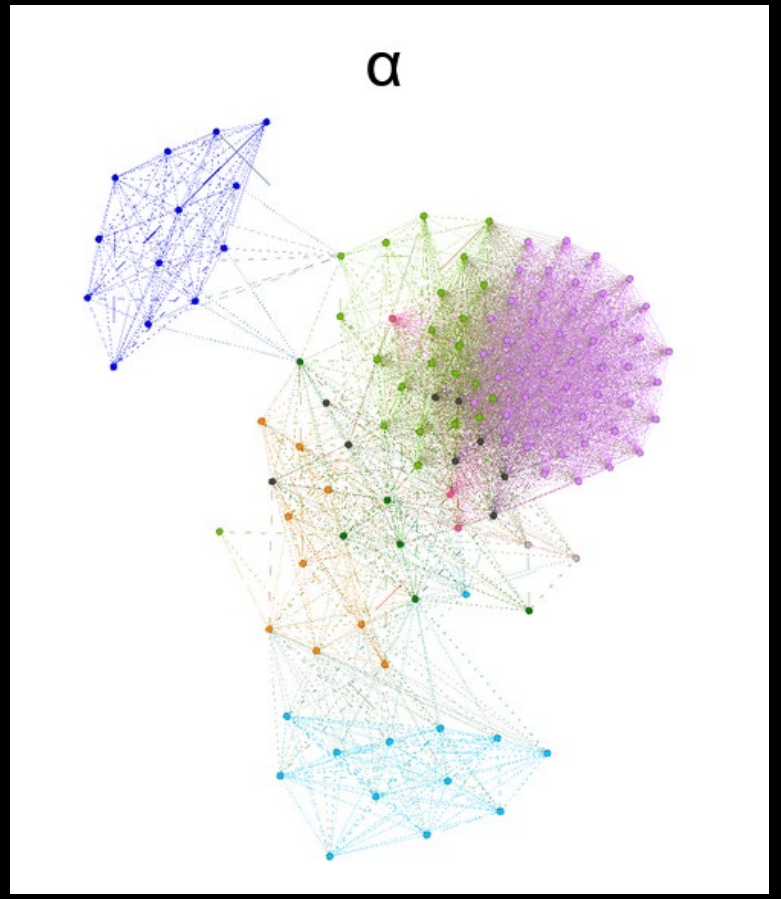
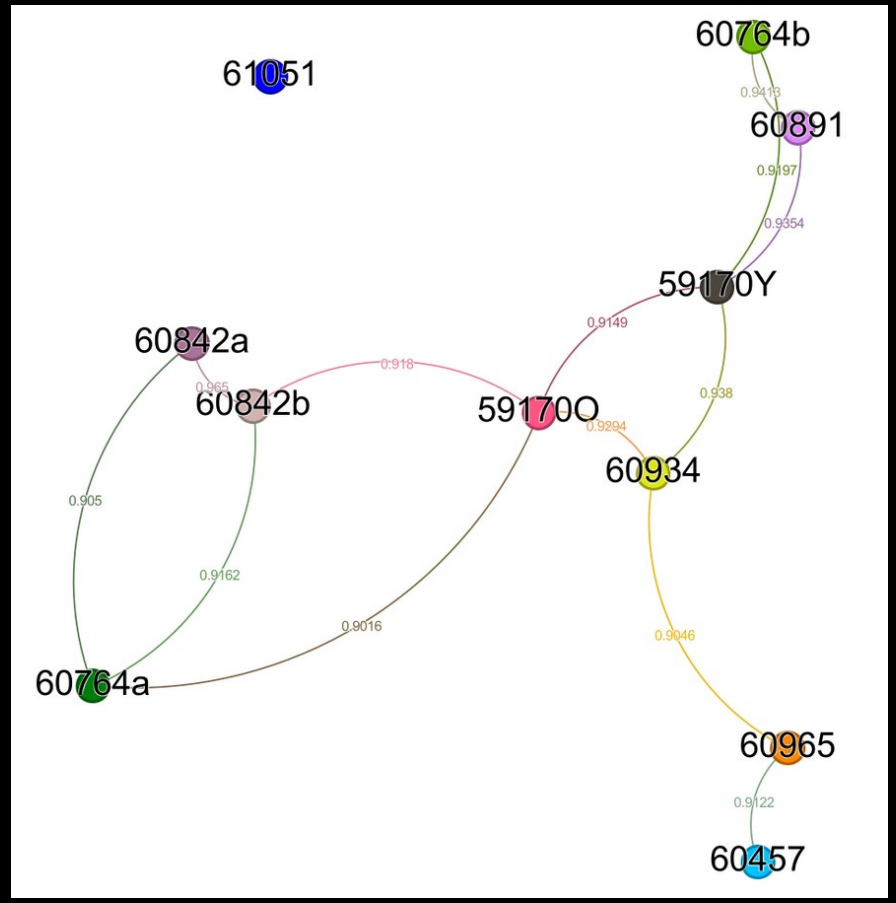
Isabelle Marthot-Santaniello

The Digital Paleography of Greek Papyri: State of Research





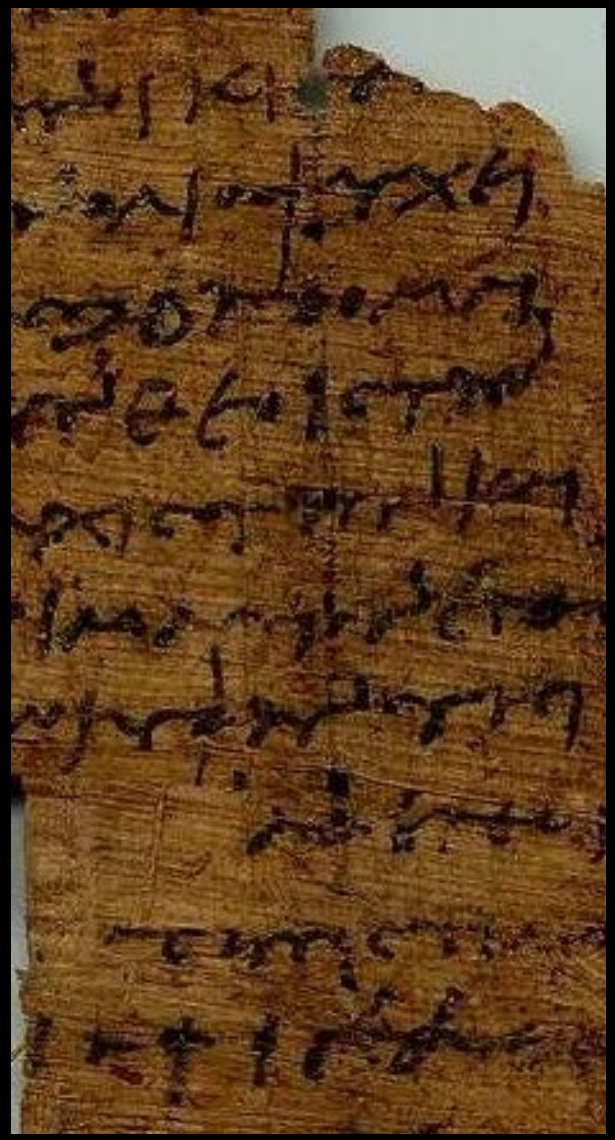
Swiss National Science Foundation

HTR and OCR from papyrus to codex
Sunoikisis, 29 April 2024

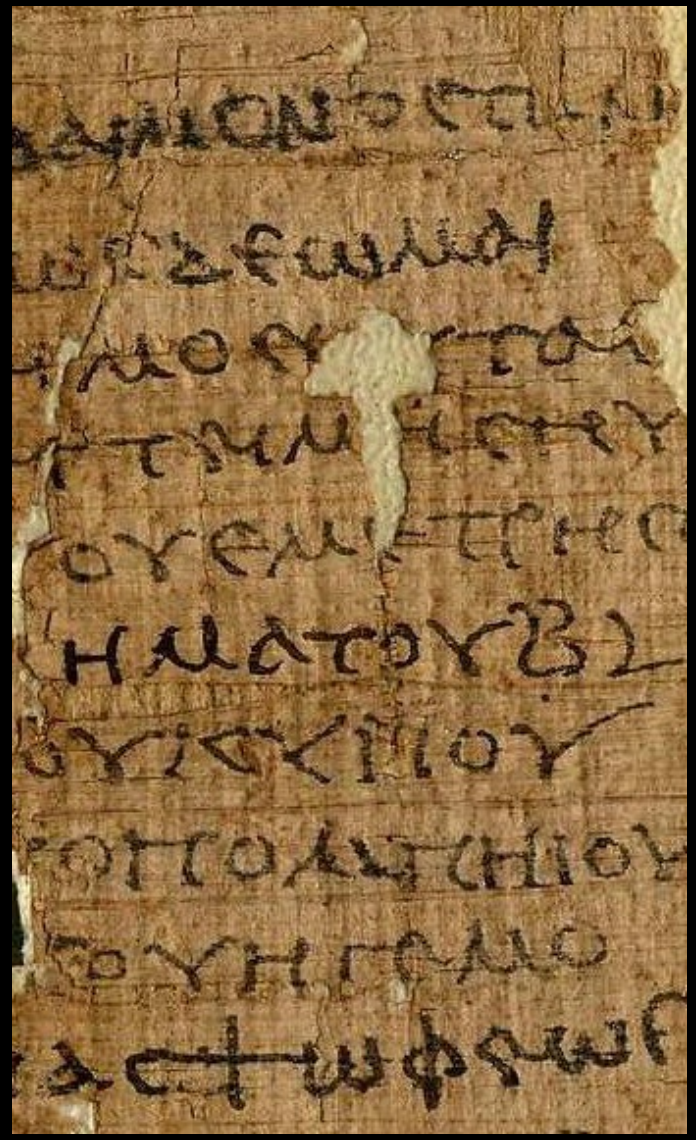
Universität Basel

The papyrological material

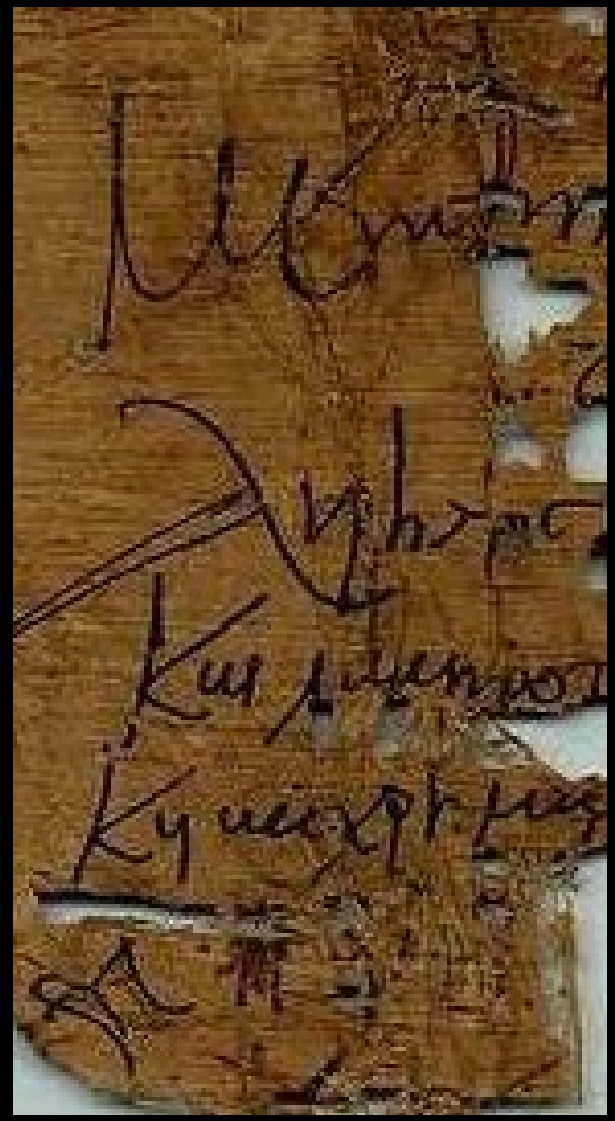
- A millenium of Greek in Egypt (from the end of the 4th BCE to the beginning of the 8th CE)



Ptolemaic
2nd – 1st c. BC



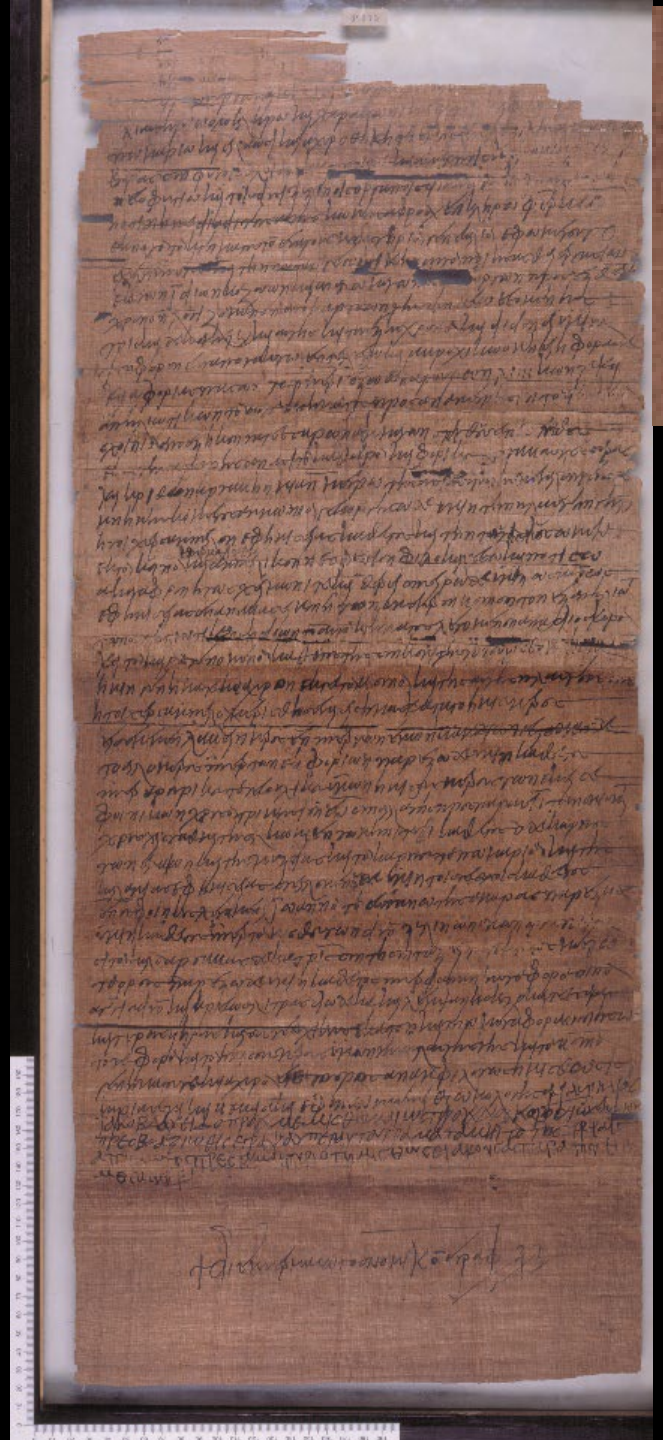
Roman
2nd – 3rd c. AD

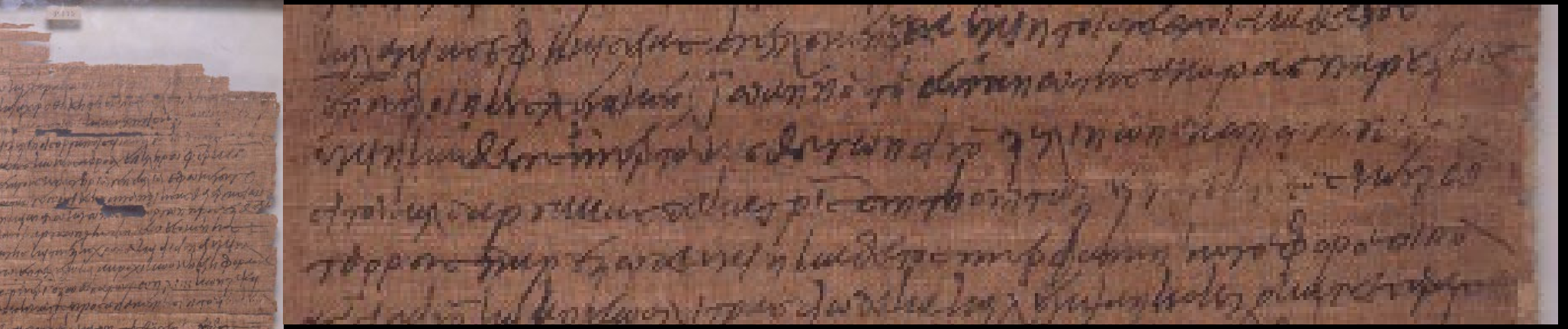


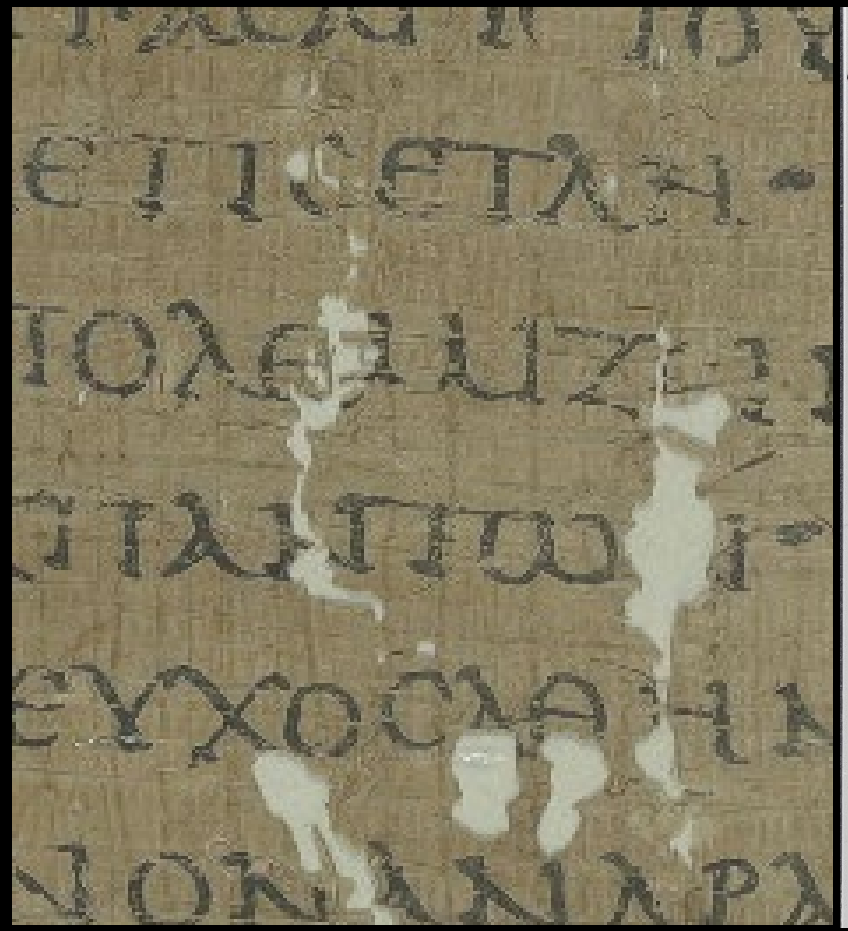
Byzantine
5th – 6th c. AD

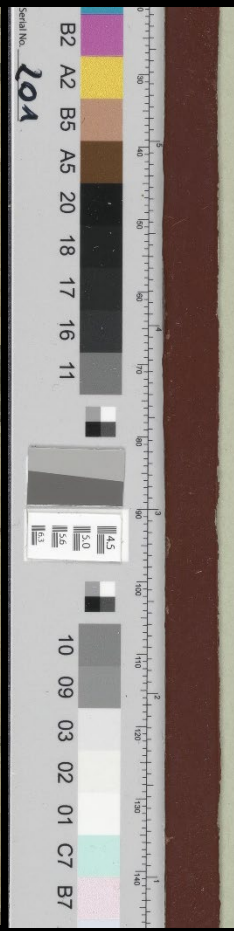
The papyrological material

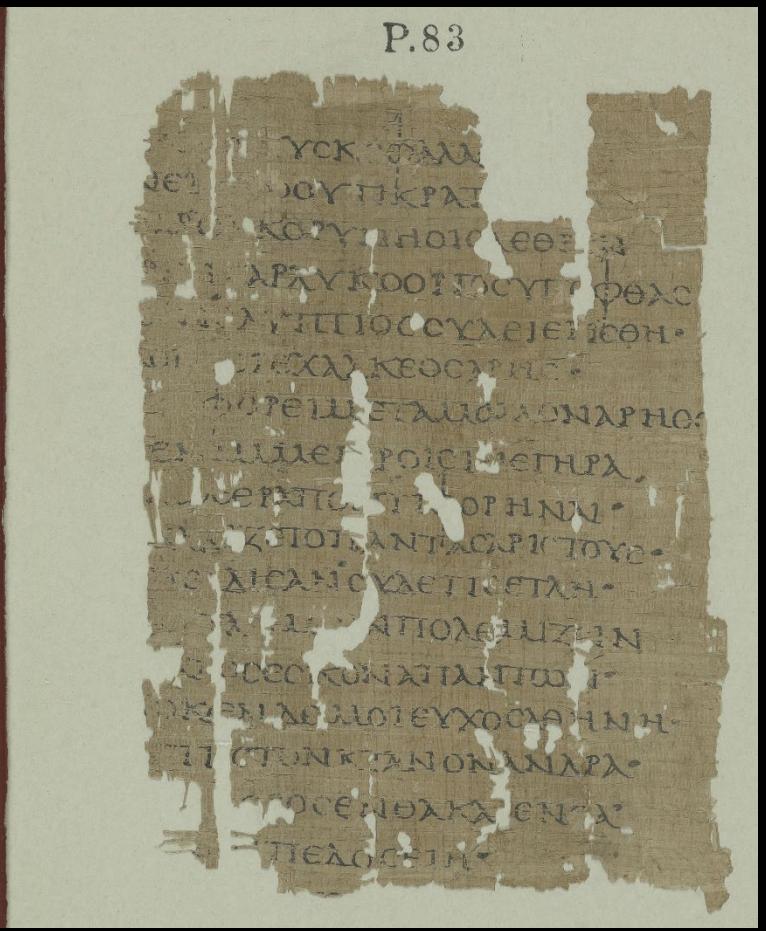
- A millenium of Greek in Egypt (from the end of the 4th BCE to the beginning of the 8th CE)
- A great variety of literary, paraliterary and documentary texts from deluxe books and official administrative communication to drafts, receipes, private accounts and school exercices on ostraca
- Various states of preservation (holes, breaks, faded or erased ink)










P.83

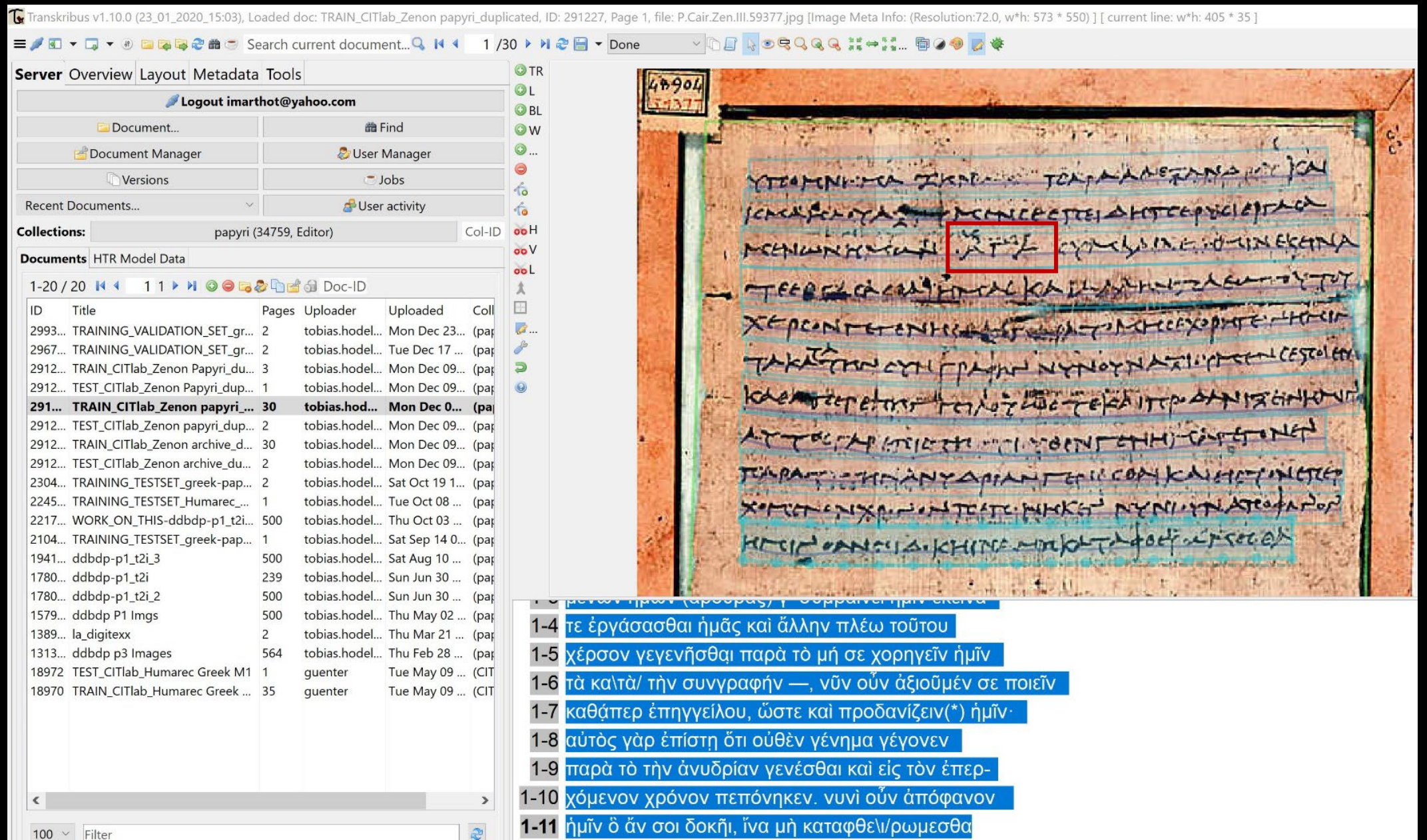
The peculiarities of (digital) papyrology

