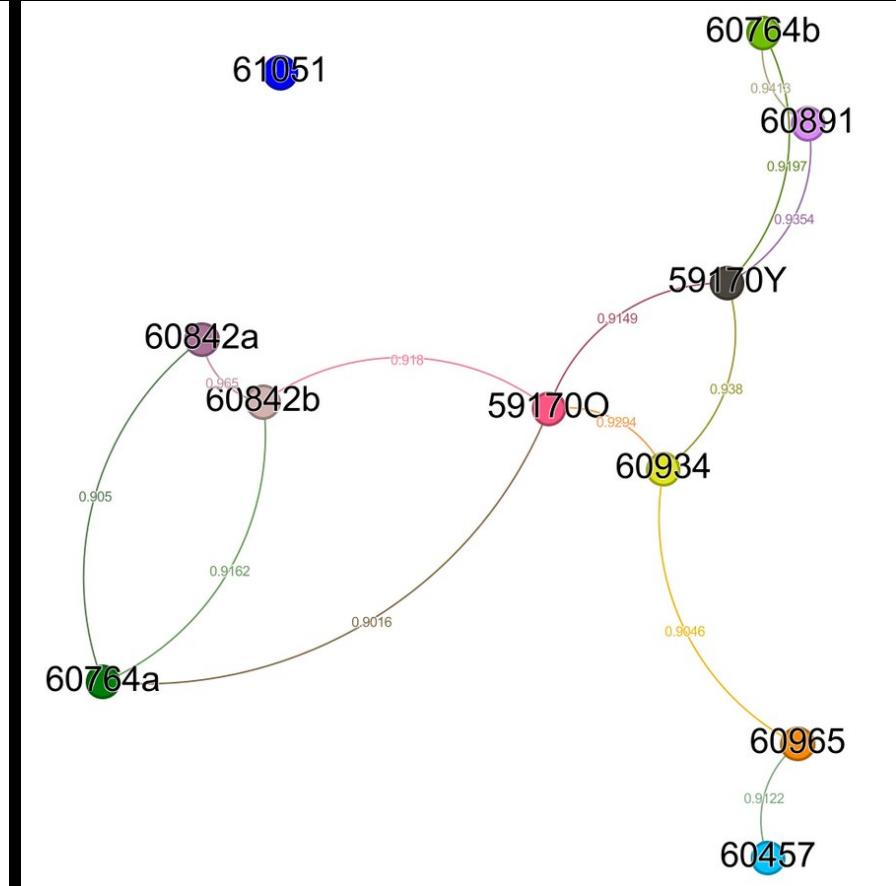
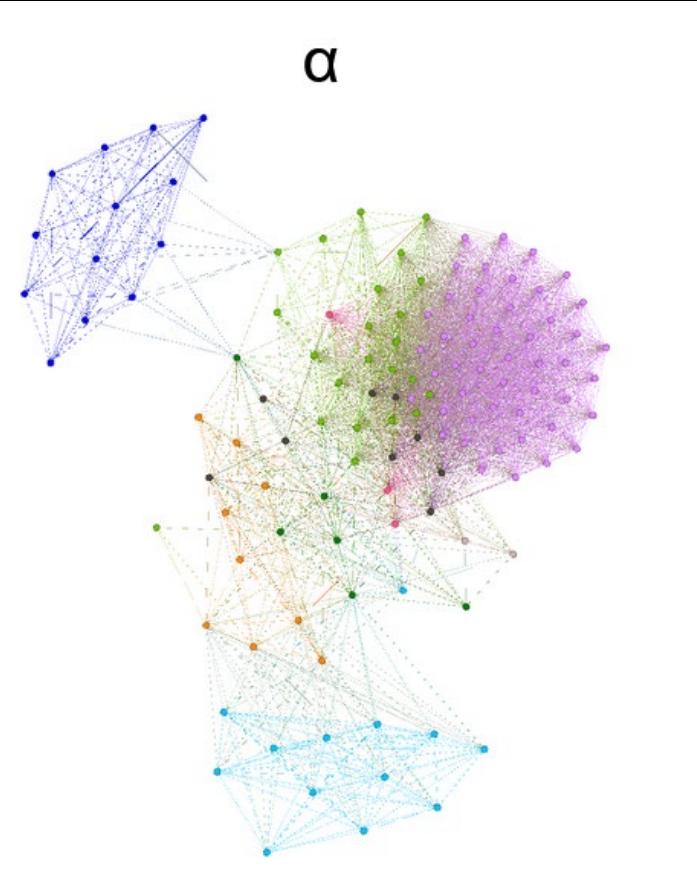


