲛ̄ⲧⲧⲁⲗⲁⲓⲡⲱⲣⲟⲥ. ⲛ̄ⲧⲛⲁⲩ, ⲁⲛ, ⲉⲡⲉⲑⲟⲟⲩ ⲉⲁⲓⲉⲓⲣⲉ
ⲛ̄ⲣⲱⲙⲉ · ⲁⲗⲗⲁ ⲫⲥⲏϩ ⲭⲉⲡⲉⲧⲛⲁϯ ⲛ̄ϩⲉⲛⲡⲉⲑⲟⲟⲩ ⲉⲡⲙⲁ ⲛ̄ϩⲉⲛ-
30 ⲡⲉⲧⲛⲁⲛⲟⲩⲟⲩ. ⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲁⲕⲓⲙ ⲁⲛ ϩⲙ̄ⲡⲉϥⲏⲓ [14] : 30

ⲁⲡⲁ ⲃⲏⲥⲁ

nt 35] *A DENUNCIATION OF AN ERRING NUN*

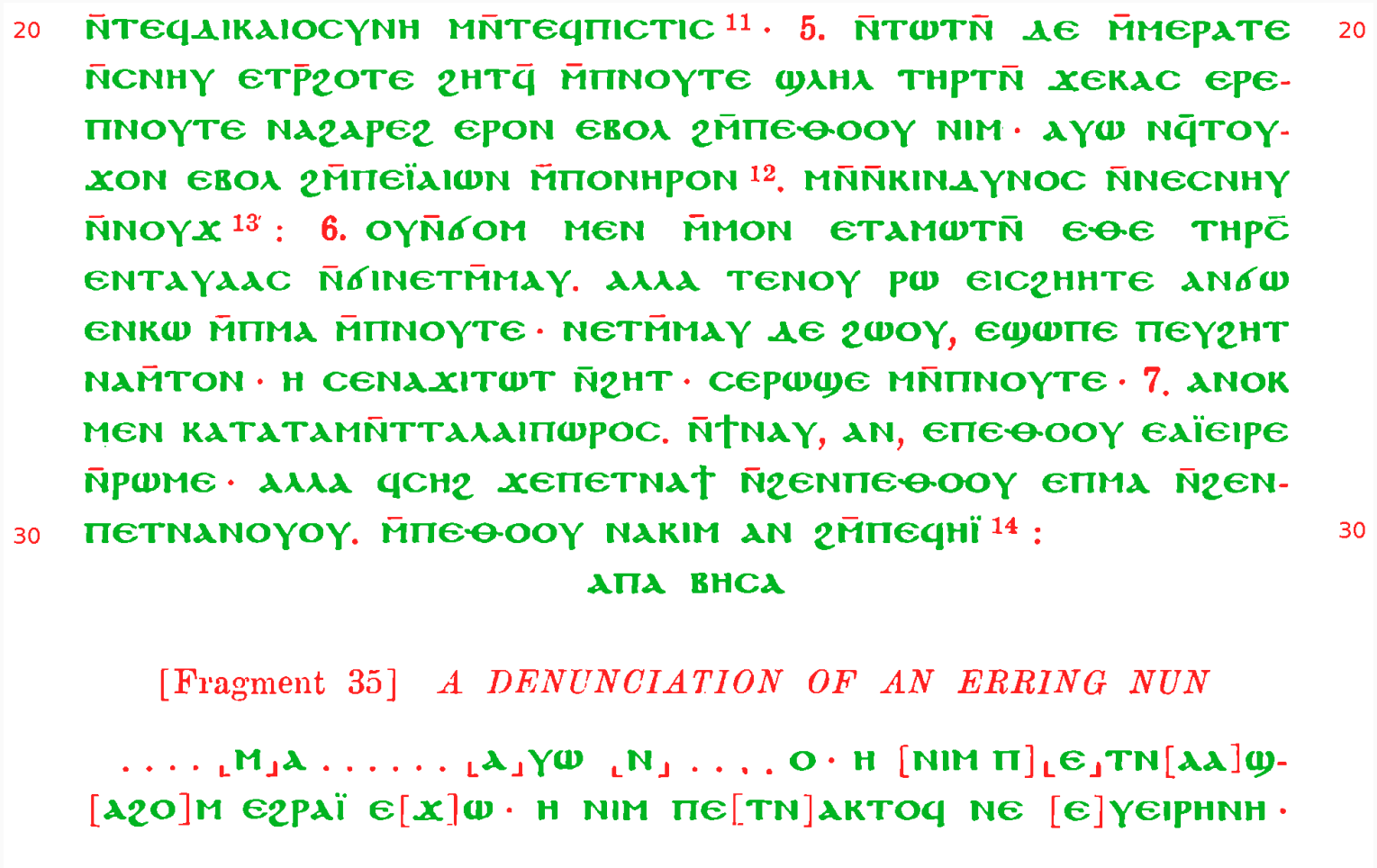
ⲗ . . . . . ⲁ ⲩⲱ ⲛ . . . . ⲟ · ⲏ [ⲛⲓⲙ ⲡ]ⲉⲧⲛ[ⲁⲁ]ⲱ-
ⲣⲁⲓ ⲉ[ⲭ]ⲱ · ⲏ ⲛⲓⲙ ⲡⲉ[ⲧⲛ]ⲁⲕⲧⲟϥ ⲛⲉ [ⲉ]ⲩⲉⲓⲣⲏⲛⲏ ·