Unpublished texts

State and access vary according to collections (restoration, digitization, currently under study)

-> No demand for HTR for now

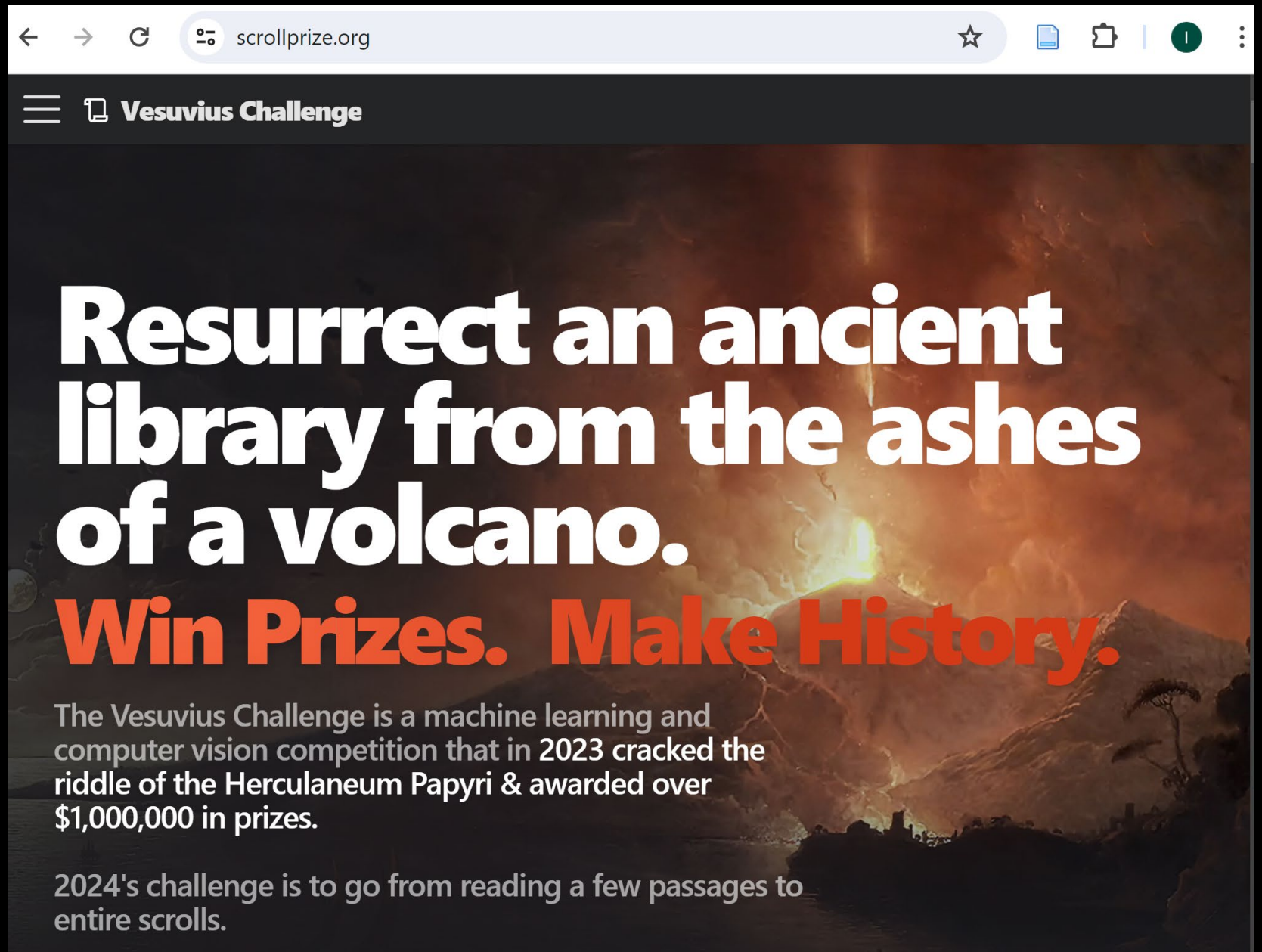
Transkribus (with Tobias Hodel)



Isabelle Marthot-Santaniello and Tobias Hodel, “Papyri, Handwritten Text Recognition, and Text Processing. State of the Art and Outlook – Approaches to a digital paleography”, Proceedings of the international workshop *In the Name of the Rose Searching for Unknown, Lost, and Forgotten Ancient Texts*, Rome, September30-October 1, 2021
Forthcoming publication

Dataset “Ground-Truthed Data Set of Zenon Papyri for Handwritten Text Recognition” https://zenodo.org/records/6565706

An exception: the Herculaneum papyri



European Commission | Horizon 2020 European Union funding for Research & Innovation

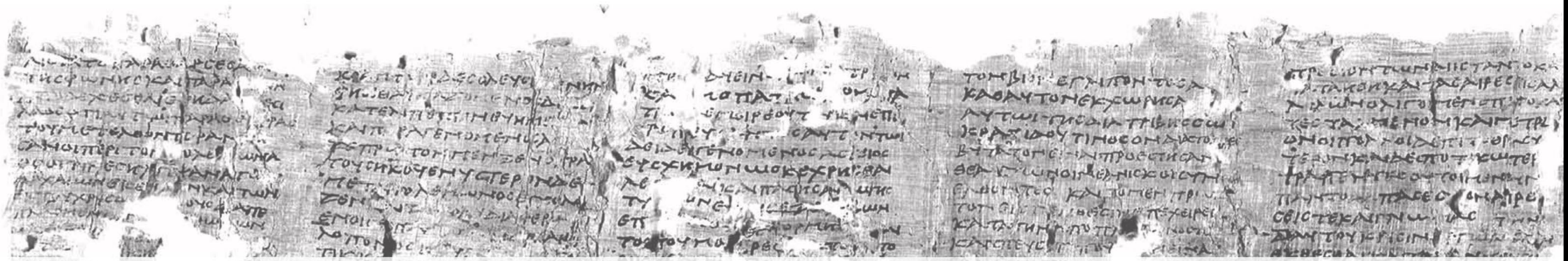
erc

GREEK SCHOOLS

Università di Pisa

Consiglio Nazionale delle Ricerche

MiC Ministero della Cultura



ERC Advanced Grant 885222-GreekSchools

THE GREEK PHILOSOPHICAL SCHOOLS

according to Europe's earliest 'history of philosophy'

Towards a new pioneering critical edition of Philodemus' *Arrangement of the Philosophers*

The peculiarities of (digital) papyrology

- 80,000 published texts

trismegistos.org

utions

KEEP TM ALI

Welcome to

TRISMEGISTOS

An interdisciplinary portal of the ancient world

TRISMEGISTOS

Combine dates, places, name Go ?

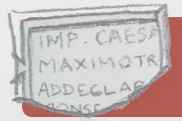
you can now also use EDCS numbers, e.g. 'EDCS-27900304'; other project IDs to follow

central search video tutorial

How to cite Trismegistos

email password Envoyer

Not registered? Subscribe here.



Texts

LDAB

Languages / scripts

Collections

Archives

People

Places

The peculiarities of (digital) papyrology

- 80,000 published texts
- 60,000 digitally encoded

Papyri.info

sign i

Browse: DDbDP HGV APIS DCLP Authors TM Number or Search: Data Bibliography

Search

New Search Search

within chars

and or not then near lex clear -

regex abbr start-not end-not

- [ ] Convert from betacode as you type
- [x] ignore capitalization
- [x] ignore diacritics/accents

(•) Text ( ) Metadata ( ) Translations

Series --- All values --- Vol.

or

Collection --- All values --- ID #

Author --- All values ---

Work --- All values ---

<<

Records per page 15 Go

Please select values from the left-hand column to return results

Selecting a value using the controls in the left-hand column will return a list of all documents that match it in the right-hand column. Once these results have been returned, the controls can be used to further refine the search with additional values. This process of adding new search constraints can be applied repeatedly until the results have been narrowed as far as desired.

More about string-search

More about searching by Series and Collection

More about searching by Provenance

More about searching by Nome

More about searching by Date

More about searching by Language

More about searching by Translation Language

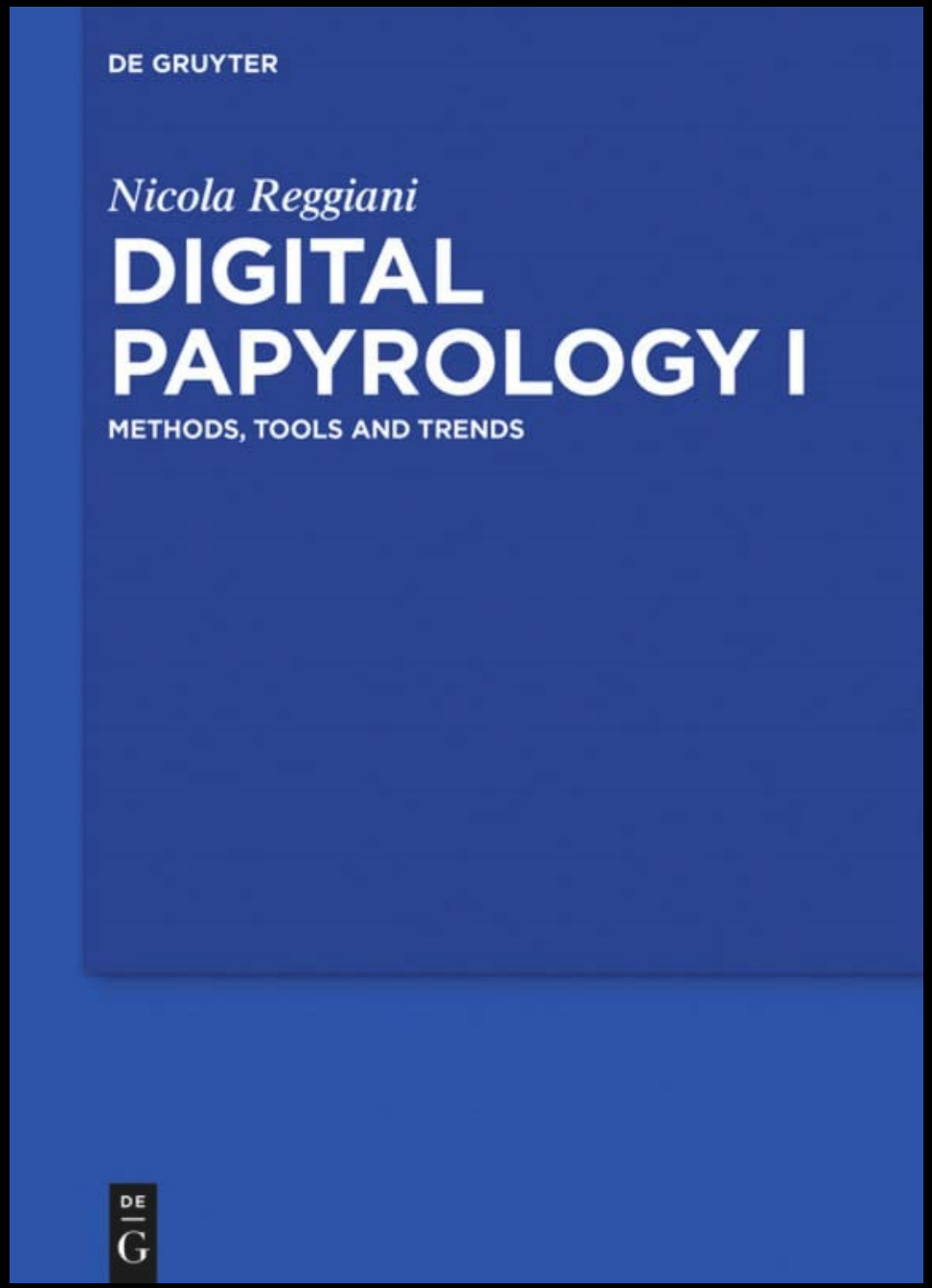
More about searching by Image

More about searching by Transcription

To remove values from your search criteria and broaden your search results, click on the 'x' in the upper-left-hand corner of the stored value when it appears. To remove all values and start again, click on 'New Search'.

The peculiarities of (digital) papyrology

- 80,000 published texts
- 60,000 digitally encoded
- 34,000 digital images



2017

https://doi.org/10.1515/9783110547474

Rather than HTR, text-image alignment (Rodney Ast, Holger Essler with E-scriptorium)

msia.escriptorium.fr

eScriptorium Home Contact

eScriptorium

eScriptorium: A Digital Text Production Pipeline for Print and Handwritten Texts using machine learning techniques.

Automatic Transcription

Apply OCR/HTR to images of printed and handwritten documents using shared open models.

Manual transcription

Make use of an ergonomic user interface leveraging modern browser technology to edit segmentations and transcriptions.

Train Models

Create new models or finetune existing ones to improve automatic recognition.

Import/ Export

Import and export models and texts transcriptions in a variety of formats. Access data through a full REST API.

Datasets of images

- Securely dated texts PapPal: https://pappal.info/
- Dateable bookhands CDDGB: https://classics.artsandsciences.baylor.edu/academics/greek-bookhands-database
- Aphrodito papyri BIPAB: http://bipab.aphrodito.info/

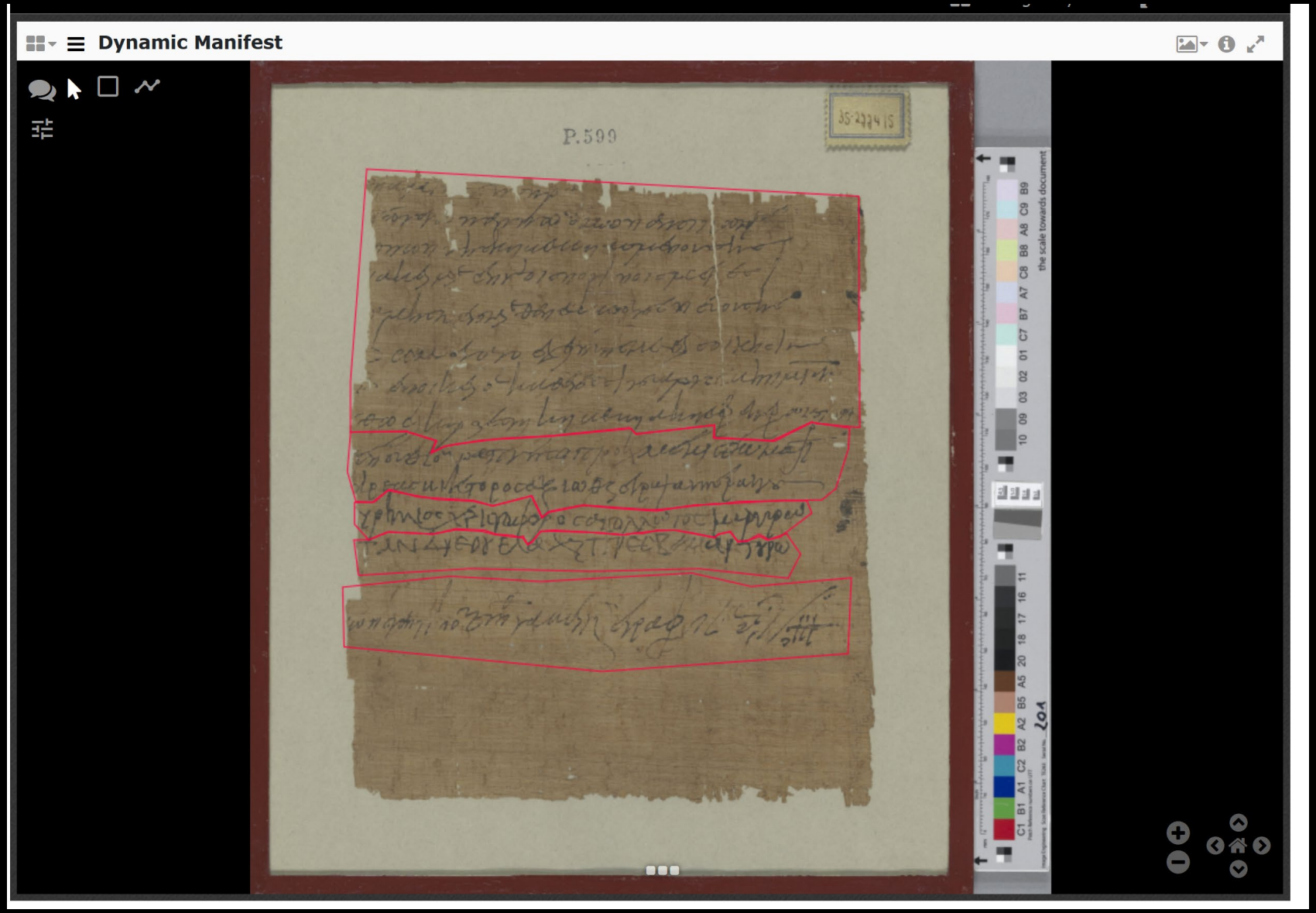
D-scribes

Reuniting fragments, identifying scribes and characterizing scripts:

The Digital Palaeography of Greek and Coptic Papyri

SNSF Ambizione project 2018-2023

Writers in Dioscorus archive




Andreas_6.jpg


Andreas_8.jpg


Dioscorus_1.jpg


Dioscorus_2.jpg


Dioscorus_3.jpg


Dioscorus_4.jpg


Dioscorus_5.jpg


Hermauos_1.jpg


Hermauos_2.jpg


Hermauos_3.jpg


Hermauos_4.jpg


Hermauos_5.jpg

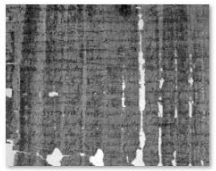

Isak_1.jpg


Isak_2.jpg


Isak_3.jpg


Isak_4.jpg


Isak_5.jpg


Kyros1_1.jpg

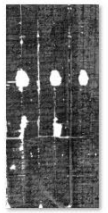

Kyros1_3.jpg


Kyros1_4.jpg


Kyros1_5.jpg


Kyros3_1.jpg

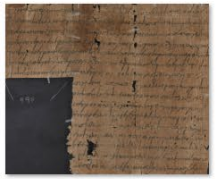

Kyros3_2.jpg


Kyros3_3.jpg


Kyros3_4.jpg


Menas_1.jpg


Menas_2.jpg


Menas_3.jpg


Menas_4.jpg

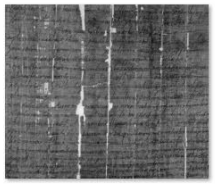

Menas_5.jpg

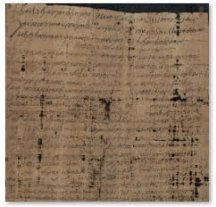

Pilatos_1.jpg


Pilatos_2.jpg


Pilatos_3.jpg

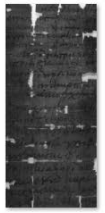

Pilatos_4.jpg


Pilatos_5.jpg


Pilatos_6.jpg


Victor_2.jpg


Victor_3.jpg


Victor_5.jpg


Victor_6.jpg

120 samples of 23 notaries

State of the art Writer Identification methods

- if enough well preserved samples
- work for inter-writer discrimination but struggle with intra-writer variations

= the networks manages to recognize Dios' handwriting because he is only known by a single long papyrus but not Abraam's known by several small texts spread over 30 years.

https://d-scribes.philhist.unibas.ch/en/case-studies/dioscorus/

Dioscorus case study

The archive of Dioscorus of Aphrodito (6th c. A.D.) is the richest papyrus archive of the Byzantine period. Gathering more than 700 texts of various content, it provides a unique opportunity to study day-to-day cursive writing and literacy in the microcosm of an Egyptian village.

From the Dioscorus archive was selected a dataset specially tailored for the task of Writer Identification, see > GRK-Papyri

> GRK-Papyri_120 is an extension of GRK-Papyri from 50 to 120 images of papyri. It was segmented into rows to form the Papy-Row dataset (> accessible here) and presented during the PatReCH workshop in ICPR 2020, see > here.

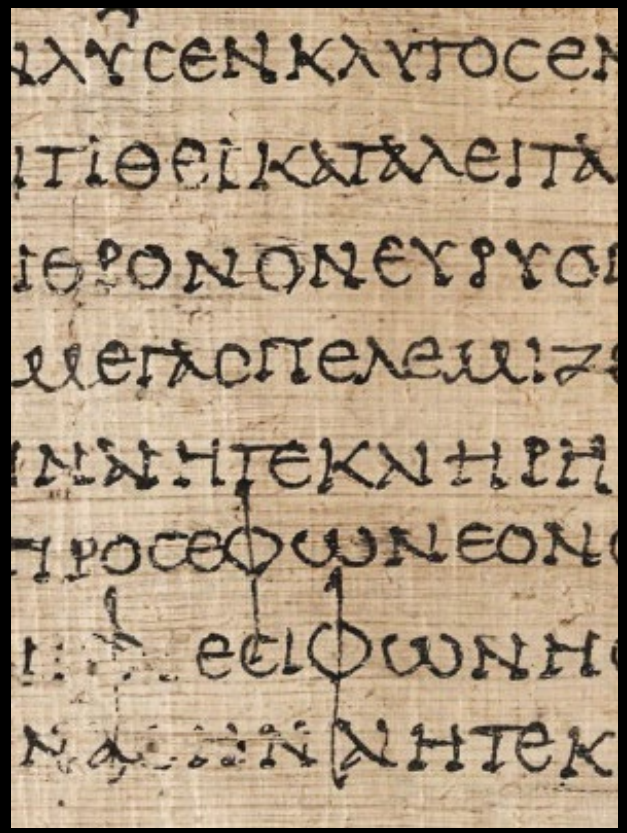
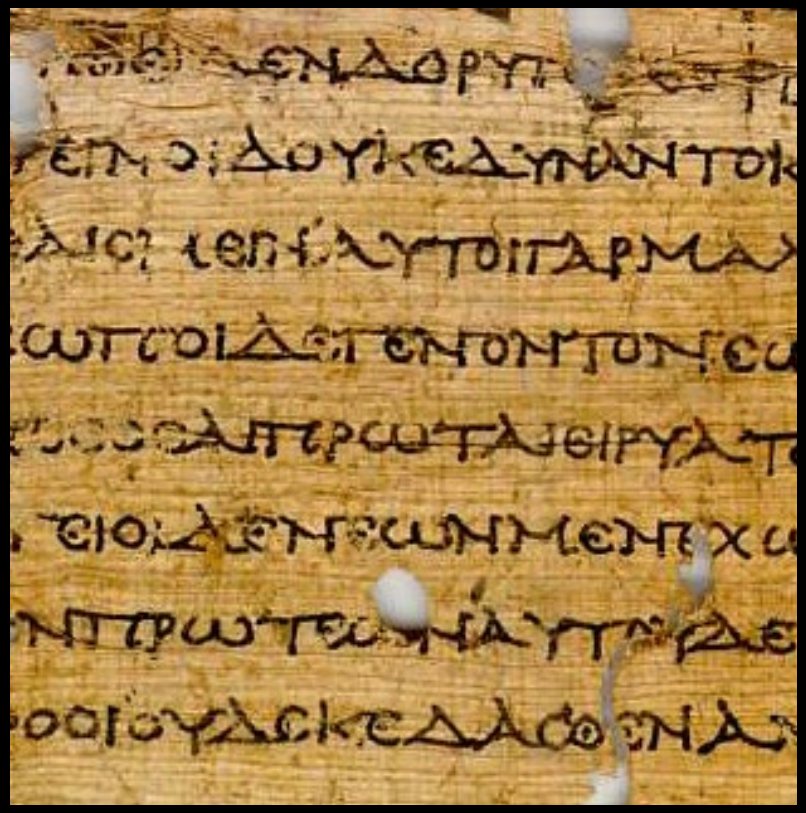
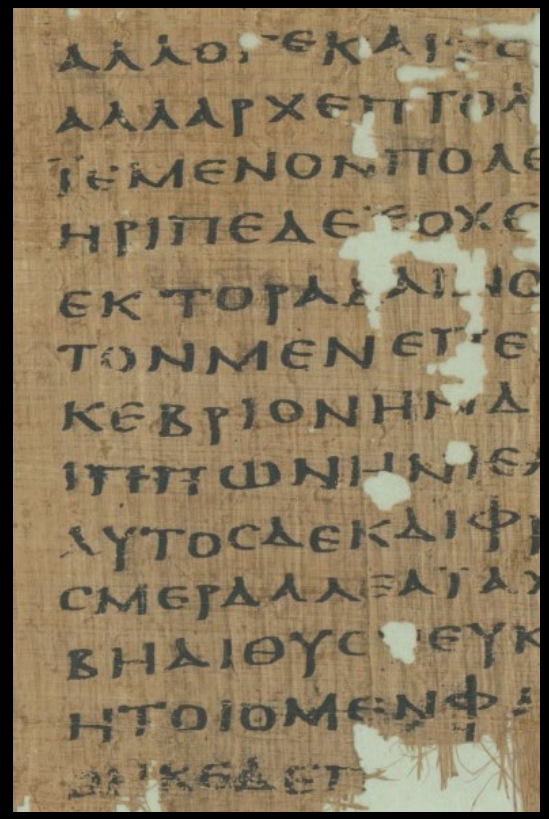
> GRK-Papyri_97 is a selection of GRK-Papyri_120 to count only 19 writers (97 images) is a study currently submitted to ICDAR 2023.