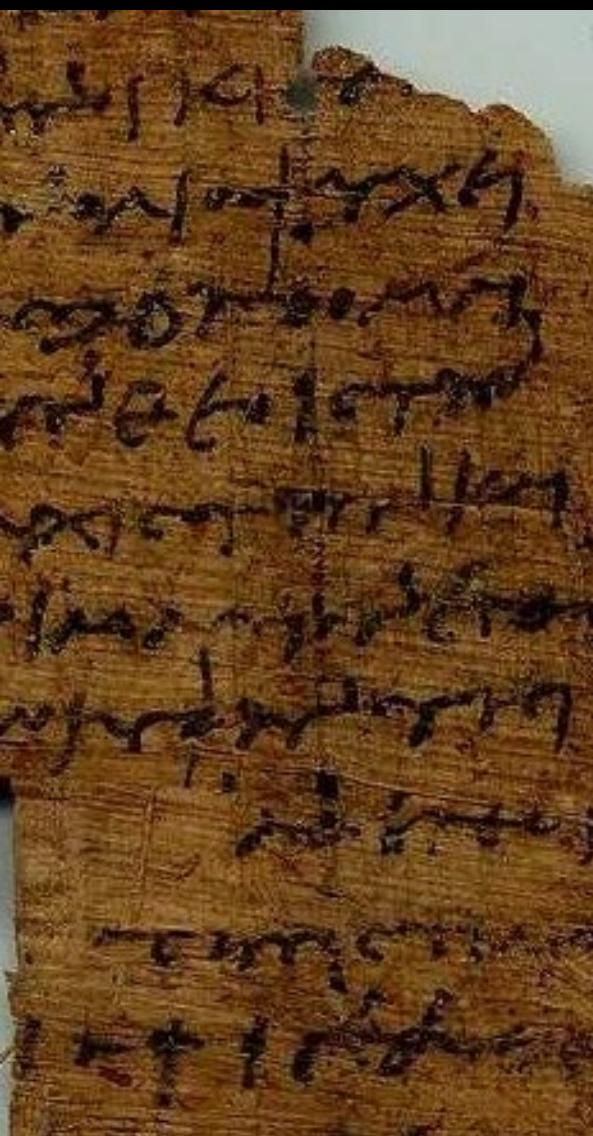
# The Digital Paleography of Greek Papyri: State of Research



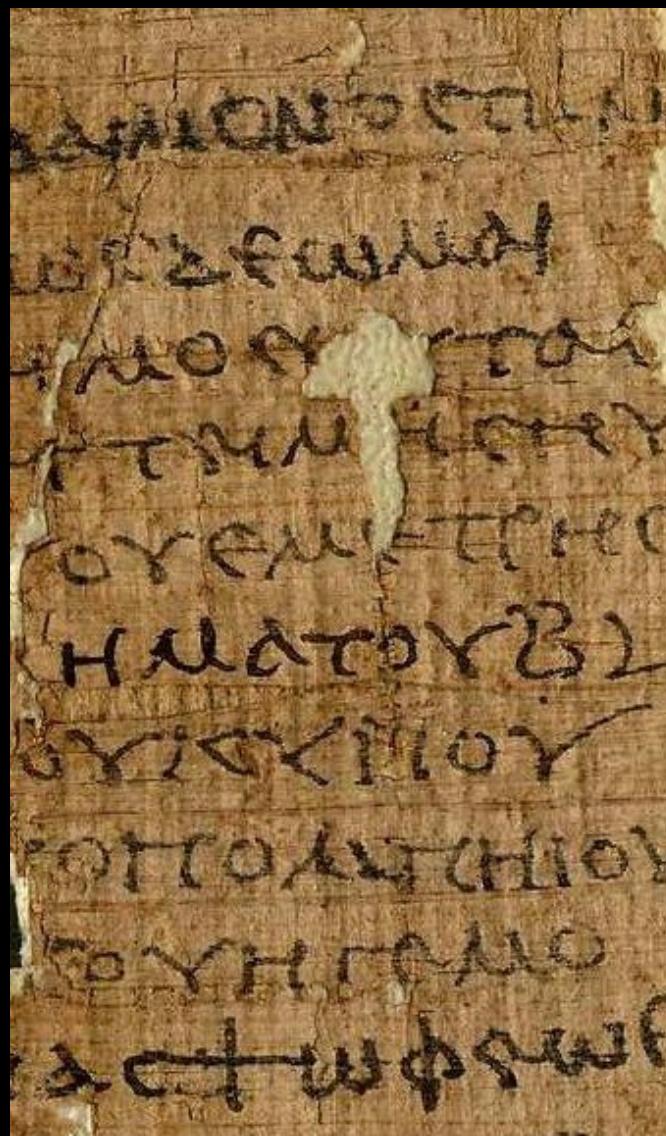
HTR and OCR from papyrus to codex  
Sunoikisis, 29 April 2024