- Might also not be included in a model

Previous work

- Coptic models created by Moheb for Tesseract (2013)
- Trained with several Coptic fonts, no non-Coptic letters support
- No support for diacritics
- Non-Coptic letters get replaced with similar Coptic letters

The good, the bad, and the problematic



20 ⲛ̄ⲧⲉϥⲇⲓⲕⲁⲓⲟⲥⲩⲛⲏ ⲙ̄ⲛ̄ⲧⲉϥⲡⲓⲥⲧⲓⲥ [11] · 5. ⲛ̄ⲧⲱⲧⲛ̄ ⲇⲉ ⲙ̄ⲙⲉⲣⲁⲧⲉ 20
ⲛ̄ⲥⲛⲏⲩ ⲉⲧⲣ̄ϩⲟⲧⲉ ϩⲛⲧⲫ̄ ⲙ̄ⲡⲛⲟⲩⲧⲉ ⲱⲗⲏⲗ ⲧⲏⲣⲧⲛ̄ ⲭⲉⲕⲁⲥ ⲉⲣⲉ-
ⲡⲛⲟⲩⲧⲉ ⲛⲁϩⲁⲣⲉϩ ⲉⲣⲟⲛ ⲉⲃⲟⲗ ϩⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲓⲙ · ⲁⲩⲱ ⲛⲫ̄ⲧⲟⲩ-
ⲭⲟⲛ ⲉⲃⲟⲗ ϩⲙ̄ⲡⲉⲓⲁⲓⲱⲛ ⲙ̄ⲡⲟⲛⲏⲣⲟⲛ [12]. ⲙ̄ⲛ̄ⲛ̄ⲕⲓⲛⲇⲩⲛⲟⲥ ⲛ̄ⲛⲉⲥⲛⲏⲩ
ⲛ̄ⲛⲟⲩⲭ [13] : 6. ⲟⲩⲛ̄ⲇⲟⲙ ⲙⲉⲛ ⲙ̄ⲙⲟⲛ ⲉⲧⲁⲙⲱⲧⲛ̄ ⲉⲑⲉ ⲧⲏⲣ̄
ⲉⲛⲧⲁⲩⲁⲁⲥ ⲛ̄ⲇⲓⲛⲉⲧⲙ̄ⲙⲁⲩ. ⲁⲗⲗⲁ ⲧⲉⲛⲟⲩ ⲣⲱ ⲉⲓⲥϩⲏⲧⲉ ⲁⲛⲇⲱ
ⲉⲛⲕⲱ ⲙ̄ⲡⲙⲁ ⲙ̄ⲡⲛⲟⲩⲧⲉ · ⲛⲉⲧⲙ̄ⲙⲁⲩ ⲇⲉ ϩⲱⲟⲩ, ⲉⲱⲱⲡⲉ ⲡⲉⲩϩⲏⲧ
ⲛⲁⲙ̄ⲧⲟⲛ · ⲏ ⲥⲉⲛⲁⲭⲓⲧⲱⲧ ⲛ̄ϩⲏⲧ · ⲥⲉⲣⲱⲱⲉ ⲙ̄ⲛ̄ⲡⲛⲟⲩⲧⲉ · 7. ⲁⲛⲟⲕ
ⲙⲉⲛ ⲕⲁⲧⲁⲧⲁⲙⲛ̄ⲧⲧⲁⲗⲁⲓⲡⲱⲣⲟⲥ. ⲛ̄ⲧⲛⲁⲩ, ⲁⲛ, ⲉⲡⲉⲑⲟⲟⲩ ⲉⲁⲓⲉⲓⲣⲉ
ⲛ̄ⲣⲱⲙⲉ · ⲁⲗⲗⲁ ⲫⲥϩ ϩⲉⲡⲉⲧⲛⲁ† ⲛ̄ϩⲉⲛⲡⲉⲑⲟⲟⲩ ⲉⲡⲙⲁ ⲛ̄ϩⲉⲛ-
30 ⲡⲉⲧⲛⲁⲛⲟⲩⲟⲩ. ⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲁⲕⲓⲙ ⲁⲛ ϩⲙ̄ⲡⲉϥⲏⲓ [14] : 30