GRK-Papyri and Papy-Row are the basis or the following publications:

Script characterization

Homer's *Iliad*:

1500 papyri across a millenium

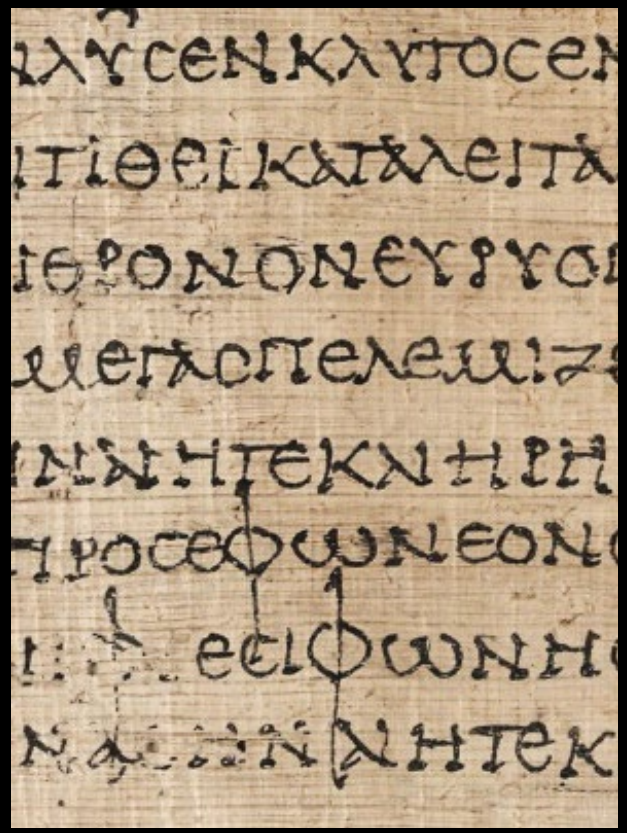
Script characterization

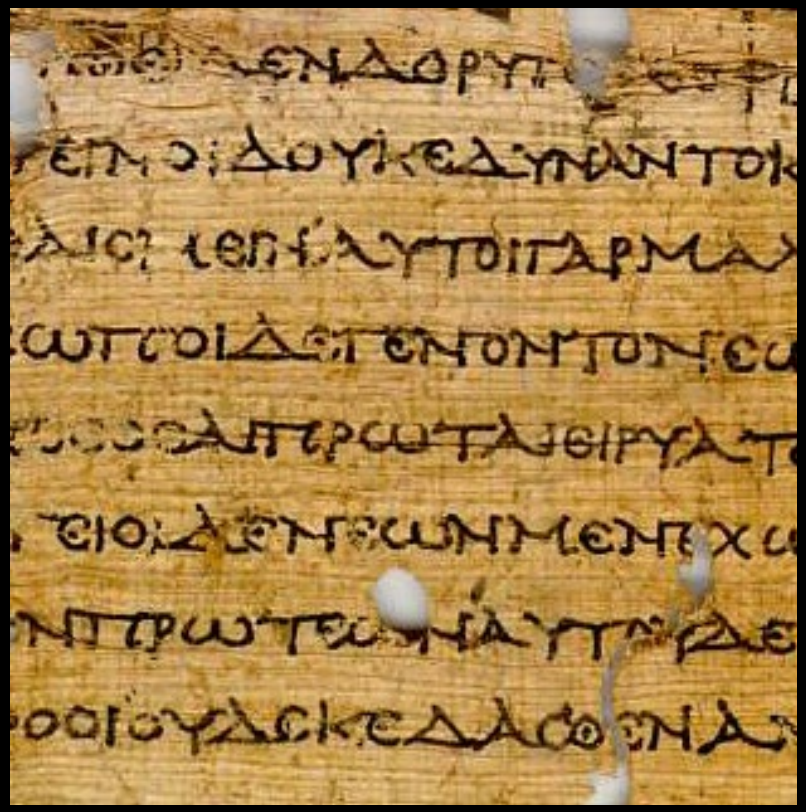
Homer's *Iliad*:

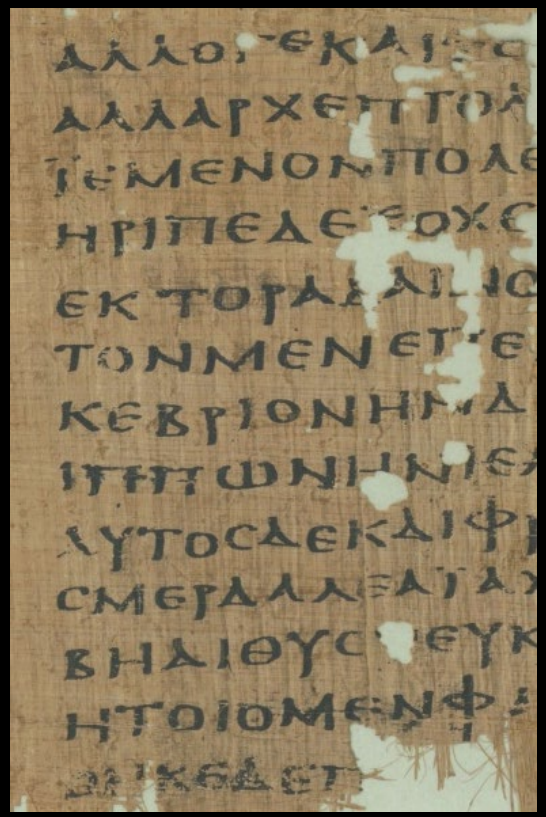
1500 papyri across a millenium

No secure date nor writer

No consensus on style typology



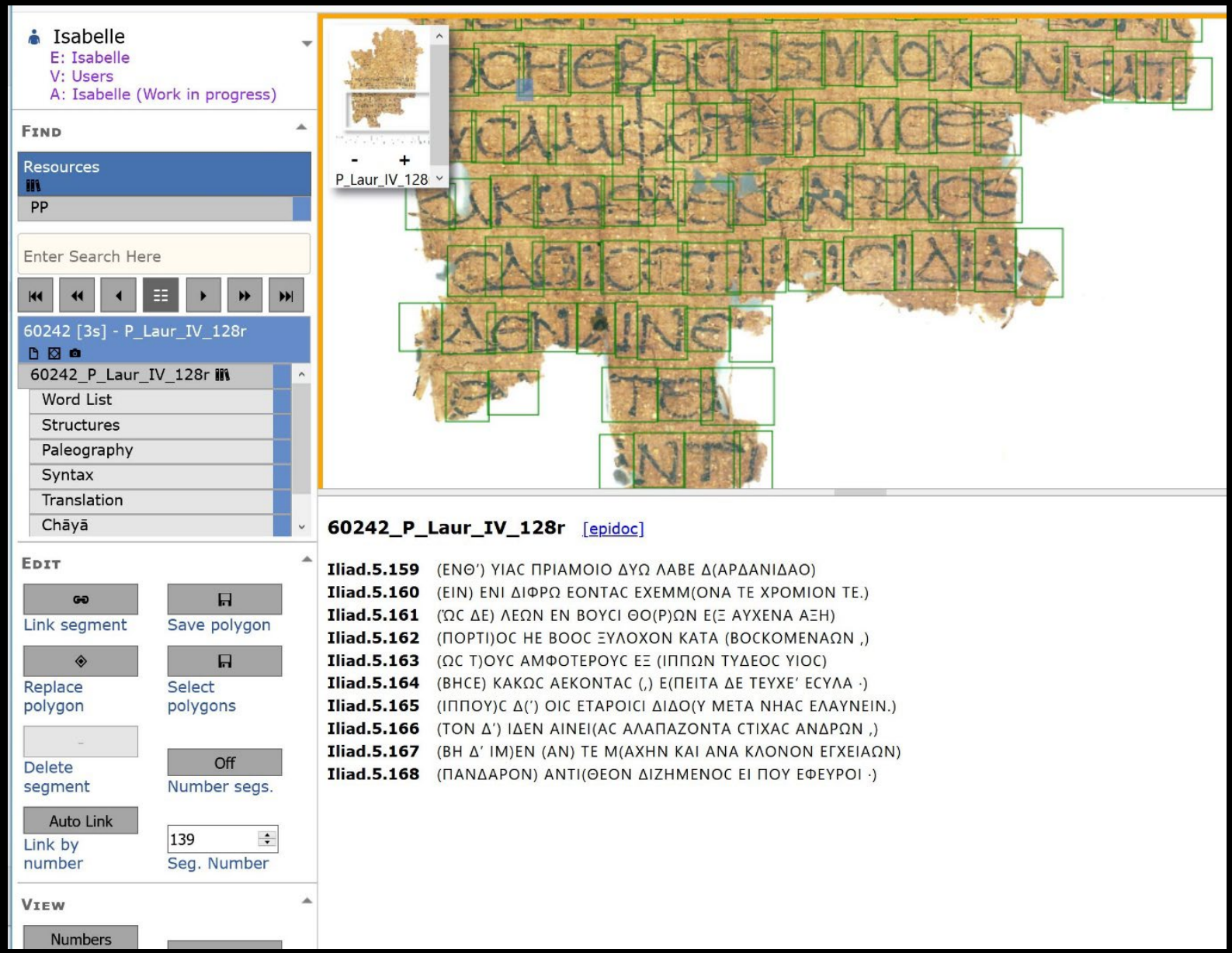




READ (O. Serbaeva, S. White)

Research Environment for Ancient Documents

READ (O. Serbaeva, S. White)



Isabelle
E: DScribesEditor
V: Users
A: Isabelle (Work in progress)

Find

Enter Search Here

60220 [4.1] - P_08440_R_001

Edit

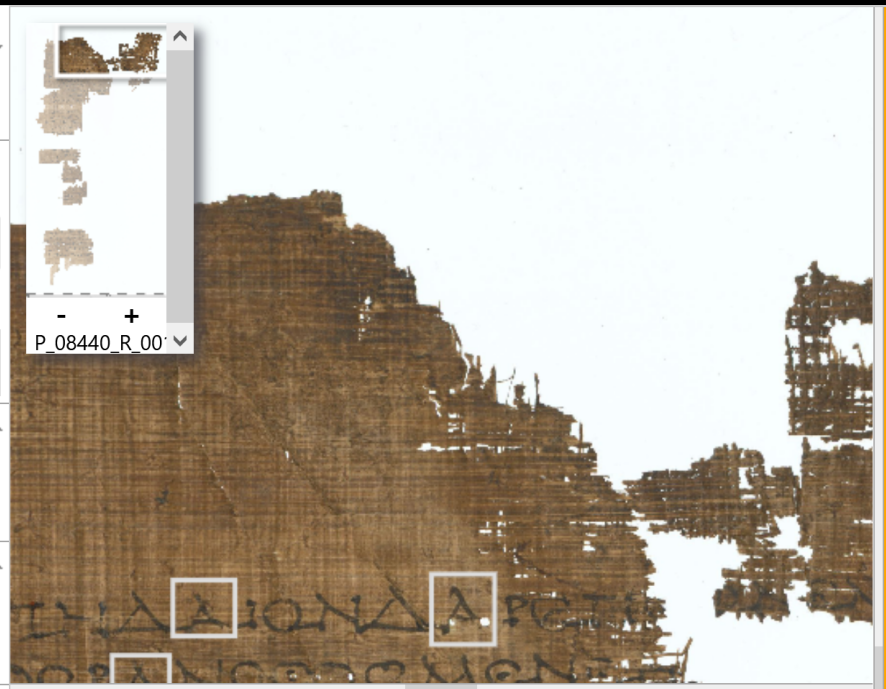
View
Edit mode

View

Properties
Download
Segments

Layout

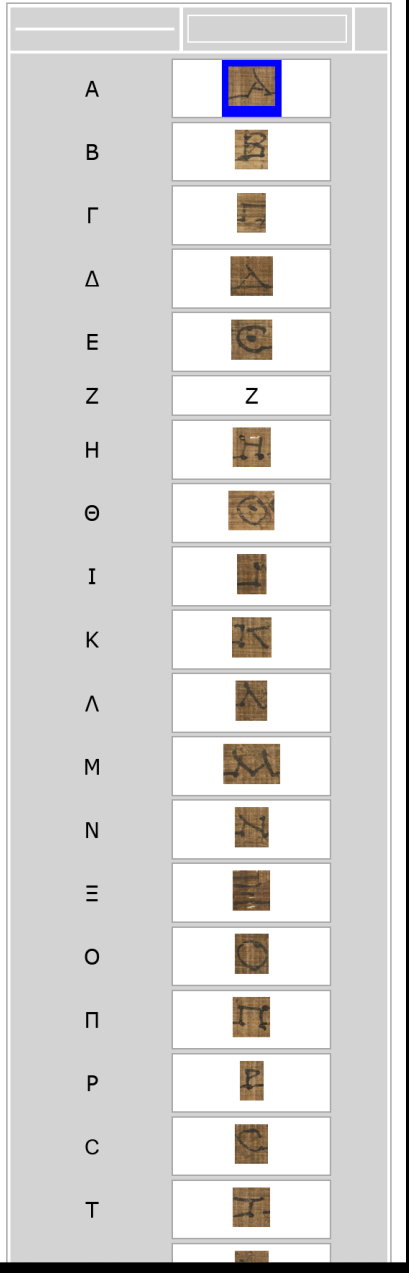
Change
Panel Layout
Off
Sync Scroll

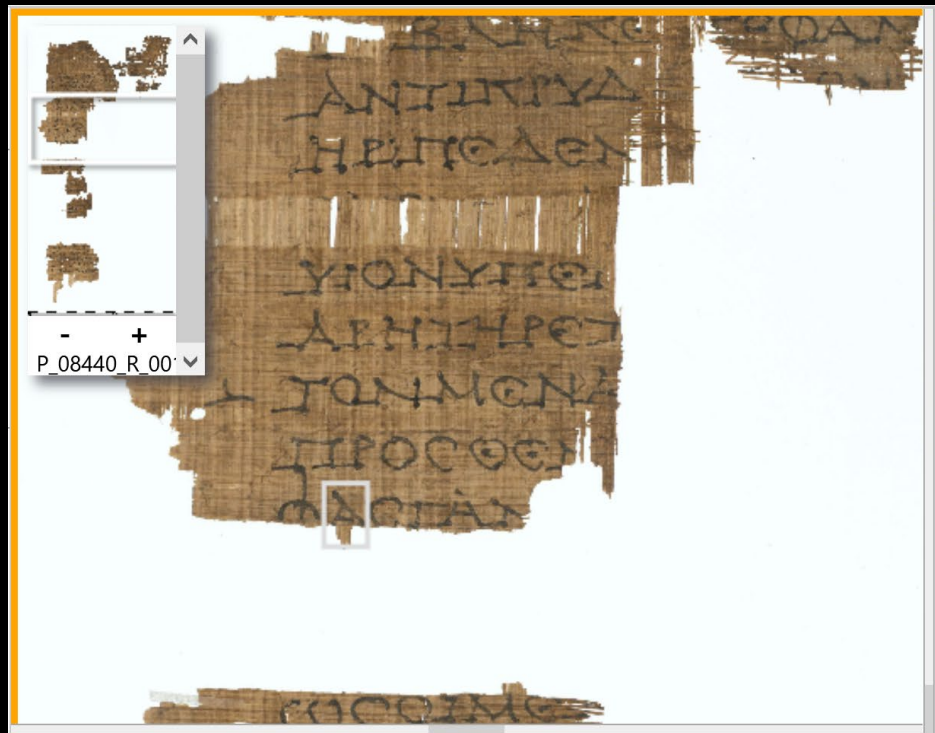


60220_P_08440_R_001 [epidoc]

Iliad.5.69 ΠΗΔΑΙΟΝ Δ(’) ΑΡ(’) ΕΠ(ΕΦΝ)Ε ΜΕΓΗΣ Α(ΝΤΗΝΟΡΟΣ ΥΙΟΝ)
Iliad.5.70 ΟΣ ΡΑ ΝΟΘΟΣ ΜΕΝ (ΕΗΝ ΠΥΚΑ Δ’ ΕΤΡΕΦΕ ΔΙΑ ΘΕΑΝΩ)
Iliad.5.71 ΙΣΑ ΦΙΛΟΙΣ(Ι) ΤΕΚΕΣΣΙ Χ(ΑΡΙΖΟΜΕΝΗ ΠΟΣΕΙ Ω)
Iliad.5.72 ΤΟΝ ΜΕΝ ΦΥΛΕΙΔΗ(Σ ΔΟΥΡΙ ΚΛΥΤΟΣ ΕΓΓΥΘΕ)Ν Ε(ΛΘΩΝ)
Iliad.5.73 (ΒΕ)ΒΛΗΚΕ(Ι) ΚΕΦΑΛ(ΗΣ) Κ(ΑΤΑ ΙΝΙΟΝ ΟΞΕΙ ΔΟΥΡΙ)
Iliad.5.74 ΑΝΤΙΚΡΥ Δ(’ ΑΝ’ ΟΔΟΝΤΑΣ ΥΠΟ ΓΛΩΣΣΑΝ ΤΑΜΕ ΧΑΛΚΟΣ)
Iliad.5.75 ΗΡΙΠΕ Δ(’) ΕΝ (ΚΟΝΙΗ ΨΥΧΡΟΝ Δ’ ΕΛΕ ΧΑΛΚΟΝ ΟΔΟΥΣΙΝ)
Iliad.5.76 (ΕΥΡΥΠΥΛΟΣ Δ’ ΕΥΑΙΜΟΝΙΔΗΣ ΥΨΗΝΟΡΑ ΔΙΟΝ)
Iliad.5.77 ΥΙΟΝ ΥΠΕΡ(ΘΥΜΟΥ ΔΟΛΟΠΙΟΝΟΣ ΟΣ ΡΑ ΣΚΑΜΑΝΔΡΟΥ)
Iliad.5.78 ΑΡΗΤΗΡ ΕΤ(ΕΤΥΚΤΟ ΘΕΟΣ Δ’ ΩΣ ΤΙΕΤΟ ΔΗΜΩ)
Iliad.5.79 ΤΟΝ ΜΕΝ Α(Ρ’ ΕΥΡΥΠΥΛΟΣ ΕΥΑΙΜΟΝΟΣ ΑΓΛΑΟΣ ΥΙΟΣ)
Iliad.5.80 ΠΡΟΣΘΕΝ (ΕΘΕΝ ΦΕΥΓΟΝΤΑ ΜΕΤΑΔΡΟΜΑΔΗΝ ΕΛΑΣ’ ΩΜΟΝ)
Iliad.5.81 ΦΑΣΓΑΝ(Ω ΑΙΞΑΣ ΑΠΟ Δ’ ΕΞΕΣΕ ΧΕΙΡΑ ΒΑΡΕΙΑΝ)
Iliad.5.84 ΩΣ ΟΙ ΜΕ(Ν ΠΟΝΕΟΝΤΟ ΚΑΤΑ ΚΡΑΤΕΡΗΝ ΥΣΜΙΝΗΝ)
Iliad.5.85 ΤΥΔΕΙΔ(ΗΝ Δ’ ΟΥΚ ΑΝ ΓΝΟΙΗΣ ΠΟΤΕΡΟΙΣΙ ΜΕΤΕΙΗ)
Iliad.5.86 ΗΕ (Μ)ΕΤΑ (ΤΡΩΕΣΣΙΝ ΟΜΙΛΕΟΙ Η ΜΕΤ’ ΑΧΑΙΟΙΣ)
Iliad.5.87 (ΘΥΝΕ Γ)ΑΡ Α(Μ ΠΕΔΙΟΝ ΠΟΤΑΜΩ ΠΛΗΘΟΝΤΙ ΕΟΙΚΩΣ)
Iliad.5.88 (ΧΕΙΜ)ΑΡΡΩ (ΟΣ Τ’ ΩΚΑ ΡΕΩΝ ΕΚΕΔΑΣΣΕ ΓΕΦΥΡΑΣ)

Paleography for 60220_P_





60220_P_08440_R_001 [epidoc]

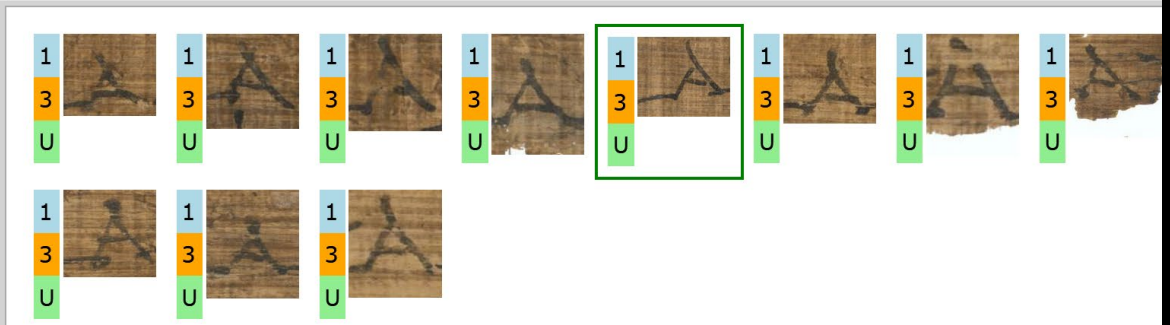
Iliad.5.69 ΠΗΔΑΙΟΝ Δ(’) ΑΡ(’) ΕΠ(ΕΦΝ)Ε ΜΕΓΗC Α(ΝΤΗΝΟΡΟC ΥΙΟΝ)
Iliad.5.70 ΟC ΡΑ ΝΟΘΟC ΜΕΝ (ΕΗΝ ΠΥΚΑ Δ’ ΕΤΡΕΦΕ ΔΙΑ ΘΕΑΝΩ)
Iliad.5.71 ΙCΑ ΦΙΛΟΙC(Ι) ΤΕΚΕCCΙ Χ(ΑΡΙΖΟΜΕΝΗ ΠΟCΕΙ Ω)
Iliad.5.72 ΤΟΝ ΜΕΝ ΦΥΛΕΙΔΗ(C ΔΟΥΡΙ ΚΛΥΤΟC ΕΓΓΥΘΕ)Ν Ε(ΛΘΩΝ)
Iliad.5.73 (ΒΕ)ΒΛΗΚΕ(Ι) ΚΕΦΑΛ(ΗC) ΚΑ(ΤΑ ΙΝΙΟΝ ΟΞΕΙ ΔΟΥΡΙ)
Iliad.5.74 ΑΝΤΙΚΡΥ Δ(’ ΑΝ’ ΟΔΟΝΤΑC ΥΠΟ ΓΛΩCCΑΝ ΤΑΜΕ ΧΑΛΚΟC)
Iliad.5.75 ΗΡΙΠΕ Δ(’) ΕΝ (ΚΟΝΙΗ ΨΥΧΡΟΝ Δ’ ΕΛΕ ΧΑΛΚΟΝ ΟΔΟΥCΙΝ)
Iliad.5.76 (ΕΥΡΥΠΥΛΟC Δ’ ΕΥΑΙΜΟΝΙΔΗC ΥΨΗΝΟΡΑ ΔΙΟΝ)
Iliad.5.77 ΥΙΟΝ ΥΠΕΡ(ΘΥΜΟΥ ΔΟΛΟΠΙΟΝΟC ΟC ΡΑ CΚΑΜΑΝΔΡΟΥ)
Iliad.5.78 ΑΡΗΤΗΡ ΕΤ(ΕΤΥΚΤΟ ΘΕΟC Δ’ ΩC ΤΙΕΤΟ ΔΗΜΩ)
Iliad.5.79 ΤΟΝ ΜΕΝ Α(Ρ’ ΕΥΡΥΠΥΛΟC ΕΥΑΙΜΟΝΟC ΑΓΛΑΟC ΥΙΟC)
Iliad.5.80 ΠΡΟCΘΕΝ (ΕΘΕΝ ΦΕΥΓΟΝΤΑ ΜΕΤΑΔΡΟΜΑΔΗΝ ΕΛΑC’ ΩΜΟΝ)
Iliad.5.81 ΦΑCΓΑΝ(Ω ΑΙΞΑC ΑΠΟ Δ’ ΕΞΕCΕ ΧΕΙΡΑ ΒΑΡΕΙΑΝ)