# The papyrological material

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



Ptolemaic  
2<sup>nd</sup> – 1<sup>st</sup> c. BC



Roman  
2<sup>nd</sup> – 3<sup>rd</sup> c. AD

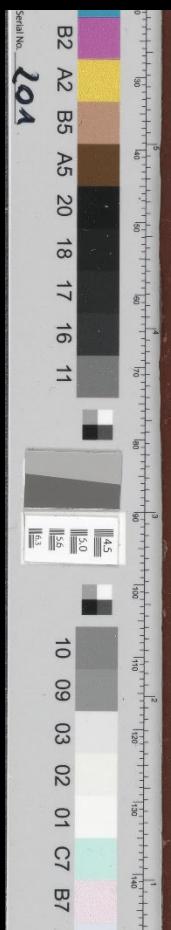


Byzantine  
5<sup>th</sup> – 6<sup>th</sup> c. AD

# The papyrological material

- A millennium 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, receipts, private accounts and school exercises on ostraca
- Various states of preservation (holes, breaks, faded or erased ink)

ΕΙΣΙΤΑΙ  
 ΤΟΥΣ ΤΟΥΣ  
 ΙΑΤΟΥΣ  
 ΕΥΧΟΣΑ Η  
 ΙΟΝΙΑ ΡΑ



P.83

ΥΚ ΣΑΛ  
 ΙΟΥΠΙΚΡΑ  
 ΚΟΥΗΝΟΙΟΕΕ  
 ΑΡΑΥΚΟΙ ΙΟΥΤ ΦΕΔ  
 ΛΙΠΙΟΣΣΥΛΕΙΕΙΘΕ  
 ΒΙ ΖΕΧΑ ΚΕΩΝΗ  
 ΑΙΡΕΤΕ ΣΑΛΟΝΑΡΗ  
 ΕΙ ΛΙΛΕ ΡΟΙΣΤΗΓΗ  
 ΚΕΡΙΟΣ ΣΙ ΟΗ ΝΗ  
 Ι ΚΟΤΟΛΑ ΣΑΡΙΟΥ  
 Ε Σ ΔΙΑΝΙΟ ΣΑ Ε Σ ΤΑ  
 Ε Σ Α Μ Π Ο Λ Ε Ι Ζ Ε Ν  
 Ε Σ Σ Σ Κ Ο Ν Ι Σ Τ Π Ο  
 Ε Κ Ε Ν Α Ρ Ο Τ Ε Υ Χ Ο Σ Α Η Ν Η  
 Ε Τ Σ Ο Ν Κ Α Ν Ο Ν Α Ρ Α  
 Ε Ο Σ Ν Ο Λ Χ Σ Ε Ν Σ Α  
 Ε Λ Ο Ο Σ Η

# 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)

Transkribus v1.10.0 (23\_01\_2020\_15:03). Loaded doc: TRAIN\_CITlab\_Zenon papyri\_duplicated, ID: 291227, Page 1, file: P.Cair.Zen.III.59377.jpg [Image Meta Info: (Resolution:72.0, w\*h: 573 \* 550) ] [ current line: w\*h: 405 \* 35 ]

Server Overview Layout Metadata Tools  
Logout imarthot@yahoo.com

Document... Find  
Document Manager User Manager  
Versions Jobs  
Recent Documents... User activity