ⲁⲡⲁ ⲃⲏⲥⲁ

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

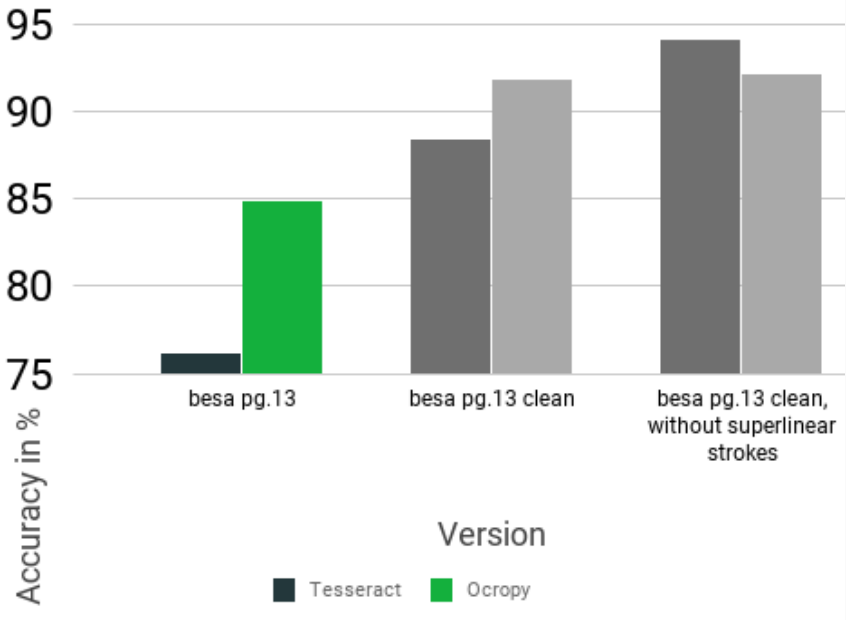
. . . . ⌞M⌟A . . . . . ⌞A⌟YW ⌞N⌟ . . . O · H [NIM P]⌞E⌟TN[AA]Ψ-
[AϩO]M ЄϩPAÏ Є[X]W · H NIM ПЄ[TN]AKTOϥ NЄ [Є]YЄIPHNH ·