Paleography for 60220_P_08440_R_001

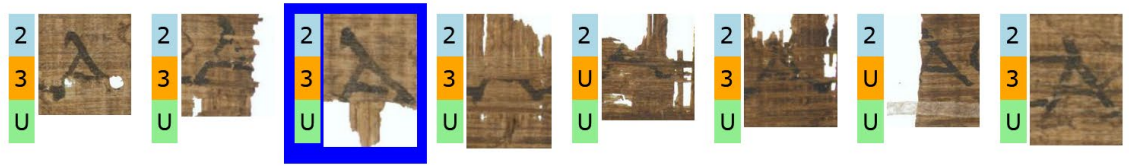
Sort: base | Tagging Orthographic Group **A**

← Previous | → Next

BT-1



BT-2



BT-3

https://showcase.d-scribes.philhist.unibas.ch/viewer

https://showcase.d-scribes.philhist.unibas.ch/viewer


ω_60468_bt1_Iliad.2.20.25.png


ω_60468_bt1_Iliad.2.22.22.png


ω_60471_bt1_Iliad.15.441.1.png


ω_60475_bt1_Iliad.17.87.38.png


ω_60476_bt1_ILIAD.22.6.29.png


ω_60476_bt1_ILIAD.22.7.33.png


ω_60476_bt1_ILIAD.22.8.32.png


ω_60476_bt1_ILIAD.22.45.6.png


ω_60476_bt1_ILIAD.22.51.9.png


ω_60476_bt1_ILIAD.22.57.3.png


ω_60479_bt1_Iliad.1.574.7.png


ω_60479_bt1_Iliad.1.577.10.png


ω_60479_bt1_Iliad.1.578.9.png


ω_60633_bt1_Iliad.12.271.14.png


ω_60633_bt1_Iliad.12.273.8.png


ω_60633_bt1_Iliad.12.274.9.png


ω_60633_bt1_Iliad.12.283.10.png


ω_60633_bt1_Iliad.12.287.13.png


ω_60633_bt1_Iliad.12.290.17.png


ω_60659_bt1_Iliad.5.117.12.png


ω_60663_bt1_Iliad.20.248.24.png


ω_60670_bt1_Iliad.10.21.32.png


ω_60701_bt1_Iliad.11.43.22.png


ω_60701_bt1_Iliad.11.49.26.png


ω_60701_bt1_Iliad.11.51.15.png


ω_60701_bt1_Iliad.11.51.24.png


ω_60732_bt1_Iliad.5.557.37.png


ω_60732_bt1_Iliad.5.558.35.png


ω_60732_bt1_Iliad.5.562.38.png


ω_60732_bt1_Iliad.5.571.25.png


ω_60732_bt1_Iliad.5.575.23.png


ω_60732_bt1_Iliad.5.577.23.png


ω_60732_bt1_Iliad.5.598.13.png


ω_60732_bt1_Iliad.5.773.25.png


ω_60732_bt1_Iliad.5.775.24.png


ω_60732_bt1_Iliad.5.824.5.png


ω_60764_bt1_Iliad.8.175.22.png


ω_60764_bt1_Iliad.8.175.38.png


ω_60764_bt1_Iliad.8.177.30.png


ω_60764_bt1_Iliad.8.314.13.png

ω_60764_bt1_Iliad.8.314.19.png

ω_60764_bt1_Iliad.8.319.4.png

ω_60812_bt1_Iliad.8.110.2.png

ω_60812_bt1_Iliad.8.112.1.png

ω_60812_bt1_Iliad.8.115.2.png

ω_60812_bt1_Iliad.8.116.5.png

ω_60842_bt1_Iliad.5.487.4.png

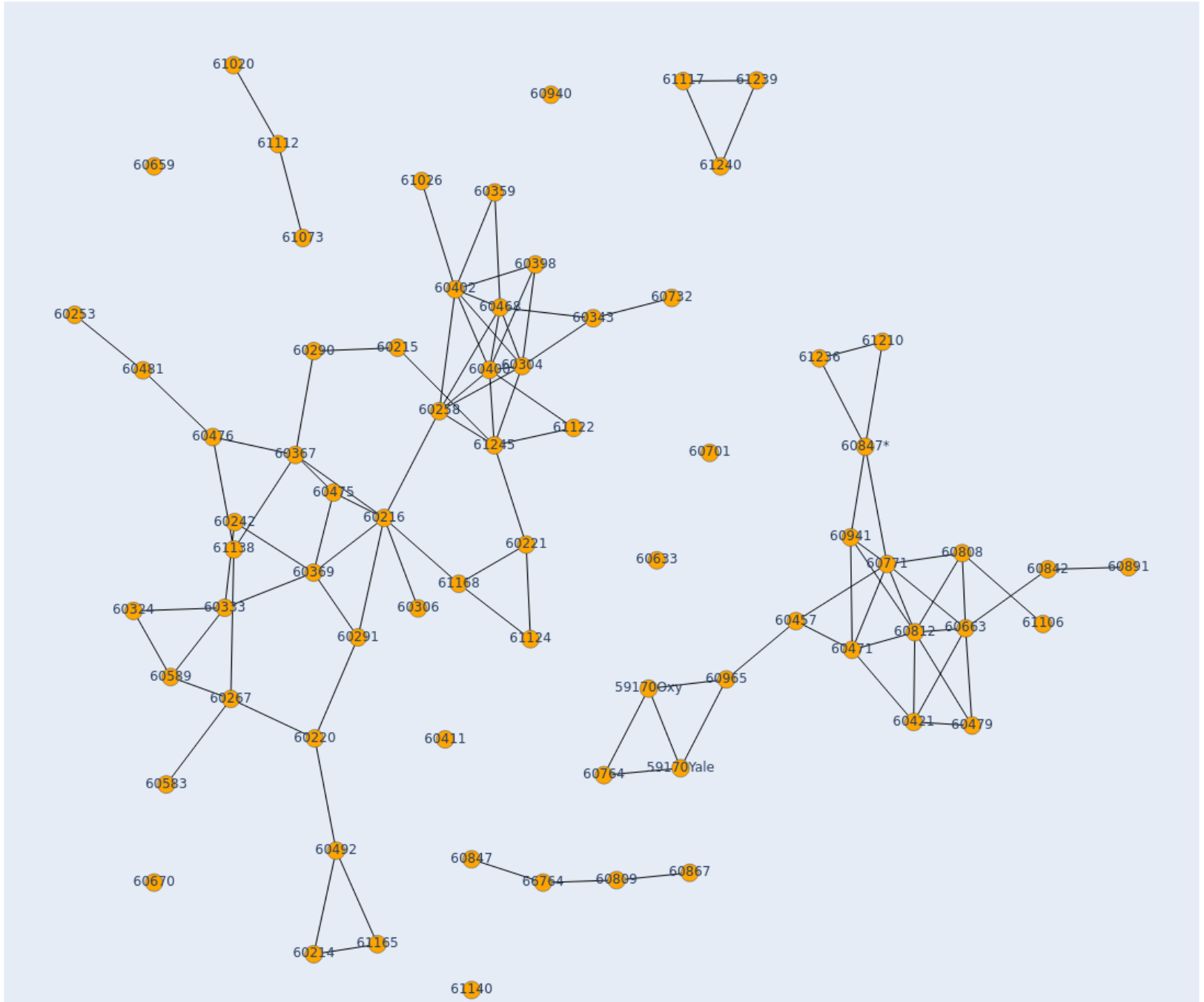
ω_60842_bt1_Iliad.5.487.6.png

I. Marthot-Santaniello, M. T. Vu, O. Serbaeva, M. Beurton-Aimar, “Stylistic Similarities in Greek Papyri Based on Letter Shapes: A Deep Learning Approach”
in M. Coustaty and A. Fornés (eds), *Document Analysis and Recognition – ICDAR 2023 Workshops*. Lecture Notes in Computer Science, vol 14193. Springer, Cham, 2023, p. 307–323.
https://doi.org/10.1007/978-3-031-41498-5_22

Similarity classification of literary papyri

- 72 literary papyri
- More than 5,000 cliplets of alphas, epsilons and mus
- AI: SimSiam network

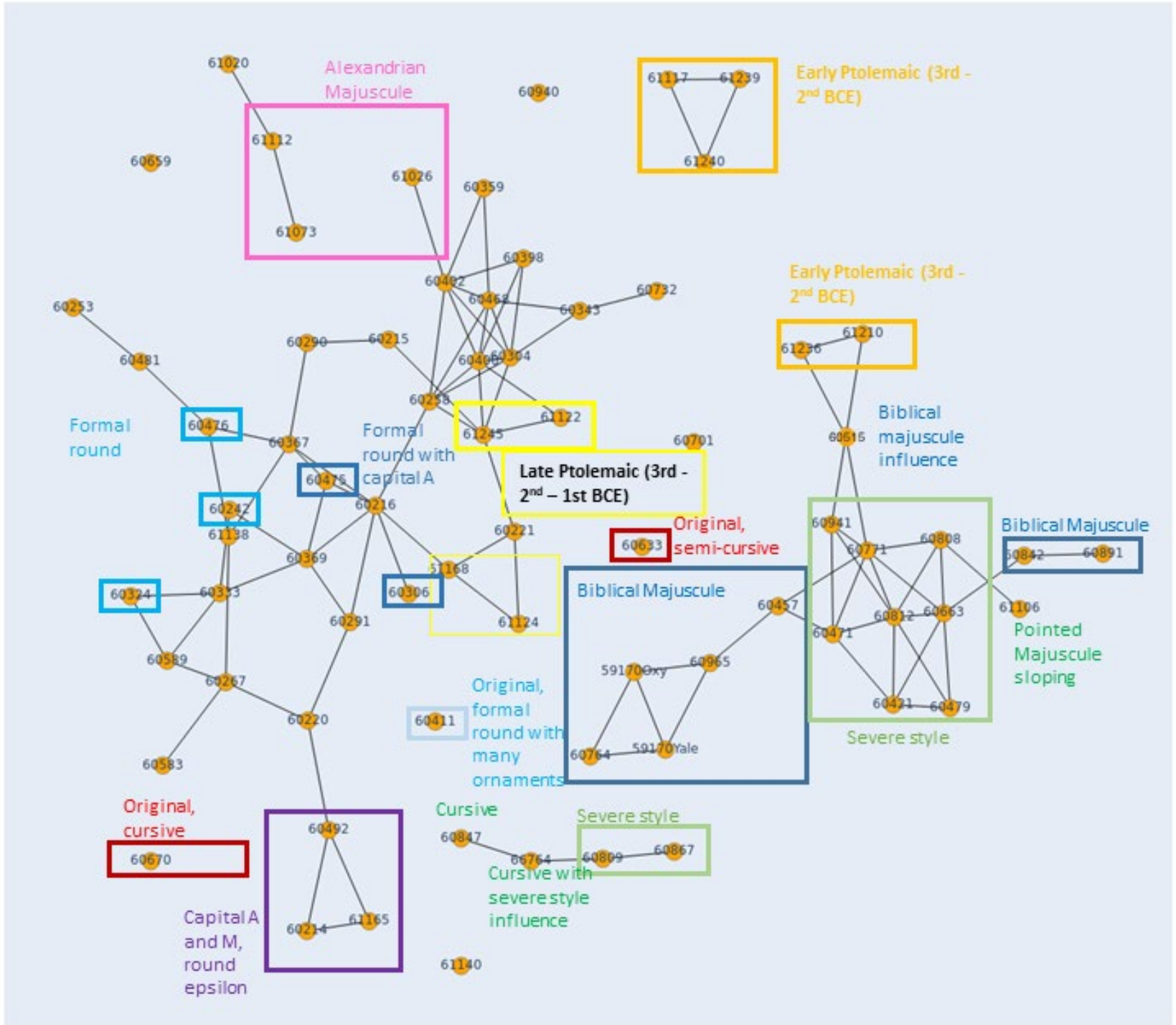
Similarity estimation results



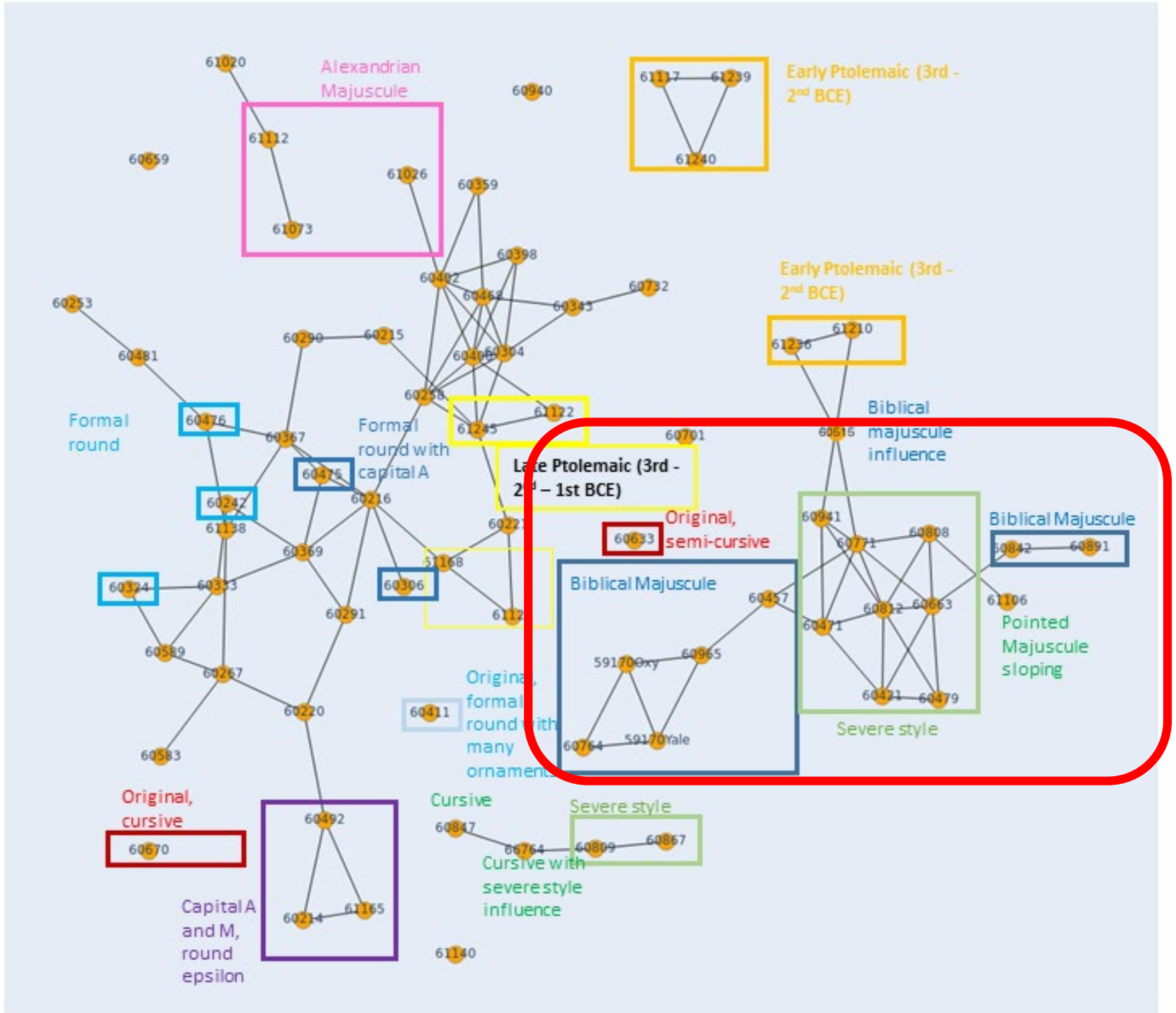
Similarity estimation results between papyri

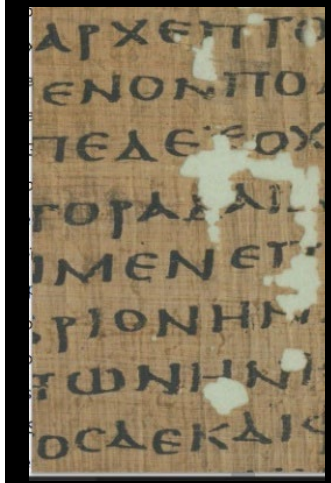
(combination of alpha, epsilon and mu scores)