Collections: papyri (34759, Editor) Col-ID

Documents HTR Model Data

1-20 / 20 Doc-ID

ID	Title	Pages	Uploader	Uploaded	Coll
2993...	TRAINING_VALIDATION_SET_gr...	2	tobias.hodel...	Mon Dec 23...	(pap)
2967...	TRAINING_VALIDATION_SET_gr...	2	tobias.hodel...	Tue Dec 17 ...	(pap)
2912...	TRAIN_CITlab_Zenon Papyri_du...	3	tobias.hodel...	Mon Dec 09...	(pap)
2912...	TEST_CITlab_Zenon Papyri_dup...	1	tobias.hodel...	Mon Dec 09...	(pap)
<b>291...</b>	<b>TRAIN_CITlab_Zenon papyri_d...</b>	<b>30</b>	<b>tobias.hodel...</b>	<b>Mon Dec 0...</b>	<b>(pap)</b>
2912...	TEST_CITlab_Zenon papyri_dup...	2	tobias.hodel...	Mon Dec 09...	(pap)
2912...	TRAIN_CITlab_Zenon archive_d...	30	tobias.hodel...	Mon Dec 09...	(pap)
2912...	TEST_CITlab_Zenon archive_du...	2	tobias.hodel...	Mon Dec 09...	(pap)
2304...	TRAINING_TESTSET_greek-pap...	2	tobias.hodel...	Sat Oct 19 1...	(pap)
2245...	TRAINING_TESTSET_Humarec...	1	tobias.hodel...	Tue Oct 08 ...	(pap)
2217...	WORK_ON_THIS-ddbdp-p1_t2i...	500	tobias.hodel...	Thu Oct 03 ...	(pap)
2104...	TRAINING_TESTSET_greek-pap...	1	tobias.hodel...	Sat Sep 14 0...	(pap)
1941...	ddbdp-p1_t2i_3	500	tobias.hodel...	Sat Aug 10 ...	(pap)
1780...	ddbdp-p1_t2i	239	tobias.hodel...	Sun Jun 30 ...	(pap)
1780...	ddbdp-p1_t2i_2	500	tobias.hodel...	Sun Jun 30 ...	(pap)
1579...	ddbdp P1 Img	500	tobias.hodel...	Thu May 02 ...	(pap)
1389...	la_digitexx	2	tobias.hodel...	Thu Mar 21 ...	(pap)
1313...	ddbdp p3 Images	564	tobias.hodel...	Thu Feb 28 ...	(pap)
18972	TEST_CITlab_Humarec Greek M1	1	guenter	Tue May 09 ...	(CIT)
18970	TRAIN_CITlab_Humarec Greek ...	35	guenter	Tue May 09 ...	(CIT)

100 Filter

The screenshot shows the Transkribus application interface. On the left, there's a navigation bar with links like 'Logout imarthot@yahoo.com', 'Document...', 'Find', 'Document Manager', 'User Manager', 'Versions', 'Jobs', 'Recent Documents...', and 'User activity'. Below that is a 'Collections' section showing 'papyri (34759, Editor)' and a 'Documents' section listing various files with their details. The main area displays a papyrus fragment with handwritten text in black ink on a light background. The text is partially obscured by a blue rectangular box. A red rectangular box highlights a specific word in the middle of the page. At the bottom, there's a list of numbered items, likely related to the transcription or analysis of the document.

44904

1-4 τε ἐργάσασθαι ἡμᾶς καὶ ἄλλην πλέω τοῦτου

1-5 χέρσον γεγενῆσθαι παρὰ τὸ μή σε χορηγεῖν ἡμῖν

1-6 τὰ κατὰ/ τὴν συνγραφήν —, νῦν οὖν ἀξιοῦμέν σε ποιεῖν

1-7 καθάπερ ἐπηγγείλου, ὥστε καὶ προδανίζειν(\*) ἡμῖν·

1-8 αὐτὸς γὰρ ἐπίστη ὅτι οὐθὲν γένημα γέγονεν

1-9 παρὰ τὸ τὴν ἀνυδρίαν γενέσθαι καὶ εἰς τὸν ἐπερ-

1-10 χόμενον χρόνον πεπόνηκεν. νυνὶ οὖν ἀπόφανον

1-11 ἡμῖν ὃ ἂν σοι δοκῇ, ἵνα μὴ καταφθεῖ/ρωμεσθα

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, September 30–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

The screenshot shows a web browser window with the URL [scrollprize.org](https://scrollprize.org) in the address bar. The page title is "Vesuvius Challenge". The main headline reads: "Resurrect an ancient library from the ashes of a volcano." Below it, in red text, is "Win Prizes. Make History." A descriptive paragraph explains: "The Vesuvius Challenge is a machine learning and computer vision competition that in 2023 cracked the riddle of the Herculaneum Papyri & awarded over \$1,000,000 in prizes." At the bottom, another text block states: "2024's challenge is to go from reading a few passages to entire scrolls." The background of the page features a dramatic illustration of Mount Vesuvius erupting.

scrollprize.org

Vesuvius Challenge

# Resurrect an ancient library from the ashes of a volcano.

## Win Prizes. Make History.

The Vesuvius Challenge is a machine learning and computer vision competition that in 2023 cracked the riddle of the Herculaneum Papyri & awarded over \$1,000,000 in prizes.