- Even clean scans still contain noise

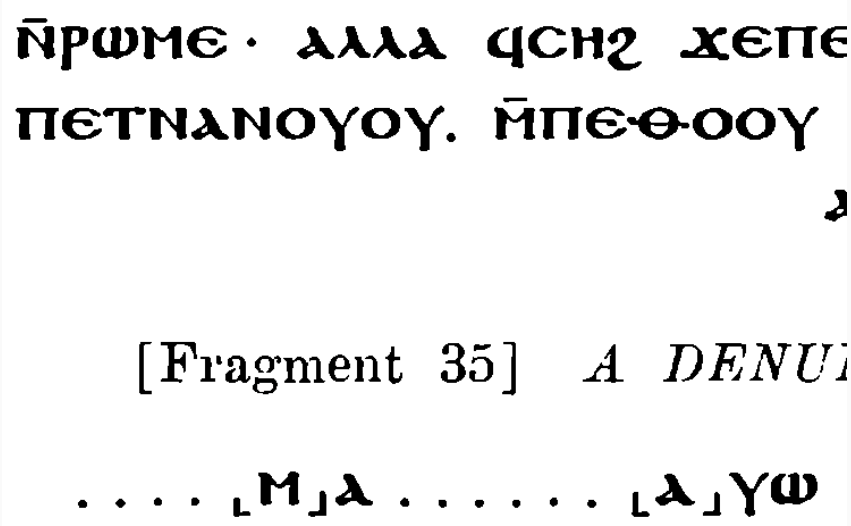
Results

Without line numbers

Accuracy



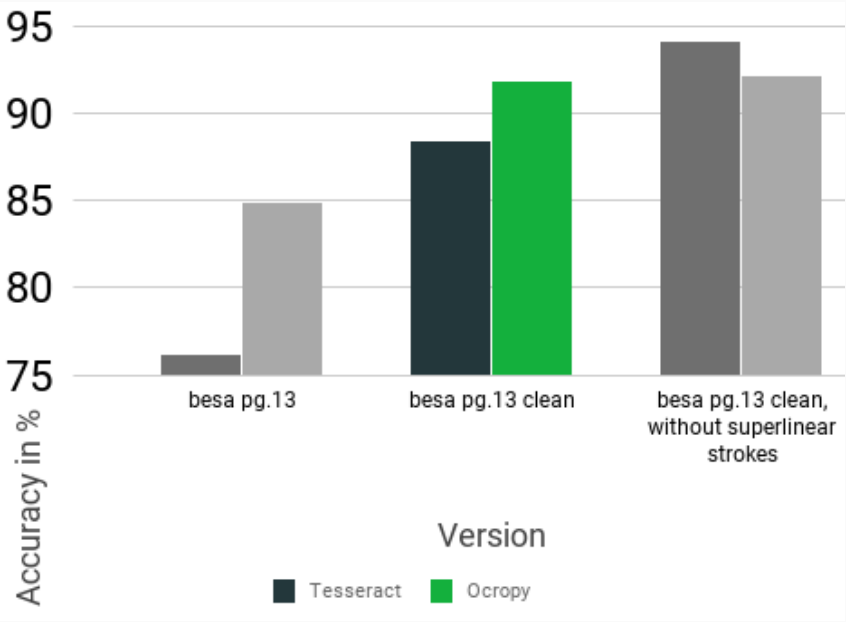
Input



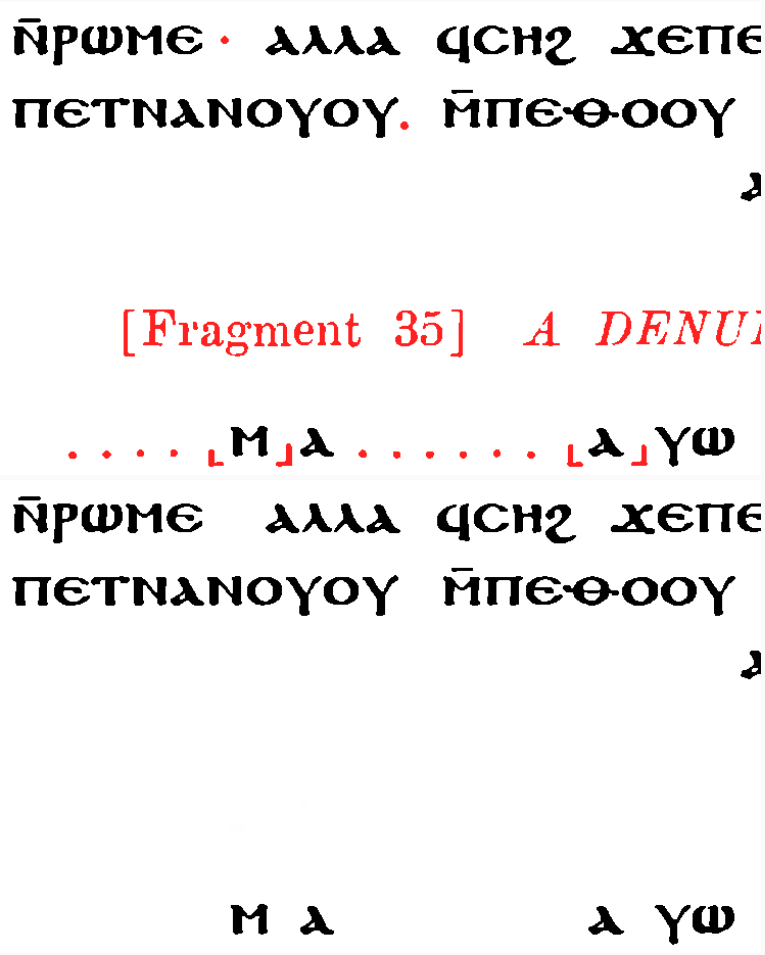
- Without time consuming pre-processing
- **Ocropy** model trained on 10 pages more accurate
- Non-multilingual **Tesseract** model less accurate