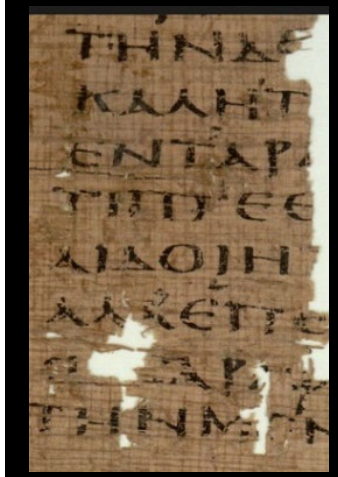
Stylistic interpretation

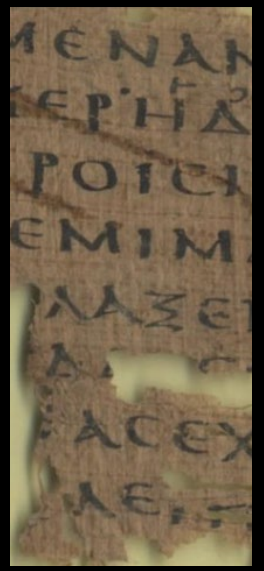


Stylistic interpretation





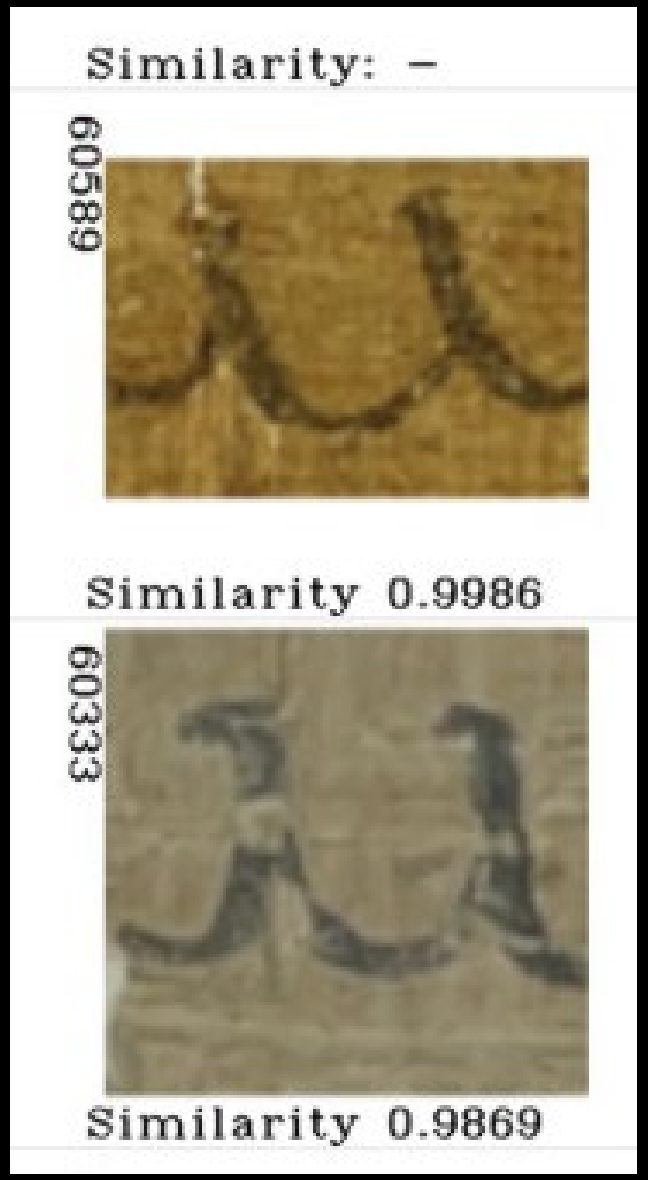


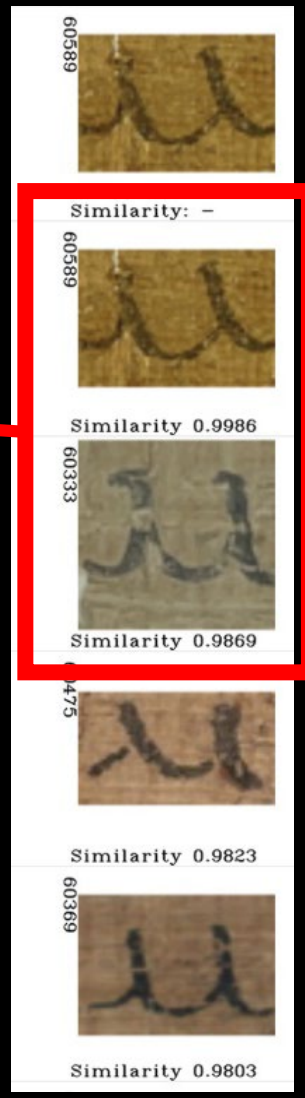


Similarity estimation results

Similarity estimation results

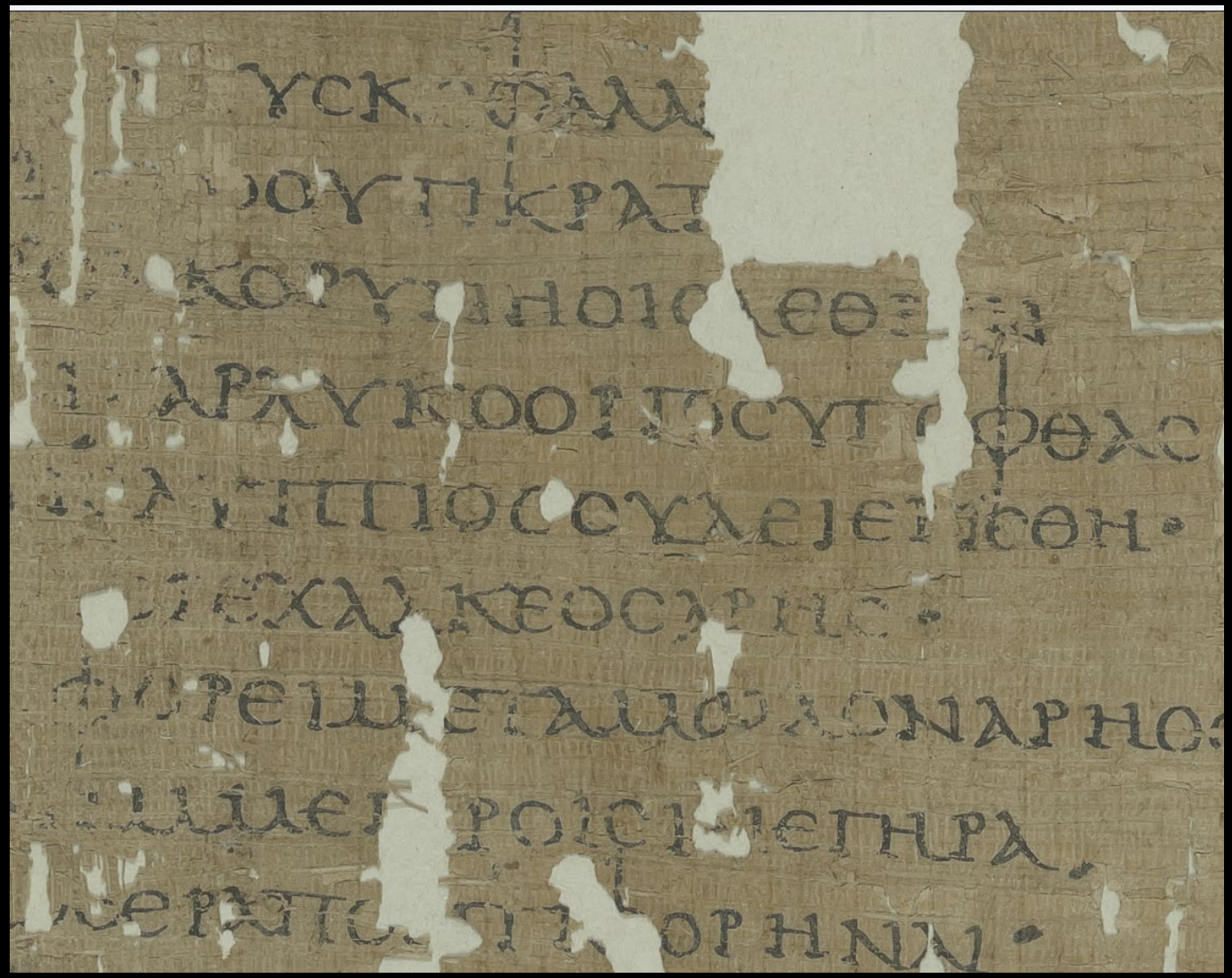
Similarity estimation results



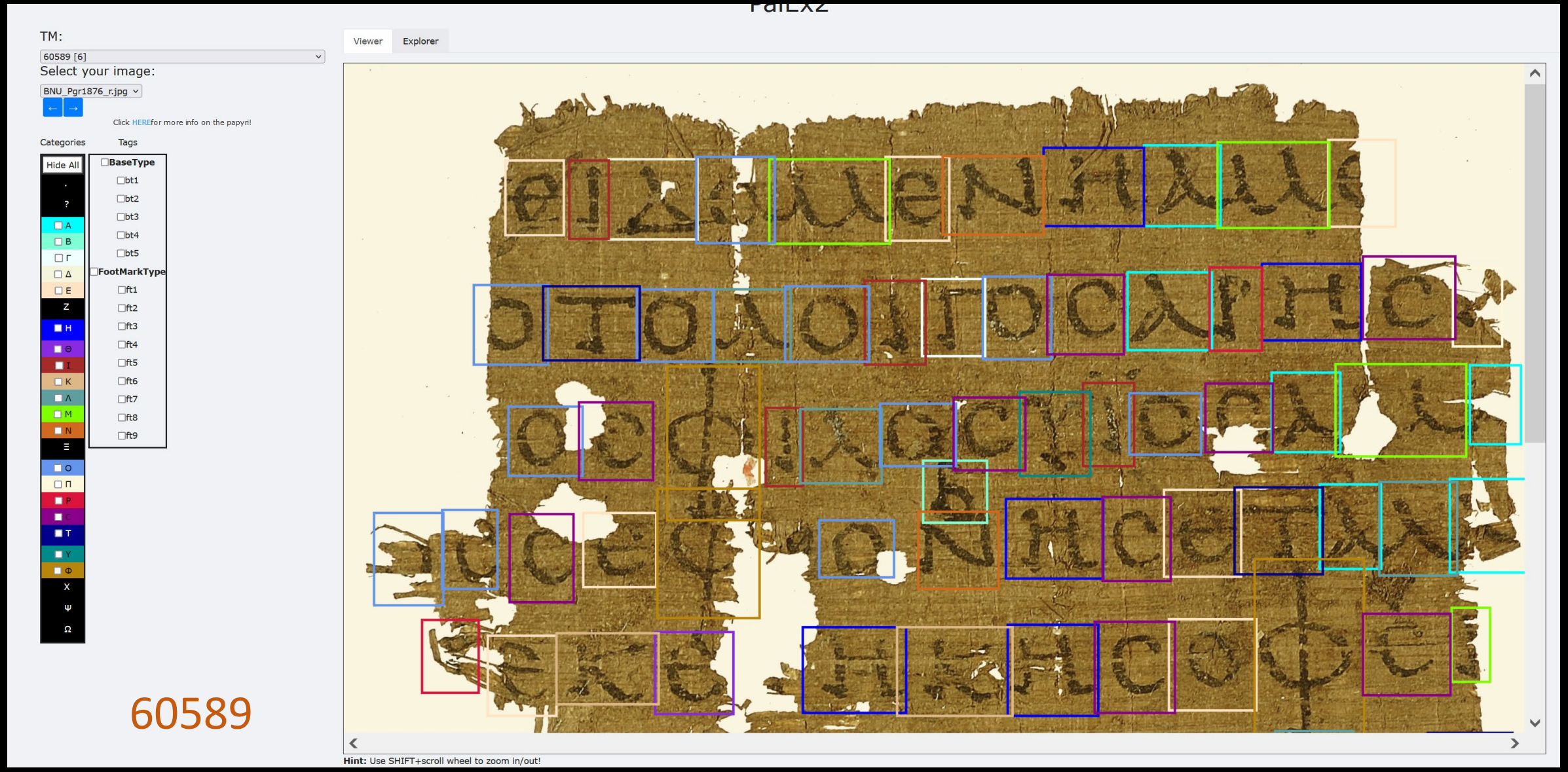


60589

60333



https://showcase.d-scribes.philhist.unibas.ch/viewer



60589

TM:

60589 [6]

Select your image:

BNU_Pgr1876_r.jpg

← →

Click HEREfor more info on the papyri!

Categories Tags

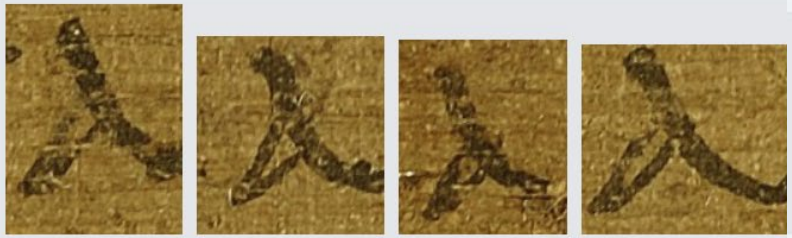
Hide All

.
?
A
B
Γ
Δ
E
Z
H
Θ
I
K
Λ
M
N
Ξ
O
Π
P
C
T
Y
Φ
X
Ψ
Ω

BaseType
bt1
bt2
bt3
bt4
bt5
FootMarkType
ft1
ft2
ft3
ft4
ft5
ft6
ft7
ft8
ft9

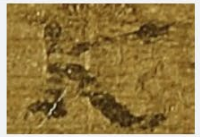
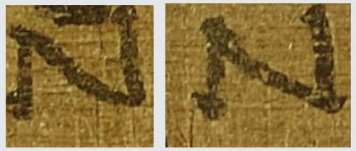
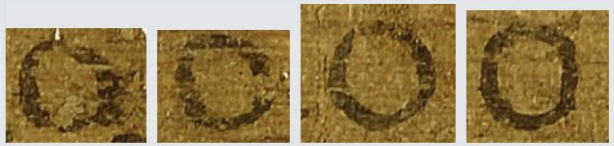
Viewer Explorer

Categories	BaseType
	bt1
.	
?	
A	
B	
Γ	
Δ	
E	
Z	
H	
Θ	
I	

A		
B		
Γ		
Δ		
E		
Z		
H		
Θ		
I		

60589

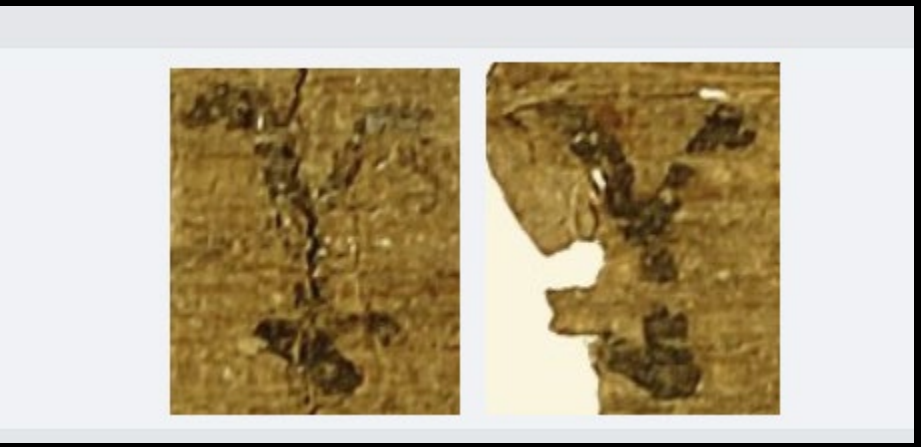
60333

Κ	
Λ	
Μ	
Ν	
Ξ	
Ο	
Π	
Ρ	
Σ	
Τ	
Χ	

60589

60333

60333



60589

ICDAR2023 Competition on Detection and Recognition of Greek Letters on Papyri

TM131_P.Batav.6_frag

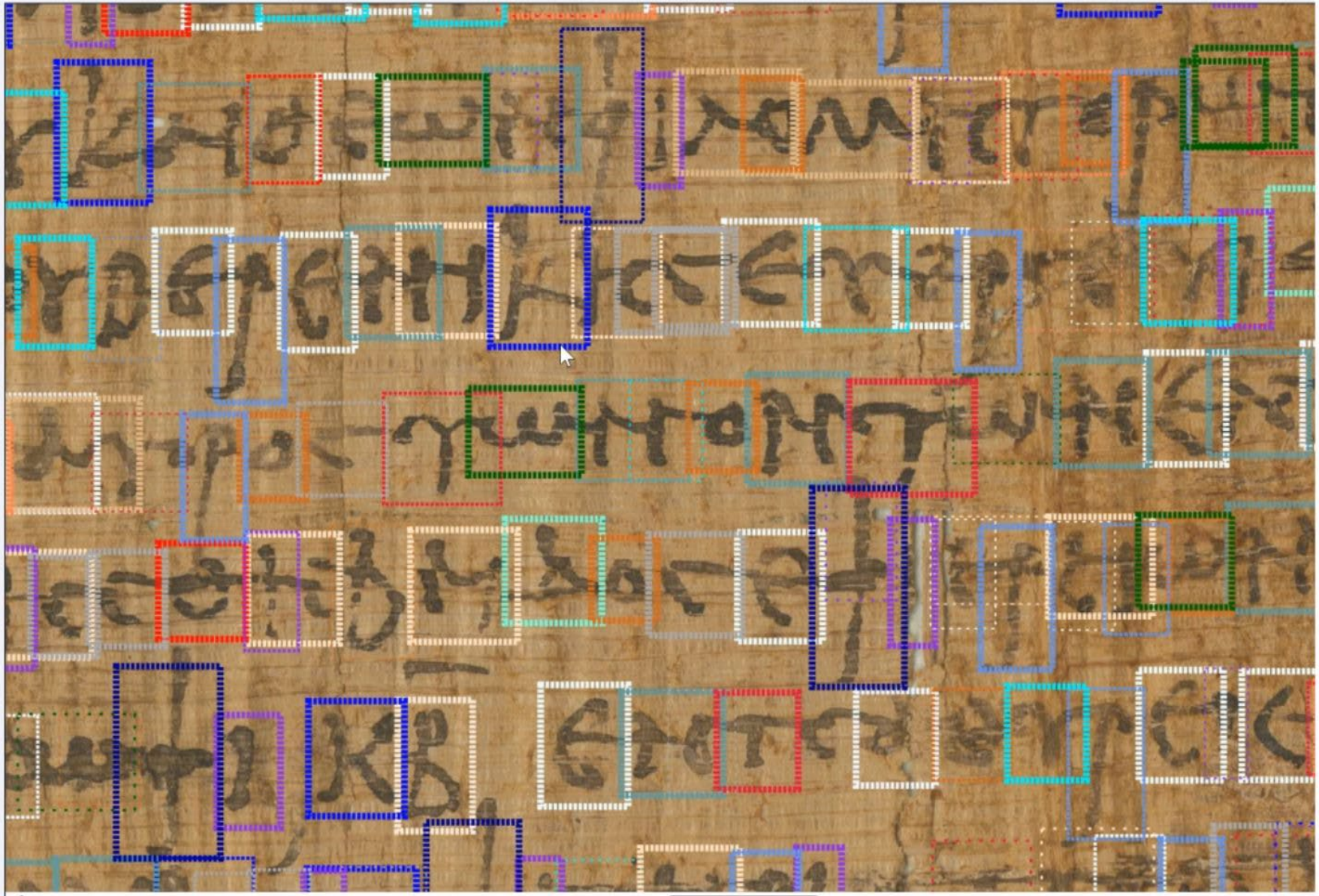
← →

Click HEREfor more info on the papyri!

Categories

Hide All

- A
- B
- Γ
- Δ
- E
- Z
- H
- Θ
- K
- I
- Λ
- M
- N
- Ξ
- O
- Π
- P
- Σ
- T
- Y
- Φ
- X
- Ψ

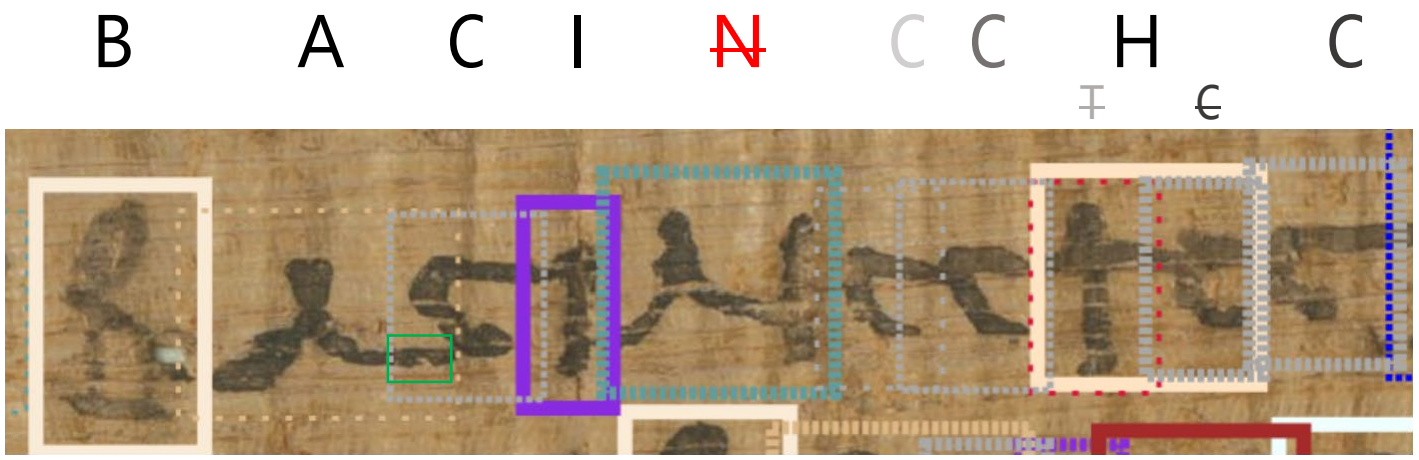


Hint: Use SHIFT+scroll wheel to zoom in/out!

Human tool (READ) vs AI Model (PapyTwin)

READ

PapyTwin



Categories

Hide All

- A
- B
- Γ
- Δ
- E
- Z
- H
- Θ
- K
- I
- Λ
- M
- N
- Ξ
- O
- Π
- P
- Σ
- T
- Y
- Φ
- X
- Ψ
- C
- Ω

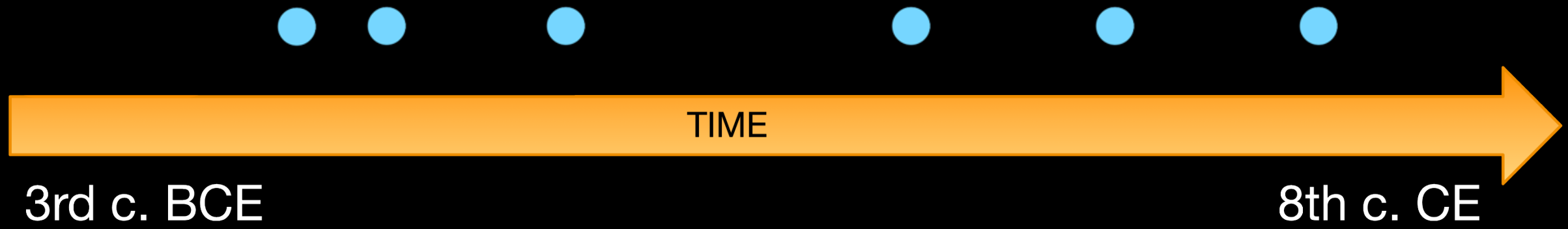
EGRAPSA:
Retracing the evolutions of handwritings in Greco-Roman Egypt thanks to digital palaeography

June 2023-May 2028
SNSF, University of Basel

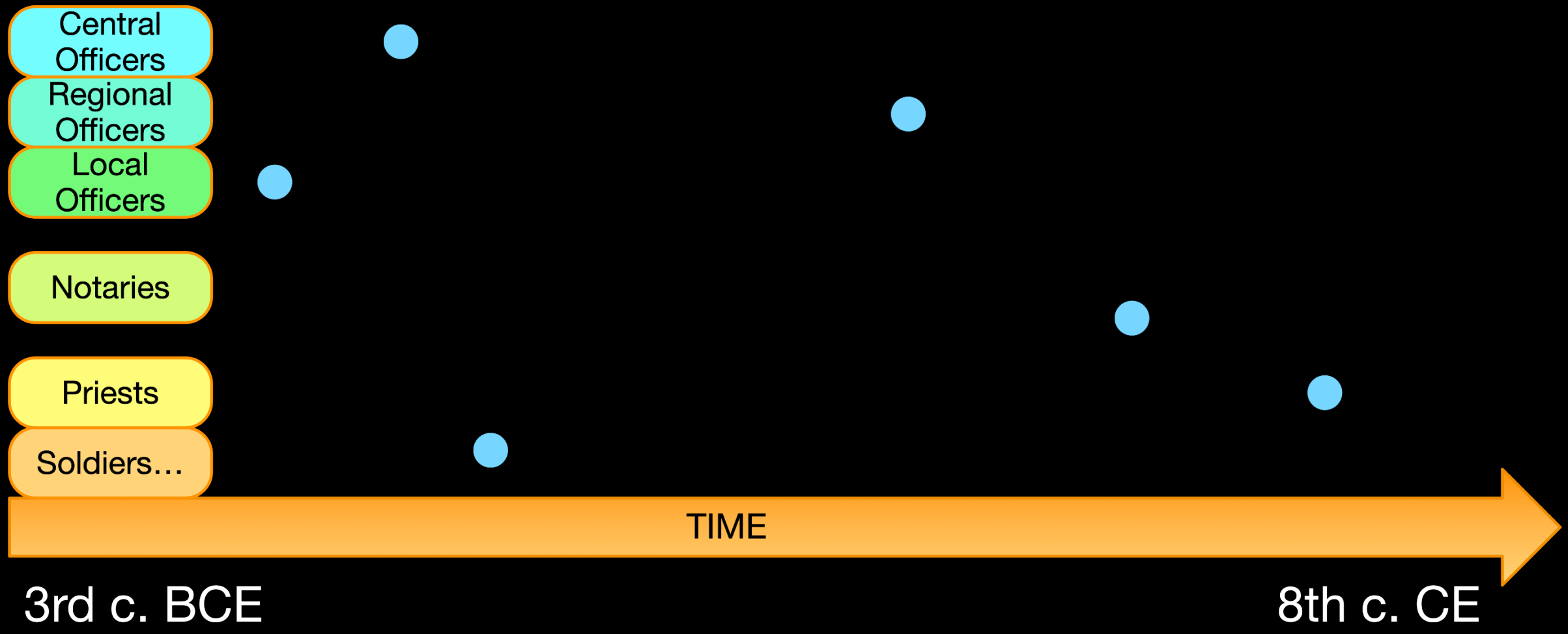
Project goal of EGRAPSA (“I have written”)

A comprehensive history
of the evolutions
of Greek handwritings
in their social and cultural contexts

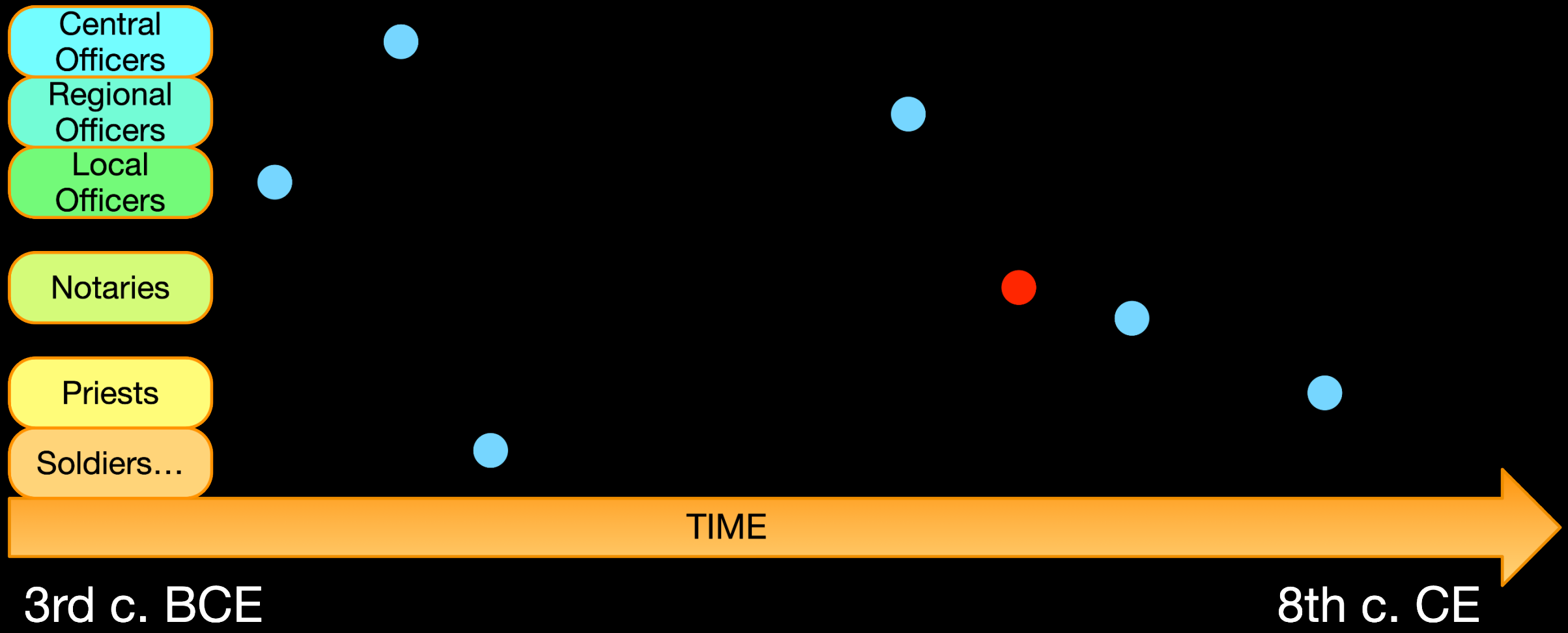
Diachronic perspective: sound anchor points from dated texts



Diachronic perspective: sound anchor points from dated texts + Synchronic perspective: socio-cultural variations among writers



Diachronic perspective: sound anchor points from dated texts + Synchronic perspective: socio-cultural variations among writers