2024's challenge is to go from reading a few passages to entire scrolls.



Horizon 2020  
European Union funding  
for Research & Innovation



GREEK SCHOOLS



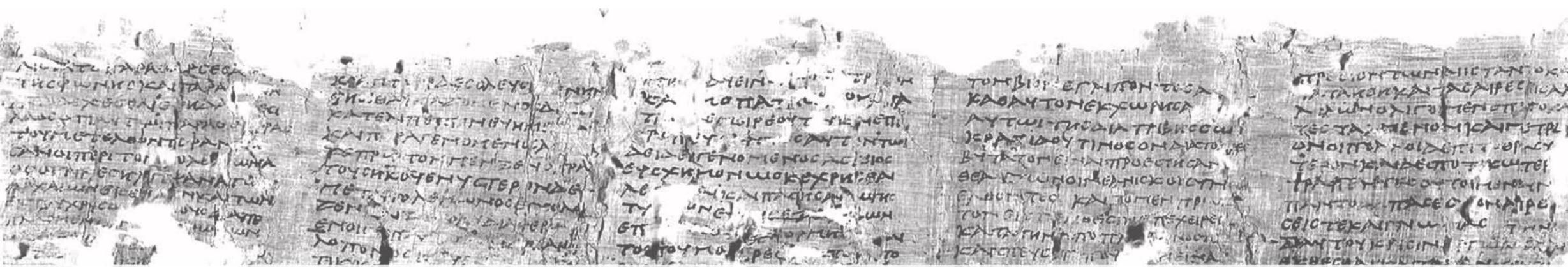
UNIVERSITÀ DI PISA



Consiglio Nazionale  
delle Ricerche



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

solutions



TRISMEGISTOS

Welcome to

# TRISMEGISTOS

An interdisciplinary portal of the ancient world

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](#).



LDAB



Languages / scripts



Collections



Archives



People



Places

# The peculiarities of (digital) papyrology

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

Browse: DDbDP HGV APIS DCLP Authors TM Number

or Search: Data Bibliography

## Search

New Search

Search

within  chars        
    

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

Text  Metadata  Translations

Series  
or--- All values ---  Vol. 

Collection

--- All values ---  ID # 

Author

--- All values --- 

Work

--- All values --- Records per page  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

DE GRUYTER

*Nicola Reggiani*

# DIGITAL PAPYROLOGY I

METHODS, TOOLS AND TRENDS

DE  
G

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.

The screenshot shows the eScriptorium website interface. At the top, there's a navigation bar with a logo, 'eScriptorium', 'Home', and 'Contact' links. Below the header, the title 'eScriptorium' is prominently displayed, followed by a subtitle: 'eScriptorium: A Digital Text Production Pipeline for Print and Handwritten Texts using machine learning techniques.' The page is divided into four main sections, each featuring a circular image and a title:

- Automatic Transcription:** Shows a circular image of a handwritten document with Arabic text. The text below the image reads: 'Apply OCR/HTR to images of printed and handwritten documents using shared open models.'
- Manual transcription:** Shows a circular image of a printed document with Arabic text. The text below the image reads: 'Make use of an ergonomic user interface leveraging modern browser technology to edit segmentations and transcriptions.'
- Train Models:** Shows a circular image of a handwritten document with a green outline highlighting specific text segments. The text below the image reads: 'Create new models or finetune existing ones to improve automatic recognition.'
- Import/Export:** Shows a circular image of a printed document with a green outline highlighting specific text segments. The text below the image reads: 'Import and export models and texts transcriptions in a variety of formats. Access data through a full REST API.'

On the right side of the page, there's a vertical sidebar with several small icons and labels, likely for navigation or additional features. The overall layout is clean and professional, designed to showcase the capabilities of the eScriptorium platform.