Without line numbers

Accuracy



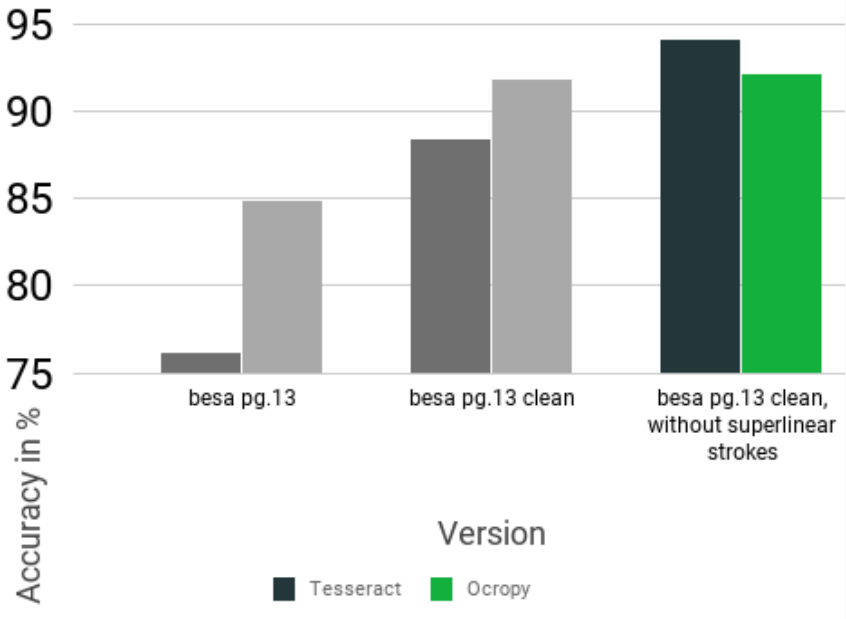
Input



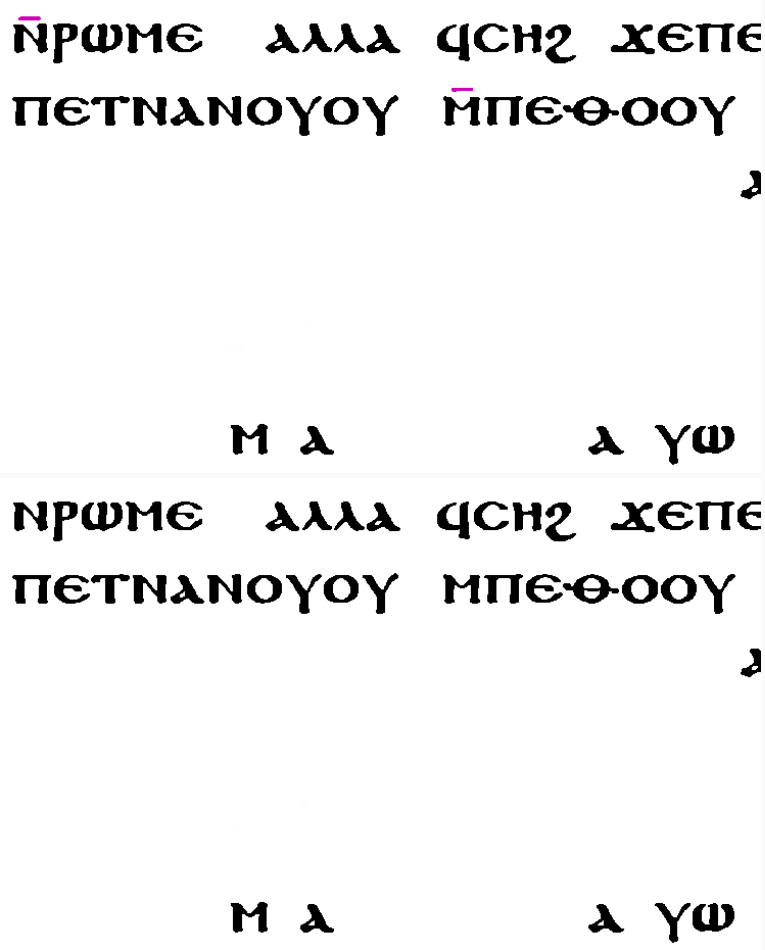
- Without non-Coptic letters
- Difference results mostly from diacritics