Concluding remarks

- Datasets need to be carefully prepared

Concluding remarks

- Datasets need to be carefully prepared

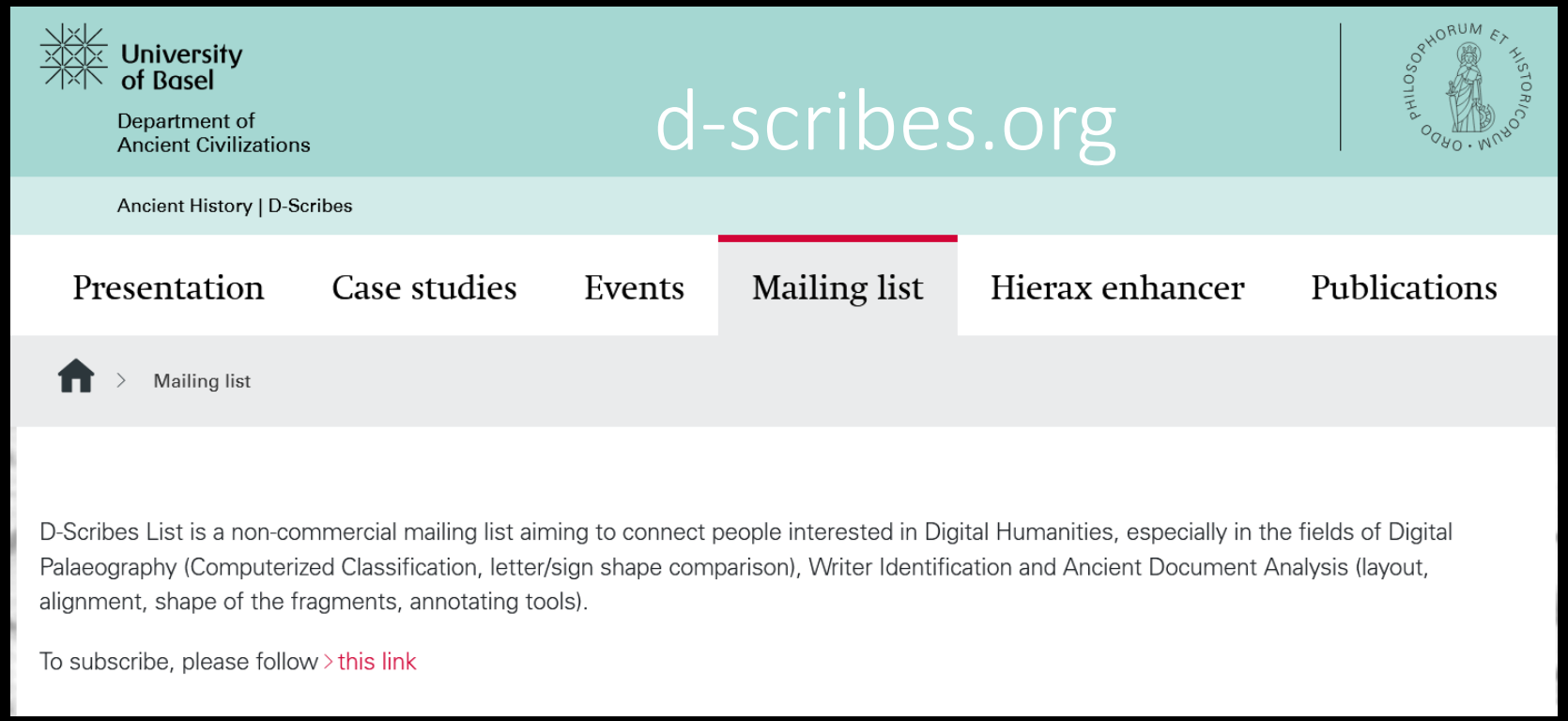
Concluding remarks

- Datasets need to be carefully prepared
- Challenge of interdisciplinary research (different speeds, in-depth analysis)

Concluding remarks

- Datasets need to be carefully prepared
- Challenge of interdisciplinary research (different speeds, in-depth analysis)
- Benefit both for Papyrology and Computer Science

D-scribes mailing list



Aims at connecting people interested in Digital Humanities:

- Digital Palaeography (Computerized Classification, letter/sign shape comparison)
- Writer Identification
- Ancient Document Analysis (layout, alignment, shape of the fragments, annotating tools)

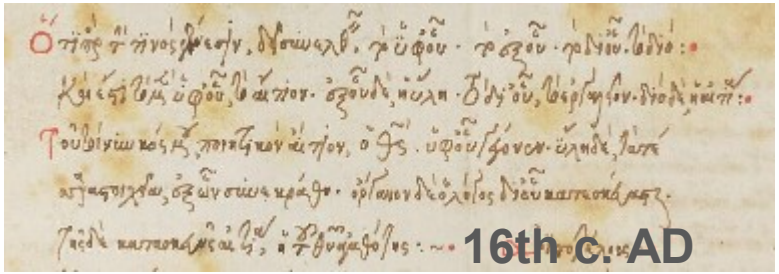
HTR and OCR from papyrus to codex

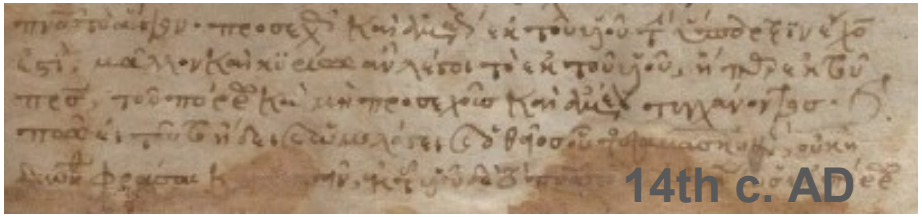
SunoikisisDC Digital Classics and Byzantine Studies: Session 4

Paraskevi Platanou (National and Kapodistrian University of Athens)
&
John Pavlopoulos (Athens University of Economics and Business)

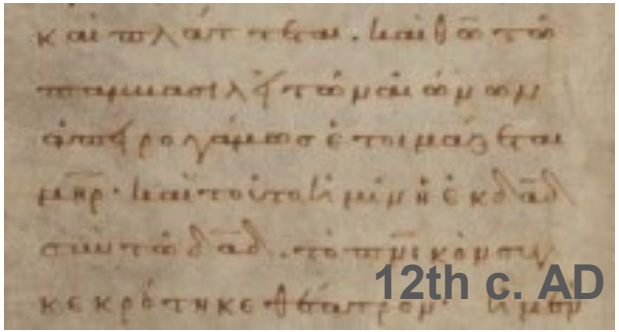
Byzantine language

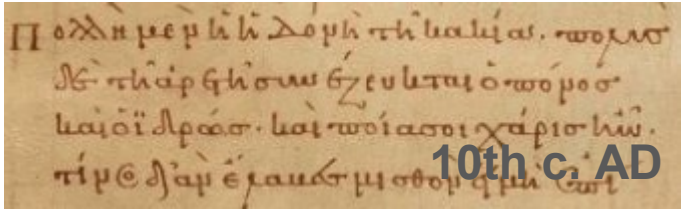
“A linguistic variety in its own right”

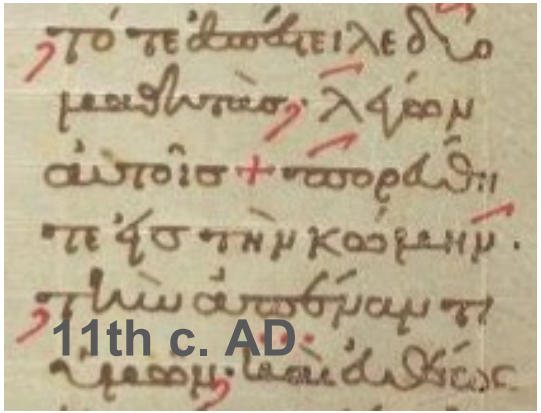

16th c. AD


14th c. AD

Byzantine manuscripts and scripts


12th c. AD


10th c. AD


11th c. AD

Paraskevi Platanou, John Pavlopoulos, and Georgios Papaioannou. 2022.
Handwritten Paleographic Greek Text Recognition: A Century-Based Approach.
In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6585–6589, Marseille, France. European Language Resources Association.

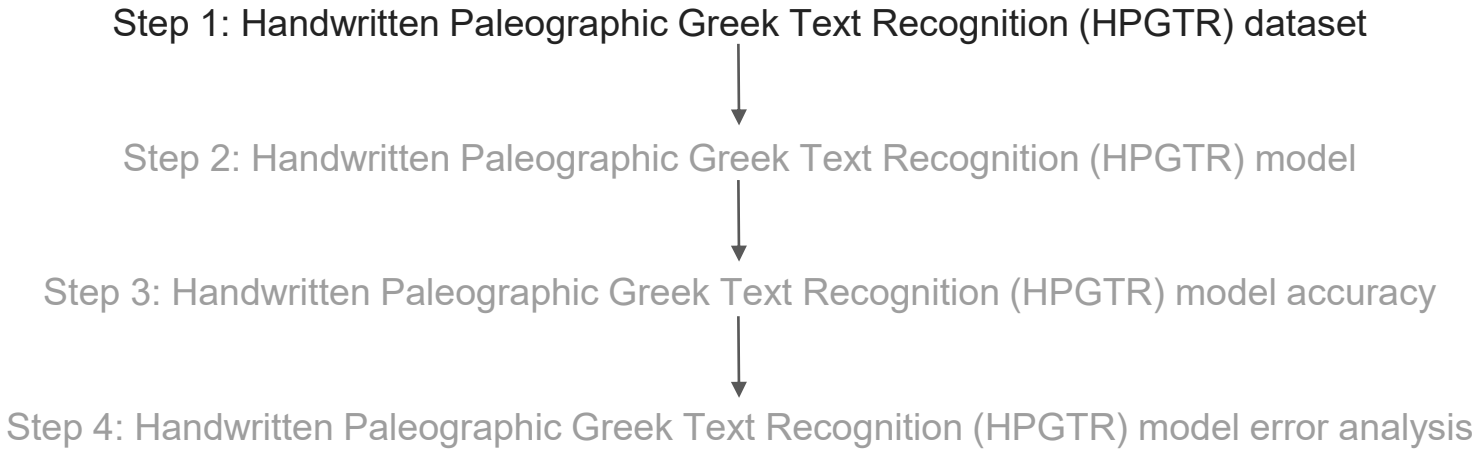


Fig.1

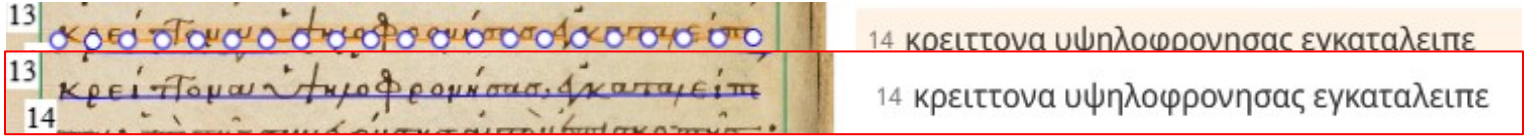
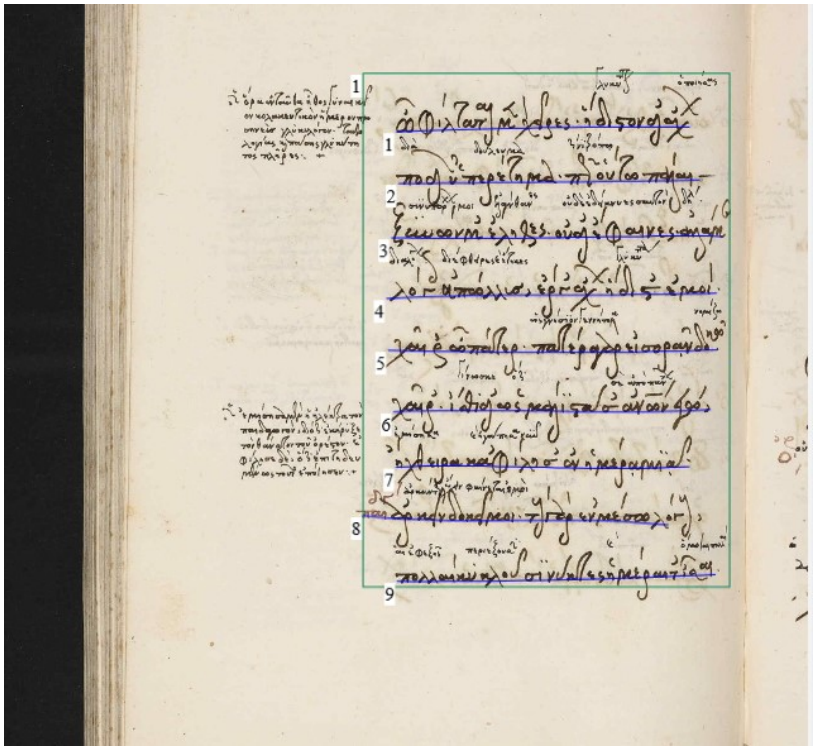


Fig.2



How do we effectively transcribe?

How do we manage effectively the layout?

Paraskevi Platanou, John Pavlopoulos, and Georgios Papaioannou. 2022.
Handwritten Paleographic Greek Text Recognition: A Century-Based Approach.
In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6585–6589, Marseille, France. European Language Resources Association.

Step 1: Handwritten Paleographic Greek Text Recognition (HPGTR) dataset

↓

Step 2: Handwritten Paleographic Greek Text Recognition (HPGTR) model

↓

Step 3: Handwritten Paleographic Greek Text Recognition (HPGTR) model accuracy

↓

Step 4: Handwritten Paleographic Greek Text Recognition (HPGTR) model error analysis

Paraskevi Platanou, John Pavlopoulos, and Georgios Papaioannou. 2022.
Handwritten Paleographic Greek Text Recognition: A Century-Based Approach.
In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6585–6589, Marseille, France. European Language Resources Association.

Step 1: Handwritten Paleographic Greek Text Recognition (HPGTR) dataset

↓

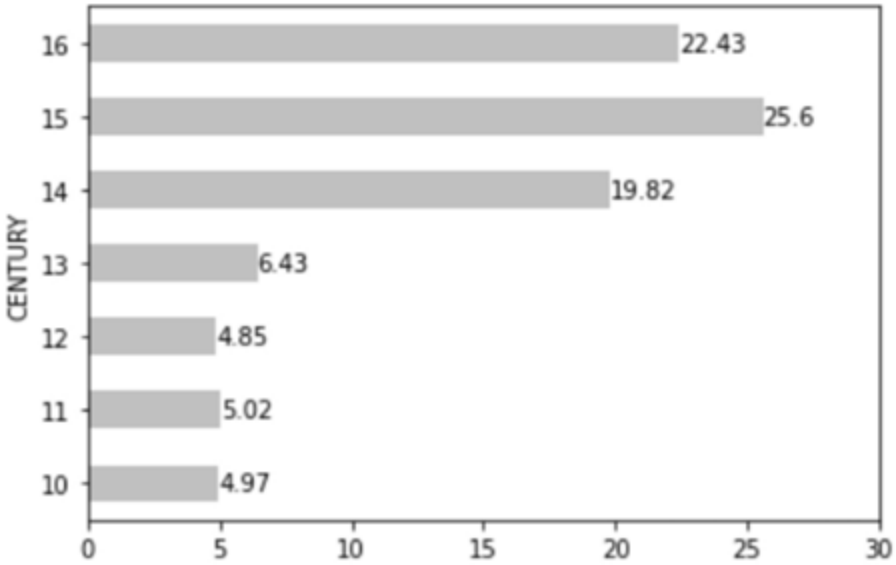
Step 2: Handwritten Paleographic Greek Text Recognition (HPGTR) model

↓

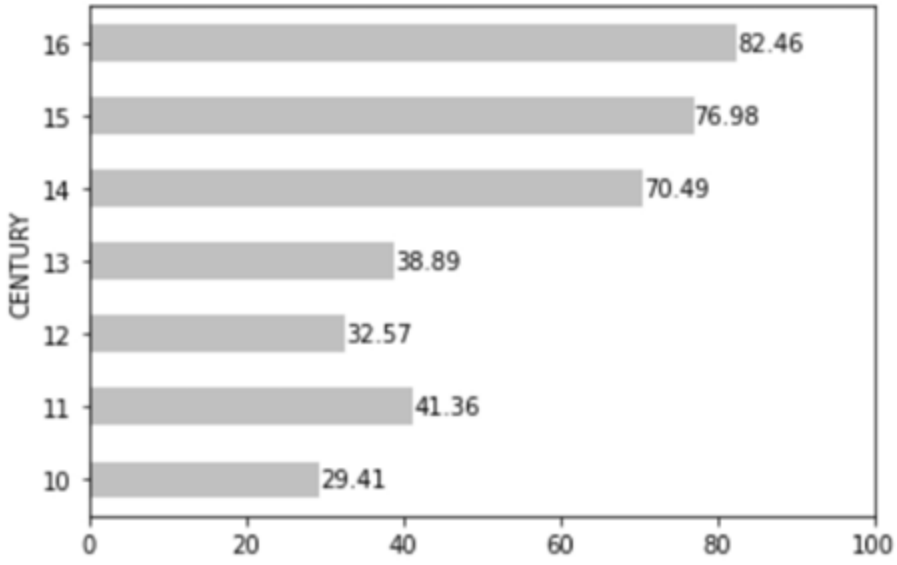
Step 3: Handwritten Paleographic Greek Text Recognition (HPGTR) model accuracy

↓

Step 4: Handwritten Paleographic Greek Text Recognition (HPGTR) model error analysis



Character error rate (percent shown horizontally) in HPGTR.N per century



Word error rate (percent shown horizontally) in HPGTR.N per century

Paraskevi Platanou, John Pavlopoulos, and Georgios Papaioannou. 2022.
Handwritten Paleographic Greek Text Recognition: A Century-Based Approach.
In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6585–6589, Marseille, France. European Language Resources Association.

Step 1: Handwritten Paleographic Greek Text Recognition (HPGTR) dataset

↓

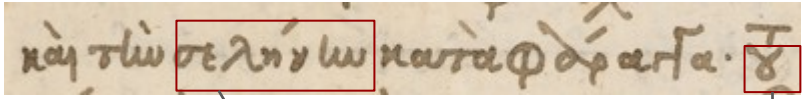
Step 2: Handwritten Paleographic Greek Text Recognition (HPGTR) model

↓

Step 3: Handwritten Paleographic Greek Text Recognition (HPGTR) model accuracy

↓

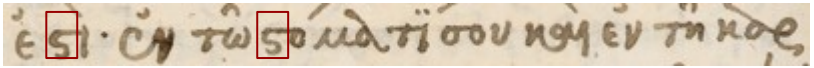
Step 4: Handwritten Paleographic Greek Text Recognition (HPGTR) model error analysis



σελήν ην

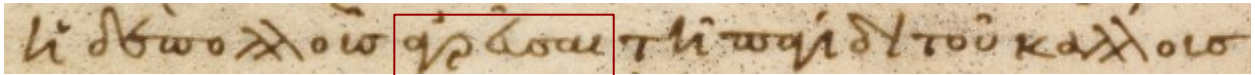
ου

σελήνην



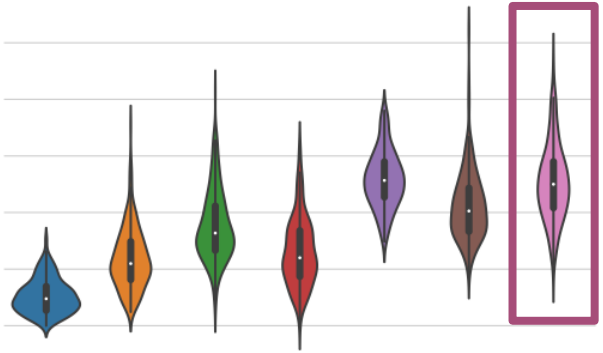
στ στ

ει



ευ

Error analysis: the 16th c. CE

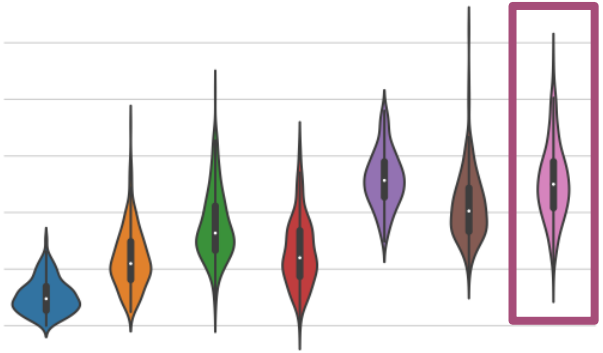


highest avg. CER

‘πλοῦς,’ → HTR → ‘πλουοο’’

Pavlopoulos, et al. 2023 EMNLP

Error analysis: the 16th c. CE

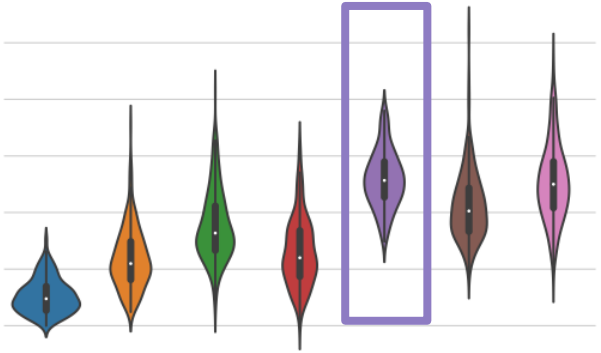


highest avg. CER

‘ῶ πατρὶς’ → HTR → ‘ὤπαρις’

Pavlopoulos, et al. 2023 EMNLP

Error analysis: the 14th c. CE



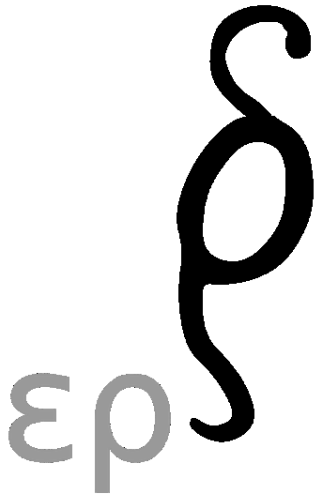
2nd highest

‘φ**έρ** ’ ἐπ ’ αὐτὴν ἴω’

HTR

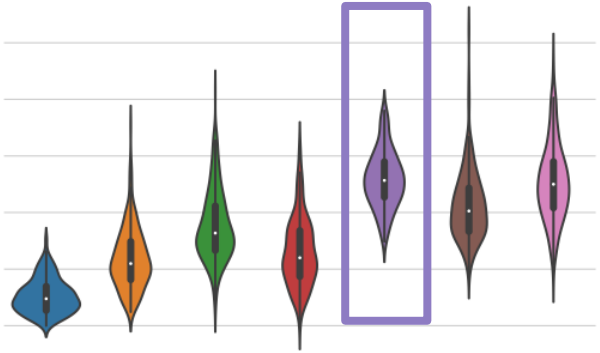
‘φ**ρ**επαυ την ιςω

The ligature of ‘ερ’ is not captured



Pavlopoulos, et al. 2023 EMNLP

Error analysis: the 14th c. CE



2nd highest

omega: horizontal eight-shaped

‘φέρ᾽ ἐπ᾽ αὐτ

R

‘φρεπαυ την ισω’

Pavlopoulos, et al. 2023 EMNLP

Detecting Erroneous HTR output

- HTR output yields diverse error rates

  ⇒ manual (tedious, expensive) correction

    ⇒ delaying the preservation of manuscripts

      ⇒ hindering the recognition of historical manuscripts

Pavlopoulos, et al. 2023 EMNLP

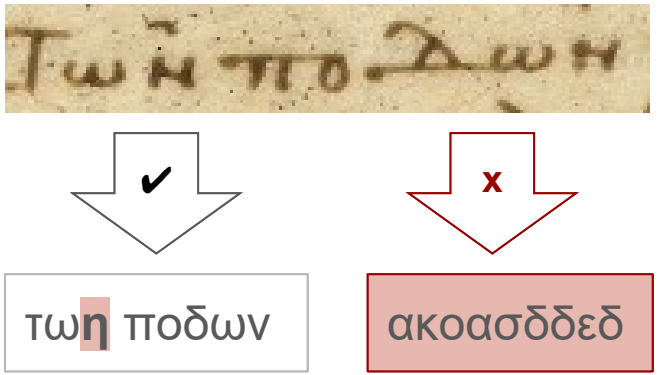
Detecting Erroneous HTR output

- HTR output yields diverse error rates
- Detecting erroneous/flawless HTR output
  - easier for sbarecmld ttxe v. non-scrambled text

Pavlopoulos, et al. 2023 EMNLP

Detecting Erroneous HTR output

- HTR output yields diverse error rates
- Detecting erroneous/flawless HTR output
  - easier for sbarecmld ttxe v. non-scrambled text