# 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/>

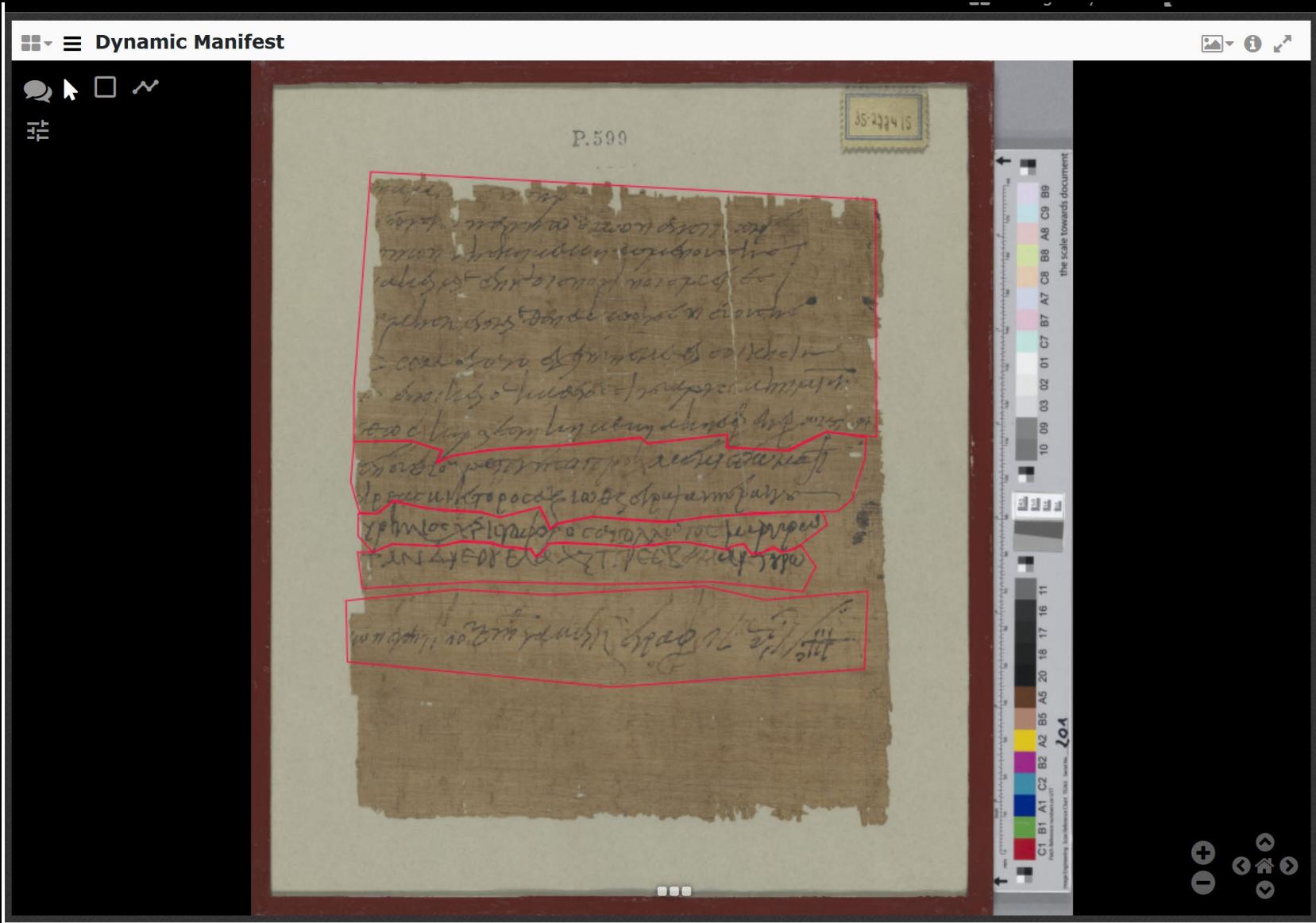