Without line numbers

Accuracy

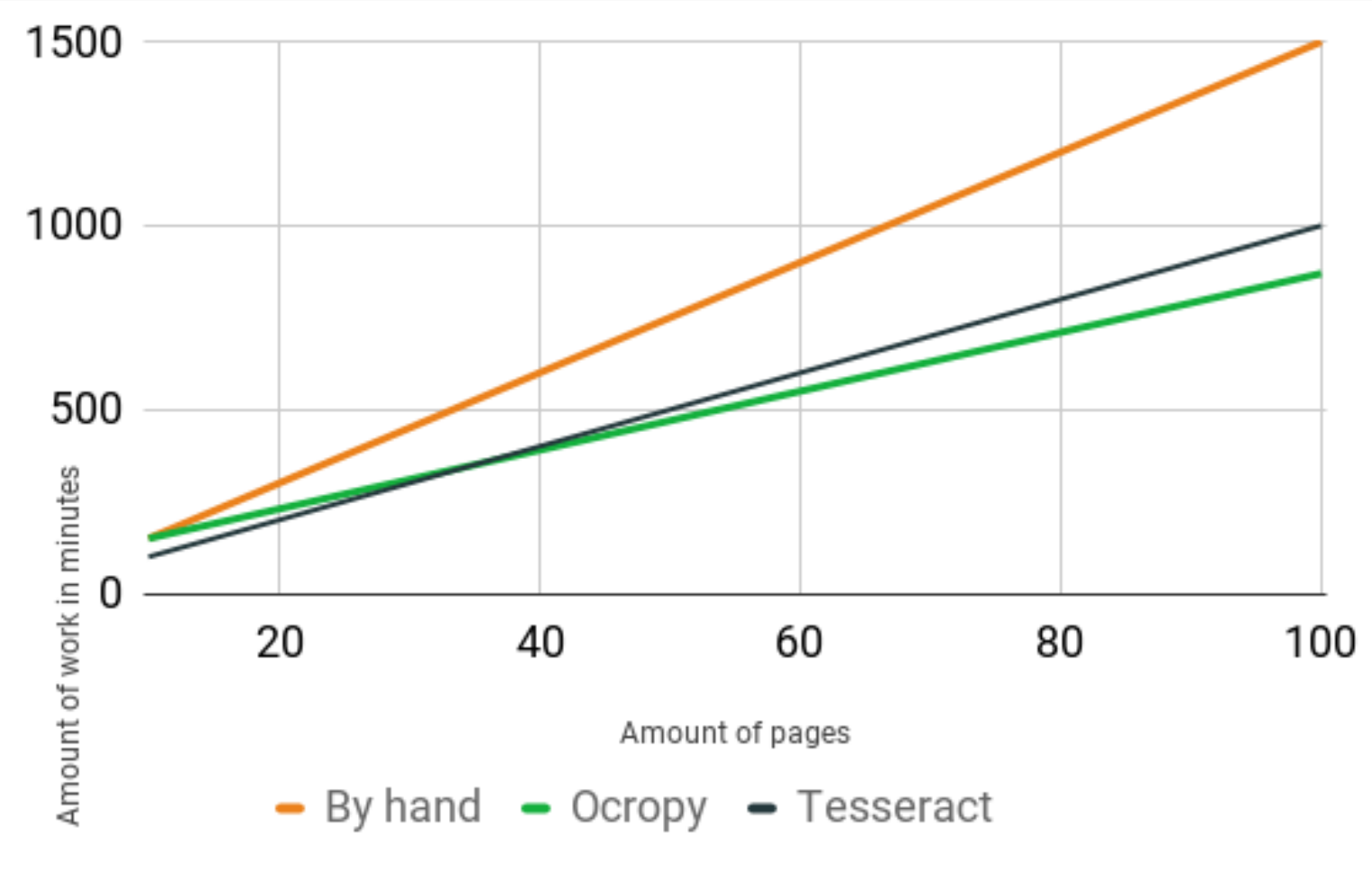


Input



- Diacritics removed
- Pure Coptic **Tesseract** model outperforms **Ocropys** mixed model

Workload comparison



- Utilising OCR decreases human workload

Roundup

- Utilisation of OCR beneficial for most clean documents
- **Tesseract** best for monolingual documents with limited fonts and font variations
- **Ocropy** best for large documents with multiple languages

Road ahead

- Approached by Google for collaboration on OCR for Coptic
- Create data set for Coptic OCR testing
- Transition from typeset to handwritten Coptic texts
- Combination of different models

Thank you.

Questions?

Get our Ocropy models at

Presentation

Kirill Bulert, So Miyagawa

Team (in alphabetical order)

Kirill Bulert, Marco Büchler, So Miyagawa.

Visit us

http://www.etrap.eu

http://www.uni-goettingen.de/de/517150.html

contact@etrap.eu

smiyaga@gwdg.de

eTRAP
Electronic Text Reuse Acquisition Project
INSTITUTE OF COMPUTER SCIENCE
GÖTTINGEN CENTRE FOR DIGITAL HUMANITIES

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

SPONSORED BY THE

Federal Ministry
of Education
and Research

KELLIA

NATIONAL
ENDOWMENT
FOR THE
HUMANITIES

BILDUNG
UND
RELIGION
SFB 1136

DFG Deutsche
Forschungsgemeinschaft

OCR References

1. Uwe Springmanns OCR Workshop
   http://www.cis.uni-muenchen.de/ocrworkshop/program.html
2. Scantailor for pre-processing
   http://scantailor.org/
3. Ocropy/Ocropus
   https://github.com/tmbdev/ocropy
4. Kraken an Ocropy fork
   http://kraken.re/
5. Tesseract OCR
   https://github.com/tesseract-ocr/
6. Moheb's Coptic Pages
   http://www.moheb.de/ocr.html

Coptic References

1. SFB 1136 (Goettingen)
   https://www.uni-goettingen.de/de/sfb-1136/521113.html
2. Digital Edition of the Coptic old testament (Goettingen)
   http://coptot.manuscriptroom.com/
3. Coptic SCRIPTORIUM (Georgetown/Pacific)
   http://copticscriptorium.org/

Licence