Pavlopoulos, et al. 2023 EMNLP

The Greek language: from antiquity to modern

10th CE to 16th c. CE

Contemporary: ~ spoken language

Ancient: including Atticised Greek

⇒ NLP with both wins

Pavlopoulos, et al. 2023 EMNLP

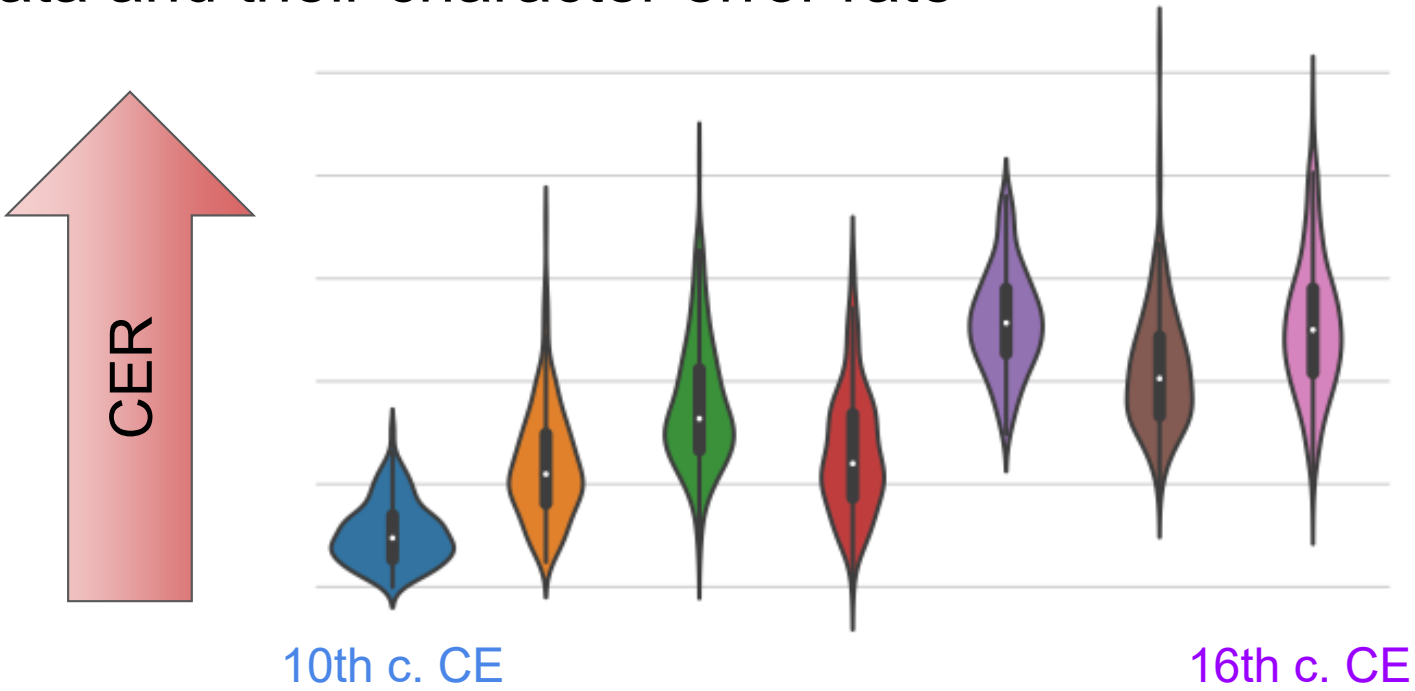
The data source

- under-trained Transkribus
- transcribed & recognised lines

Transcription	Recognition
ἐγγινομένα πάθη μὴ σβεννύντες ἀλλὰ τῆ εκλύσει (the born-in passions not extinguishing but the release)	ἐγγενομεναπαδημησμεννωτες ἀλλατῆε κλησει
τοῦ βίου τοῦ καθ ΄ εαυτοὺς πολλὰ γίνεσθαι συγχωροῦν (of the life of themselves many happening forgive)	του β ου του καλεαυτοὺς πολλαγινεσθαι συγχωρ όν
τες ἐμπυρίζουσι τὸν ἀμπελῶνα ἀλλὰ καὶ ὁ διὰ τες (- set on fire the vineyard but and the due to the)	εμπυριζου σιμαμπελῶνα ἀλλακαι ὅδξα

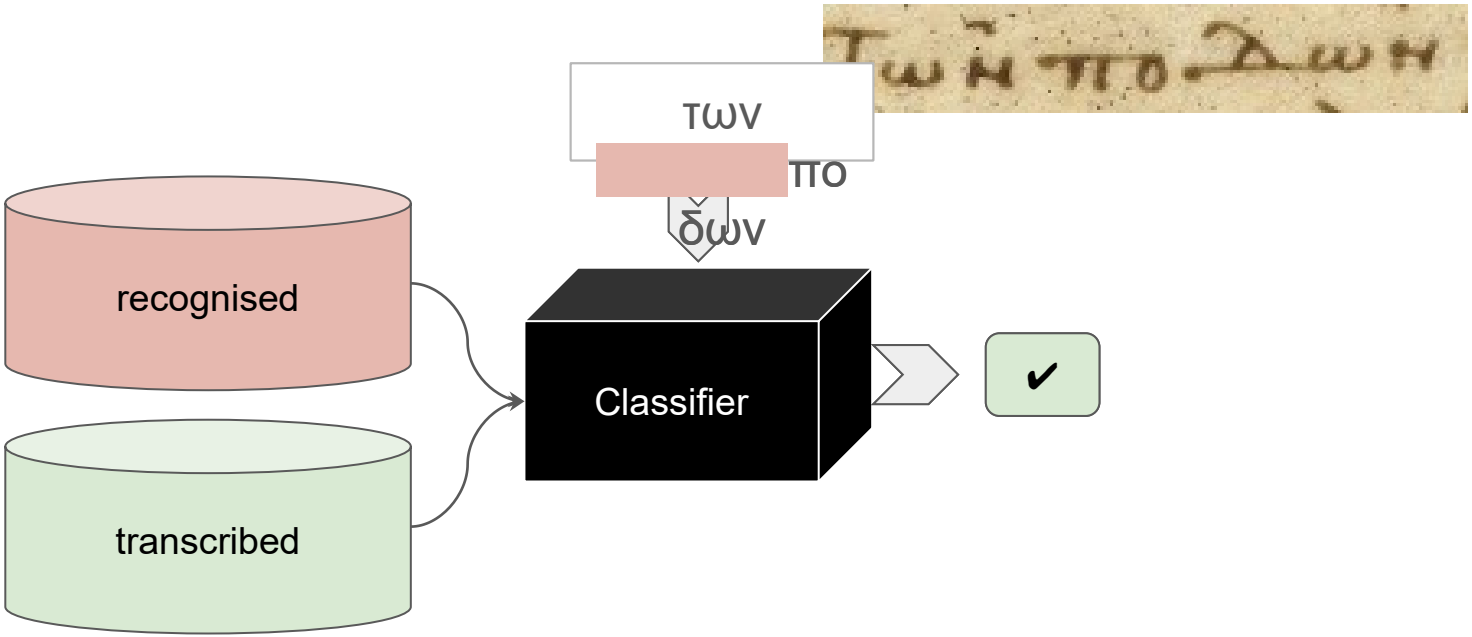
Pavlopoulos, et al. 2023 EMNLP

The data and their character error rate



Pavlopoulos, et al. 2023 EMNLP

Experiments: setup



Pavlopoulos, et al. 2023 EMNLP

Experiments: machine learning

	AP	AUC	F1 (+)	F1 (-)
Random	0.52	0.50	0.49	0.47
SVM	0.66	0.65	0.60	0.51
Forest	0.64	0.65	0.64	0.50
MLP	0.79	0.79	0.73	0.69

Pavlopoulos, et al. 2023 EMNLP

Experiments: deep learning

	AP	AUC	F1 (+)	F1 (-)
Random	0.52	0.50	0.49	0.47
SVM	0.66	0.65	0.60	0.51
Forest	0.64	0.65	0.64	0.50
MLP	0.79	0.79	0.73	0.69
GRU	0.79	0.79	0.68	0.71
GreekBERT:M	0.95	0.94	0.88	0.88
GreekBERT:M+A	**0.97**	**0.97**	**0.90**	**0.91**

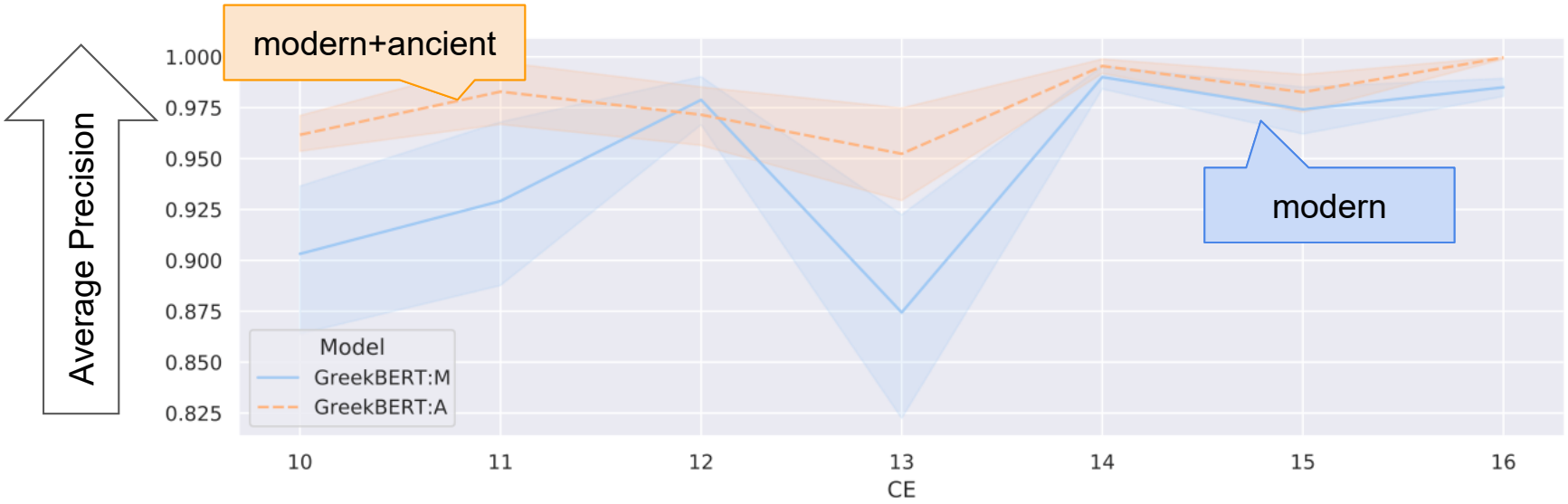
Pavlopoulos, et al. 2023 EMNLP

Experiments: deep learning

	AP	AUC	F1 (+)	F1 (-)
Random	0.52	0.50	0.49	0.47
SVM	0.66	0.65	0.60	0.51
Forest	0.64	0.65	0.64	0.50
MLP	0.79	0.79	0.73	0.69
GRU	0.79	0.79	0.68	0.71
GreekBERT:M	0.95	0.94	0.88	0.88
GreekBERT:M+A	**0.97**	**0.97**	**0.90**	**0.91**

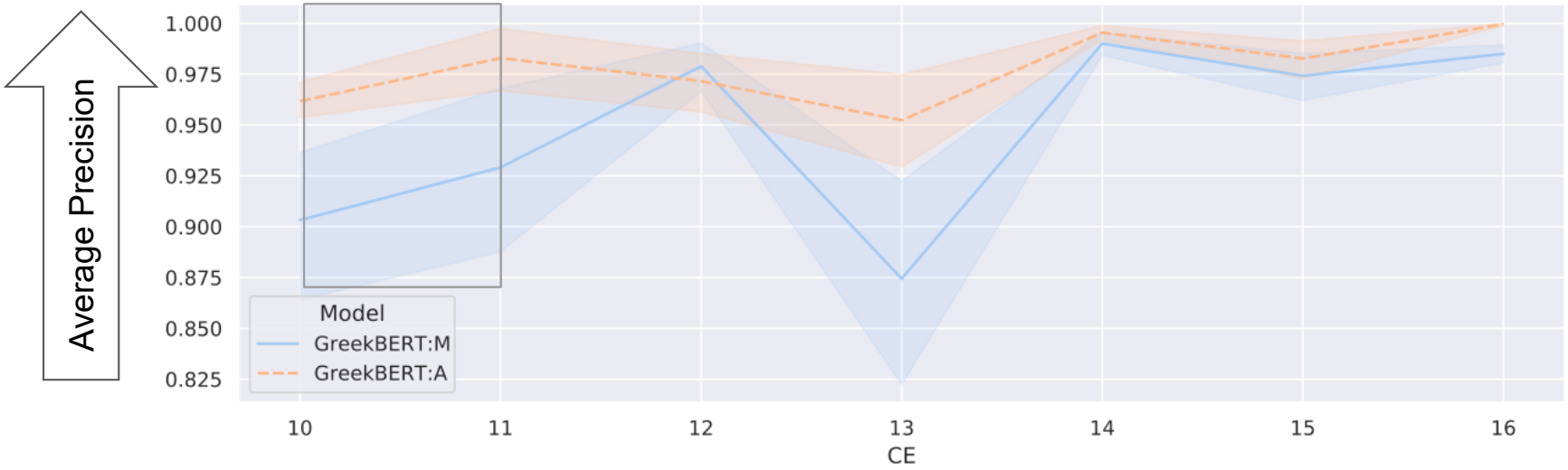
Pavlopoulos, et al. 2023 EMNLP

Experiments: evaluation per century



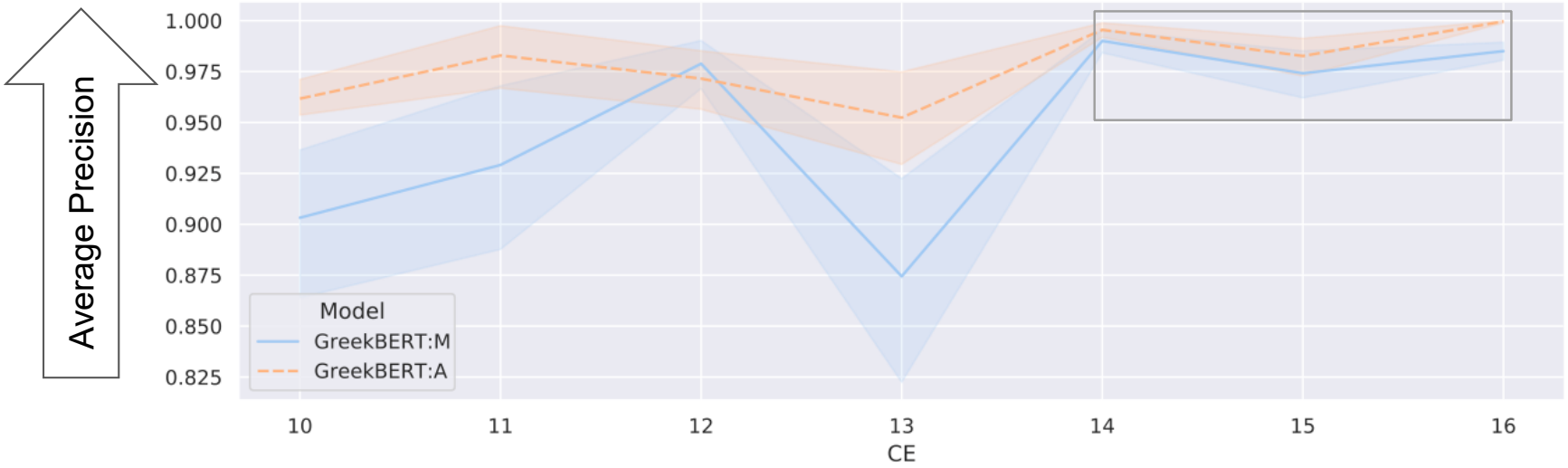
Pavlopoulos, et al. 2023 EMNLP

Experiments: evaluation per century



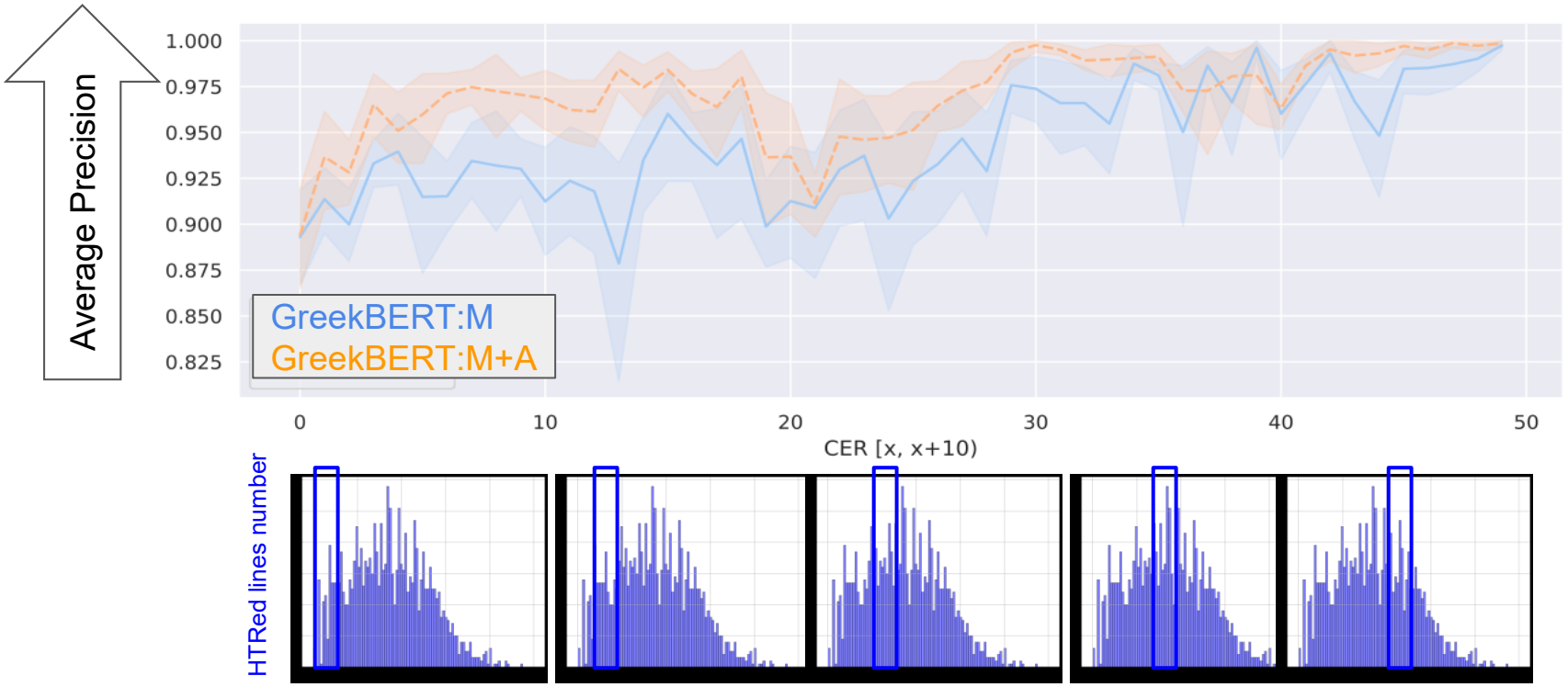
Pavlopoulos, et al. 2023 EMNLP

Experiments: evaluation per century



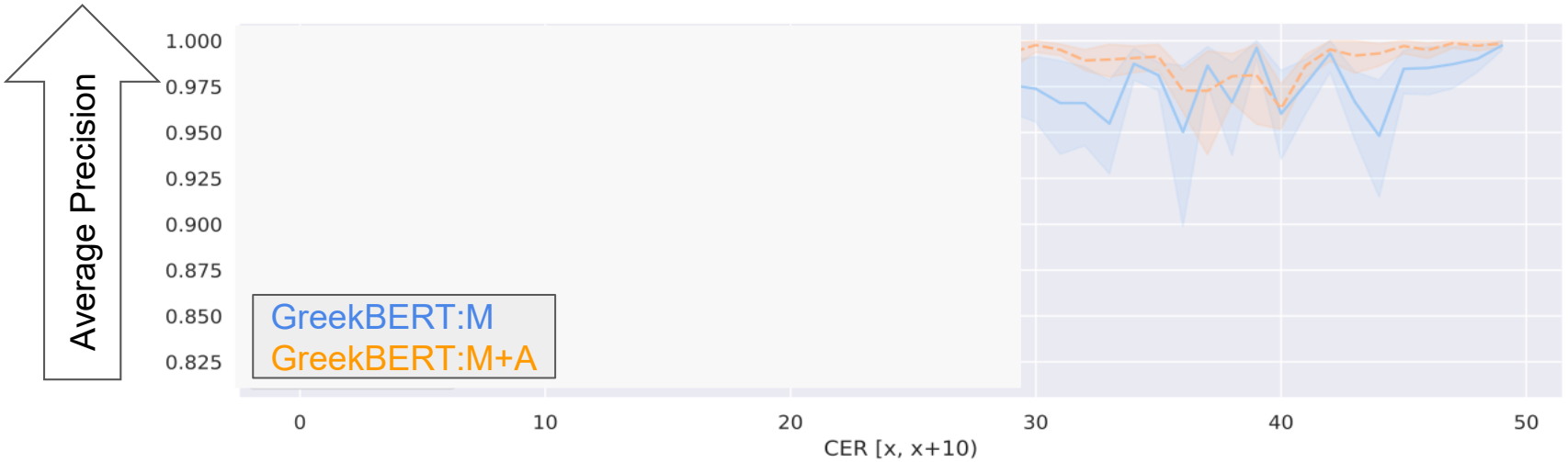
Pavlopoulos, et al. 2023 EMNLP

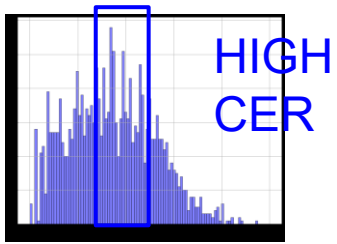
Experiments: evaluation per error zone



Pavlopoulos, et al. 2023 EMNLP

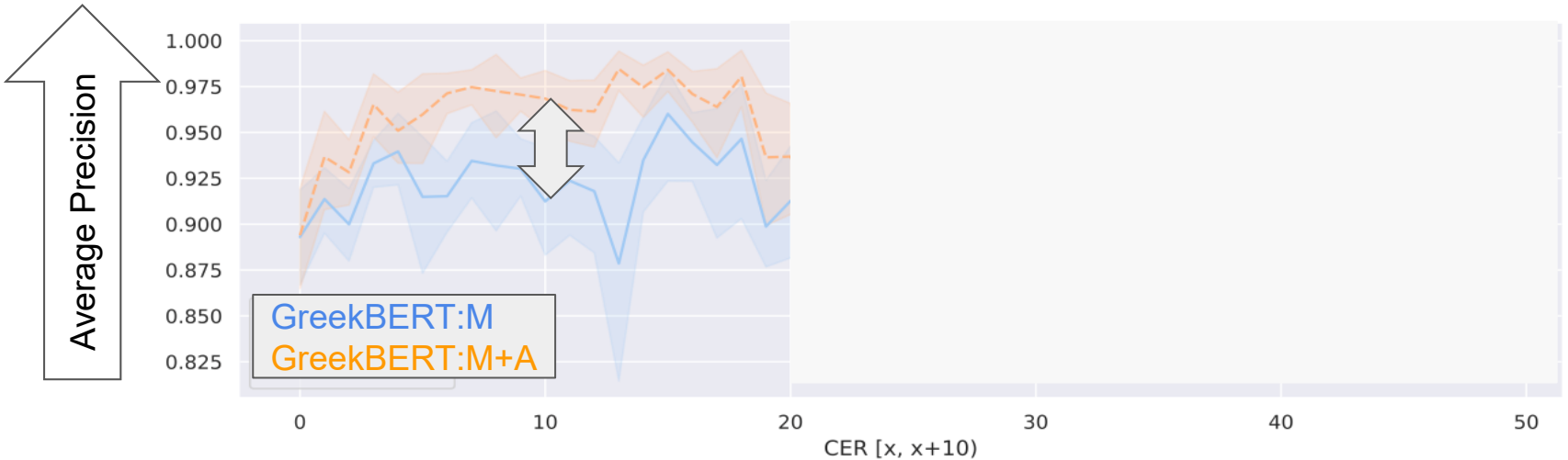
Experiments: evaluation per error zone



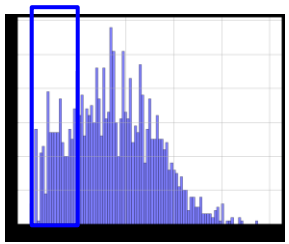


Pavlopoulos, et al. 2023 EMNLP

Experiments: evaluation per error zone



LOW CER



Pavlopoulos, et al. 2023 EMNLP

Error analysis

ἐγὼ δ ἀεί πως φιλακόλουθός εἰμι
(I am always readily following)

Pavlopoulos, et al. 2023 EMNLP

Error analysis

λα και γεδεων εκ των σκυλων των ισμαηληι
(- and Gedeon from the dogs the -)

[...] ισμαηλιτι
-κων [...]

Pavlopoulos, et al. 2023 EMNLP

Decoding Ancient Greek: Handwritting Text Recognition and the Palatine Anthology

Sunoikisis Session. HTR and OCR from papyrus to codex

Maxime Guénette
maxime.guenette@umontreal.ca
Université de Montréal

29/04/2024

Codex Palatinus graecus 23

- Palatine Anthology ≠ Greek Anthology
- Universitätsbibliothek, Ruprecht-Karls-Universität (Heidelberg)
- Xth century manuscript with old round minuscule
- Second half in Paris (BNF), *Parisinus Supplementum Graecum 384*

Issues

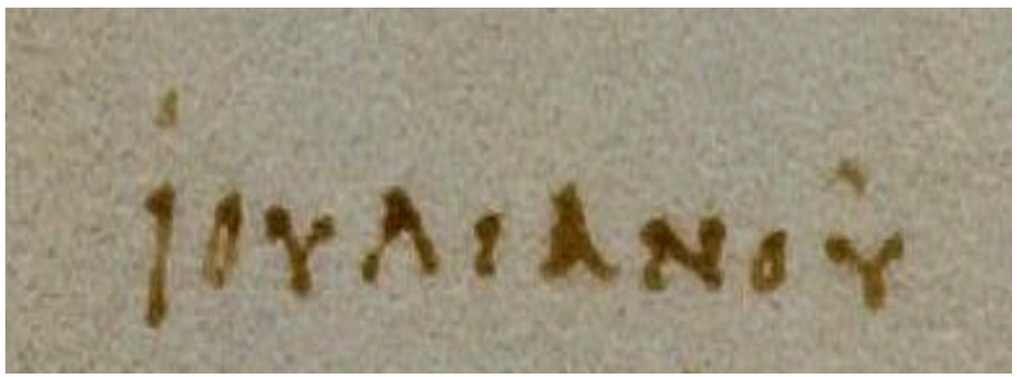
- More and more Ancient Greek manuscripts, including the Palatine Anthology, are being digitised
- ≈ 2,5% of surviving Ancient Greek literature
- Restricted access to the text of these manuscripts
- Text search and (very) limited computational linguistic studies

Solution

Handwritten Text Recognition

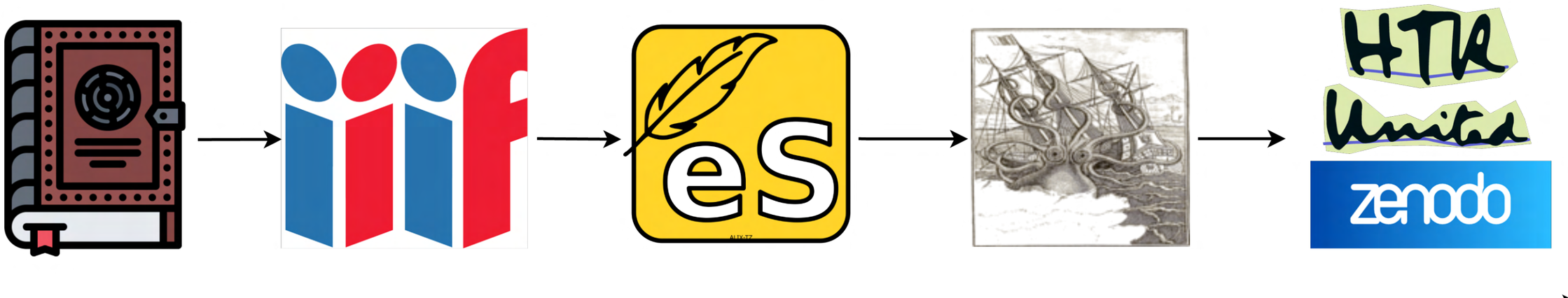
State of the art

- Very little ground truth available for Ancient Greek
- Old, outdated or unavailable HTR models
- Non-latin scripts are less studied



= ἰουλιανοῦ

Palatine Anthology pipeline



bliotheca Palatina –
gital

Watch list 0 | Login

Universitätsbibliothek Heidelberg, Cod. Pal. graec. 23

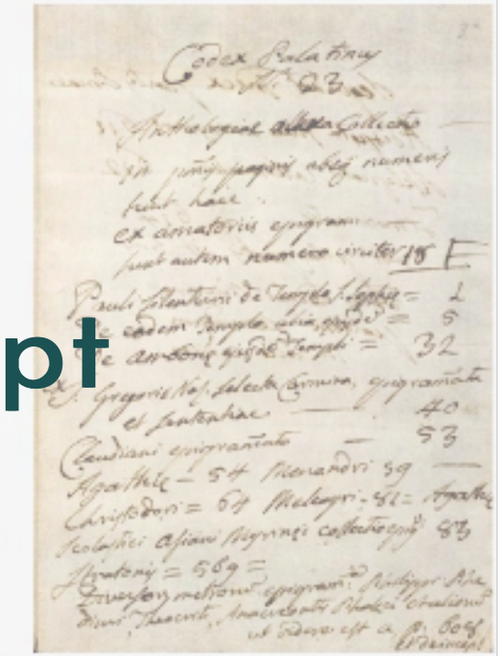
Anthologia Palatina

Konstantinopel, Mitte 10. Jh.

Teil 2 (fol. 615-662): Paris: Bibliothèque nationale de France, Supplément grec 384

- Manuscript description
- Diktyon Nr. 32453 (Pinakes entry)
- CRC 933 (Sub-Project A01, U24): Epigrams and the Byzantine Buildings of Constantinople
- Scholarly Annotations of the Digital Facsimile
- Bibliotheca Palatina

DOI / Citation link: https://doi.org/10.11588/diglit.3449
URN: urn:nbn:de:bsz:16-diglit-34495
Metadata: METS
IIIF Manifest: v2.1, v3.0



Digitisation of the manuscript

License: Public Domain Mark | Use/Order | Feedback | Download | Overview | Quire structure | 0

There are 500 annotations to individual pages of this facsimile. Individual pages with annotations are marked in "Overview" with the symbol.

- Annotation list

- You might also be interested in... (beta)

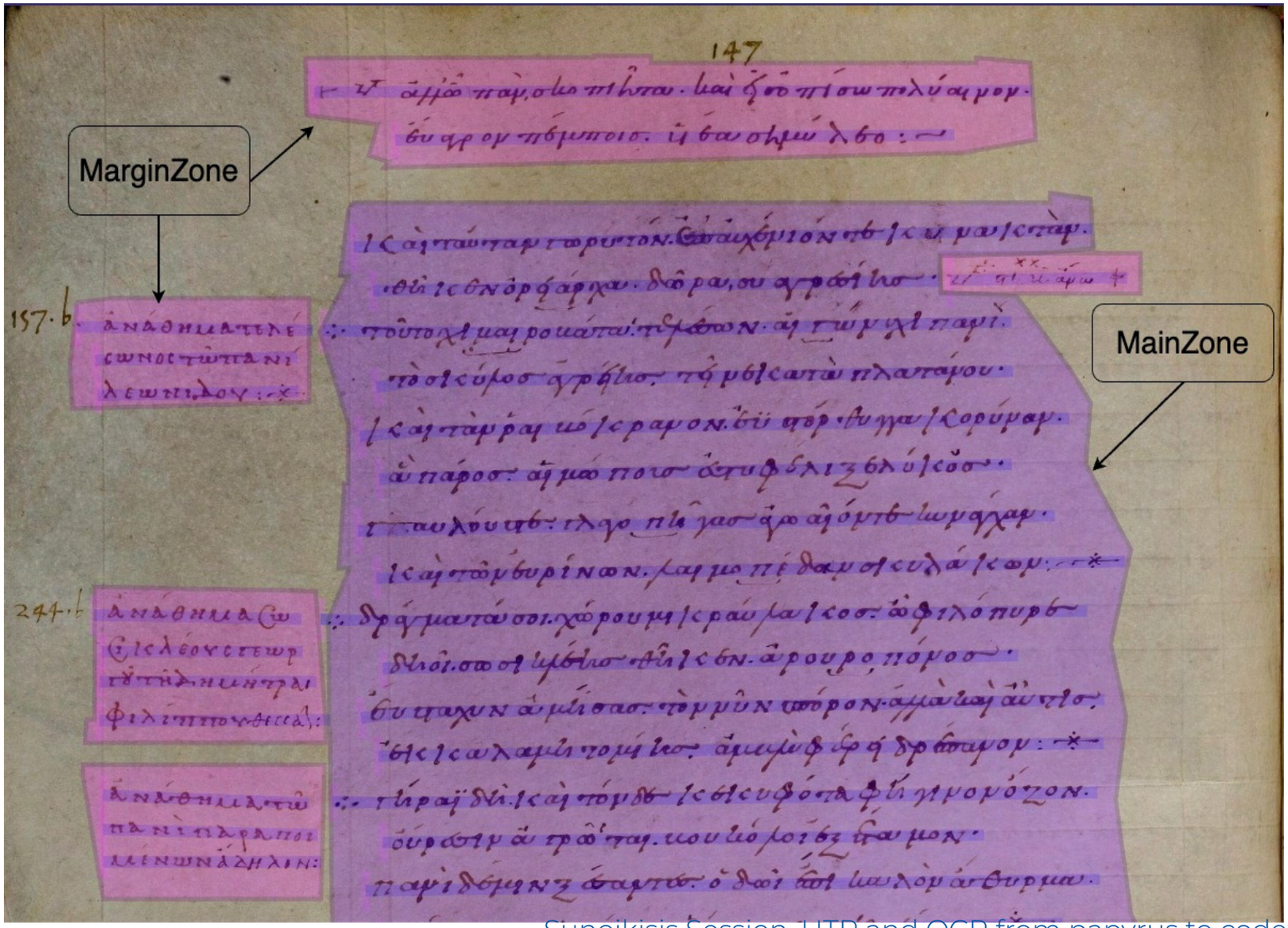
Image acquisition

- First sample of 50 pages (pp. 143-195), scribe A and corrector C
- Second sample of 20 pages (pp. 196-215), scribe A and corrector C
- Importing images using IIIF protocol

eScriptorium

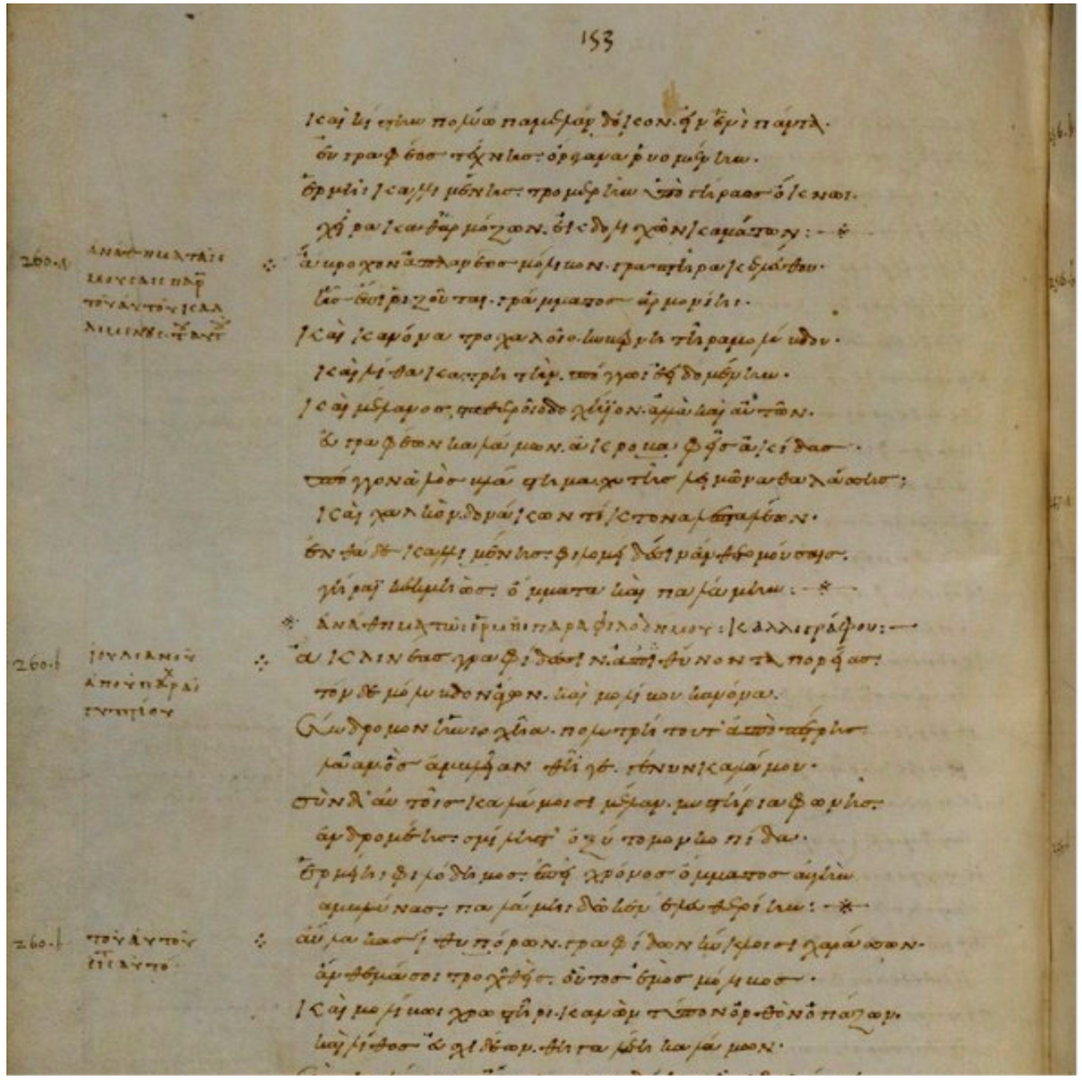
- Open-source platform
- Kraken graphical user interface
- Local or server-based installation
- Import of images (IIIF, JPG, PDF) and transcriptions (XML)
- Export of transcriptions (XML ALTO or PAGE), segmentation and handwriting recognition models

Layout Analysis



- SegmOnto ontology
- Regions: MainZone (epigram) and MarginZone (scholia)
- Lines: DefaultLine and InterlinearLine

Transcription



καὶ κίστην πολύωπα μελανδόκον· εἰν ἑνὶ πάντα·
εὐγραφέος τέχνησ· ὄργανα ῥυομένην·
ἑρμῆι καλλιμένησ· τρομερὴν ὑπὸ γήραοσ ὄκνωι
χεῖρα καθαρμόζων· ἐκ δολιχῶν καμάτων: ⁂

ἀνάθημα ταῖσ μούσαισ παρὰ τοῦ αὐτοῦ καλ λιμενουσ· του αυτου

∴ ἄβροχον ἁπλανέοσ μόλιβον· γραπτῆρα κελεύθου·
ἧσ ἔπι ῥιζοῦται· γράμματοσ ἁρμονίηι·
καὶ κανόνα τροχαλοῖο· κυβερνητῆρα μολίβδου·
καὶ λίθακα τρητήν· σπόγγωι ἐειδομένην·
καὶ μέλανοσ σταθεροῖο δοχῆϊον· ἀλλὰ καὶ αὐτῶν·
εὐγραφέων καλάμων· ἀκροβαφεῖσ ἀκίδασ·
σπόγγον ἀλὸσ βλάστημα· χυτῆσ λειμῶνα θαλάσσησ·
καὶ χαλκὸν δονάκων τέκτονα λεπταλέων·
ἐνθάδε καλλιμένησ φιλομειδέσιν ἄνθετο μούσαισ·
γήραϊ κεκμηὼσ· ὄμματα καὶ παλάμην: ⁂

⁂ ἀνάθημα τῶι ἑρμῆι παρὰ φιλοδήμου: καλλιγράφου: ~

ἰουλιανοῦ ἀπο ὑπαρ αἱ γυπτίου

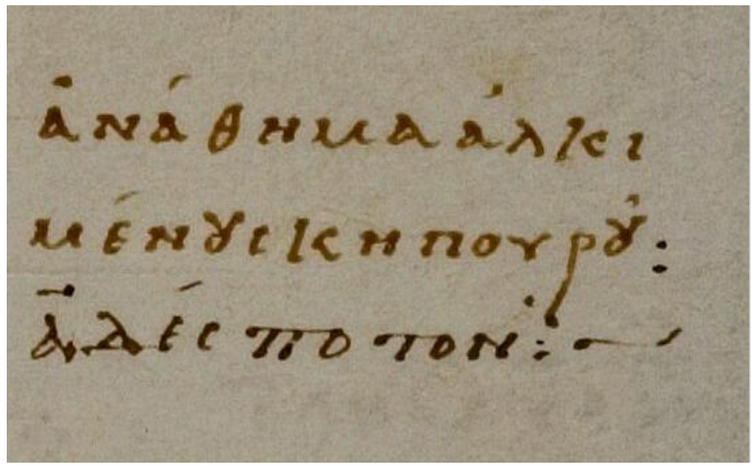
∴ ἀκλινέασ γραφίδεσσιν ἀπιθύνοντα πορείασ·
τόνδε μόλυβδον ἄγων· καὶ μολίβου κανόνα·
σύνδρομον ἡνιοχῆα· πολυτρήτου τ' ἀπὸ πέτρησ
λᾶαν· ὃσ ἀμβλεῖαν θῆγε· γένυν καλάμου·
σὺν δ' αὐτοῖσ καλάμοισι μέλαν· μυστήρια φωνῆσ·
ἀνδρομέησ· σμίλησ τ' ὀξυτόμον κοπίδα·
ἑρμείηι φιλόδημοσ· ἐπεὶ χρόνοσ ὄμματοσ αὐγὴν
ἀμβλύνασ· παλάμηι δῶκεν ἐλευθερίην: ⁂

τοῦ αὐτοῦ εἰσ τ αὐτό

∴ αὔλακασ ἰθυπόρων· γραφίδων κύκλοισι χαράσσων·
ἄνθεμά σοι τροχθεὶσ οὗτοσ ἐμὸσ μόλιβοσ·
καὶ μολέβωι χρωστῆρι· κανὼν τύπον ὀρθὸν ὀπάζων·
καὶ λίθοσ εὐσχιδέων θηγαλέη καλάμων·

Transcription guidelines

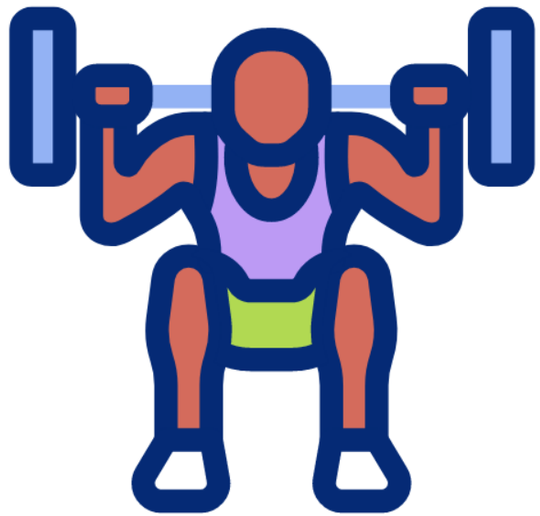
- Standardisation of sigma into “σ”
- Abbreviation are developped
- Significant symbols are transcribed
- Standardised punctuation (· and :)



ἀνάθημα ἀλκι
μένουσ κηπουροῦ:
ἀδέσποτον: ~

Model training #1

- pp. 143-195, 50 pages[1]
- *Segmenter* model
- *Recognizer* model



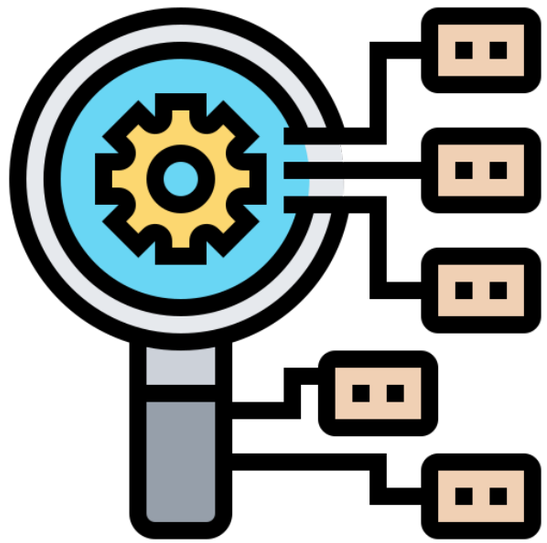
1. According to the latest pagination and that adopted by the Heidelberg Library, pages 177, 188 and 189 do not appear in the manuscript and are considered non-existent.

Segmenter

- Recognising zones of epigrams and scholia
- Recognising line types
- No metrics to calculate effectiveness

Recognizer

- Text recognition model
- Training is possible on eScriptorium
- More efficient to train directly with Kraken
- 3 models in NFC/NFD



Unicode Normalization

NFC = Normalization Form Canonical Composition

```
from unicodedata import normalize
lettre = "ὧ"
len(normalize("NFC", lettre))
```

```
1
```

ὧ = ὧ [U+1F66]

NFD = Normalization Form Canonical Decomposition

```
from unicodedata import normalize
lettre = "ὧ"
len(normalize("NFD", lettre))
```

```
3
```

ὧ = ω [U+03C9] + ̔ [U+0313] + ͂ [U+0342]

Preliminary results

Training data	Based on model	Unicode Normalization	Epochs	Accuracy score (%)
From scratch	None	NFC	47	87,36
From scratch	None	NFD	71	89,96
Fine-tuning	CREMMA Medieval	NFD	47	89,16

Model training #2

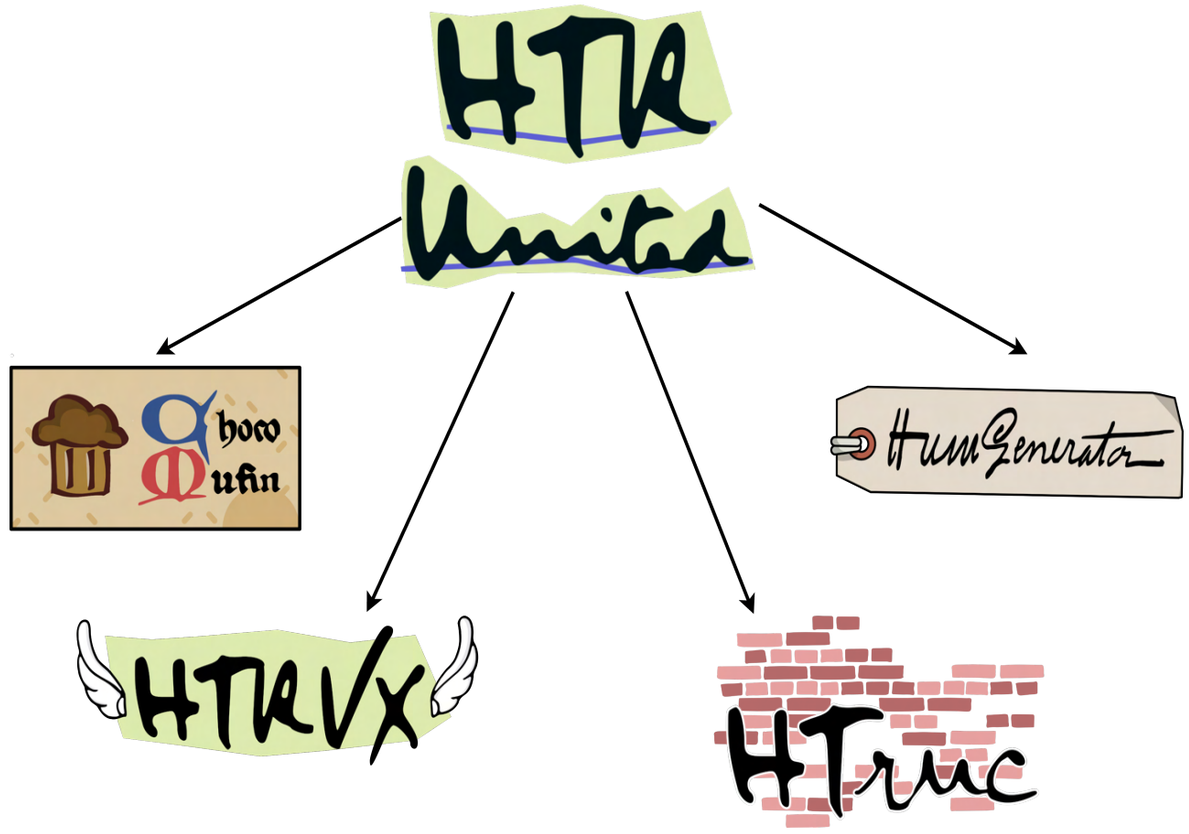
- pp. 143-215, 70 pages
- Improving *segmenter* model
- Improving *recognizer* models

Final results (v.1)

Training data	Based on model	Unicode Normalization	Epochs	Accuracy score (%)
From scratch	None	NFC	48	91,00
From scratch	None	NFD	40	90,85
Fine-tuning	CREMMA Medieval	NFD	48	91,05

FAIR data and standardised metadata

- Cataloguing ground-truth with HTR-United
- Cataloguing models with Zenodo

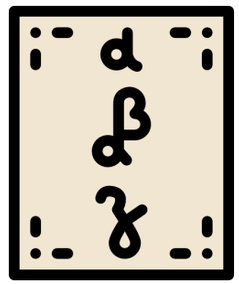


Conclusion

1. Results = Accuracy score of 90% et +

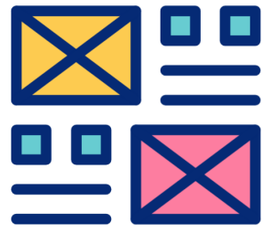
2. Amount of data = 50 p. is satisfying, but 70 p.+ is optimal

3. Training possible on models using different alphabets



Limitations

- *Segmenter* = only trained on *Codex Palatinus graecus* 23 layout



- *Recognizer* = only trained with scribe A handwritting, Xth Byzantine minuscule

Data repository

https://gitlab.huma-num.fr/ecrinum/anthologia/htr_cpgr23