# D-scribes

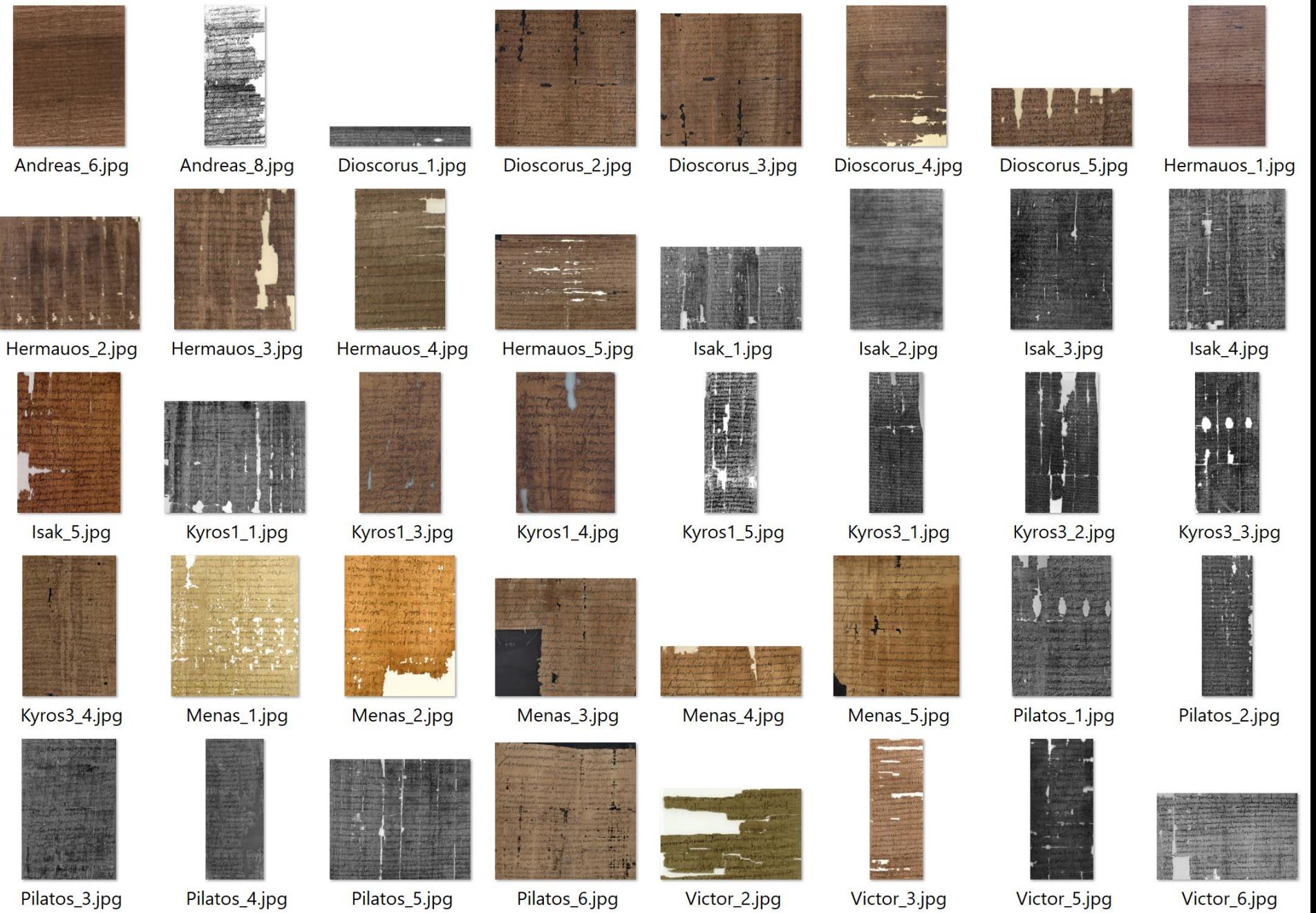
Reuniting fragments, identifying scribes and  
characterizing scripts:

The Digital Palaeography  
of Greek and Coptic Papyri

SNSF Ambizione project 2018-2023

# Writers in Dioscorus archive





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/>

Presentation

Case studies

Events

Mailing list

Resources

Publications



> Case studies > Dioscorus



GRK-Papyri

GRK-Papyri extensions

KaiRacters Dataset

## 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](#).

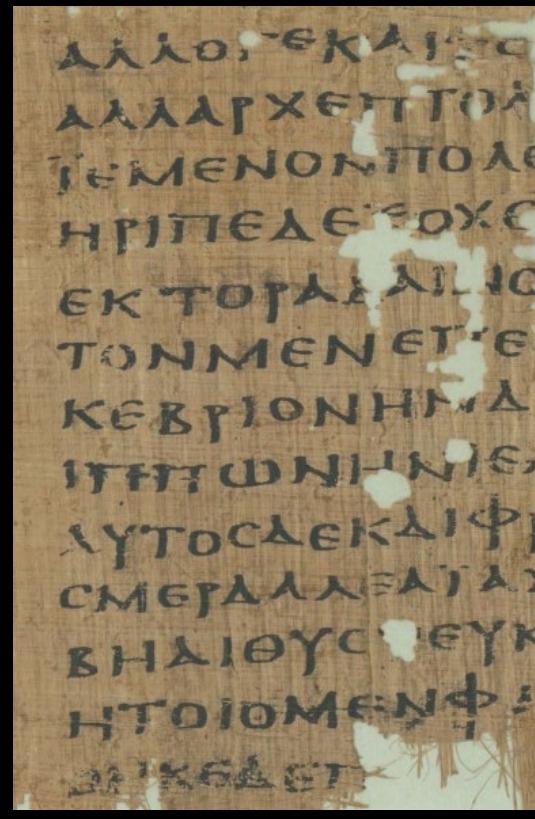
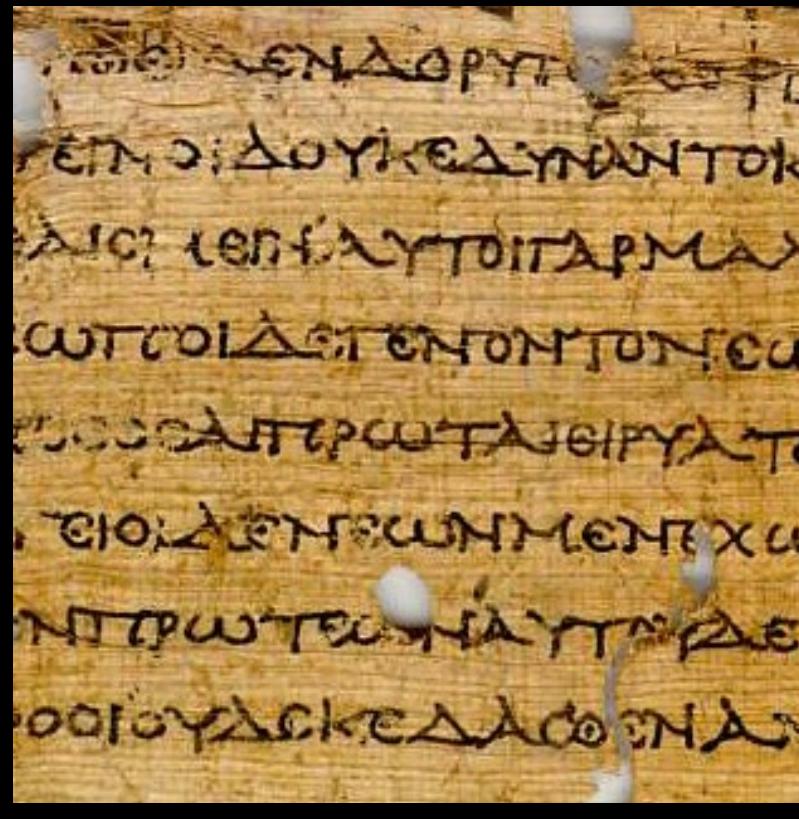
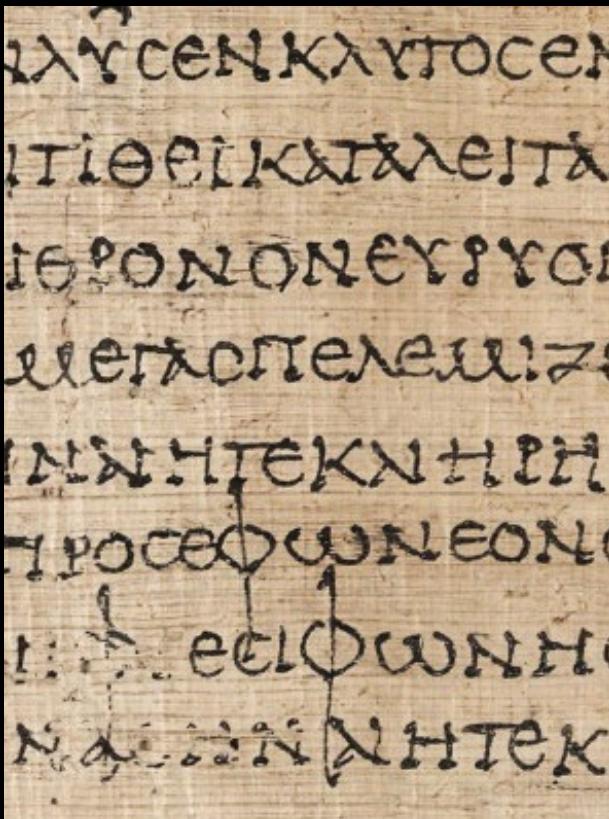
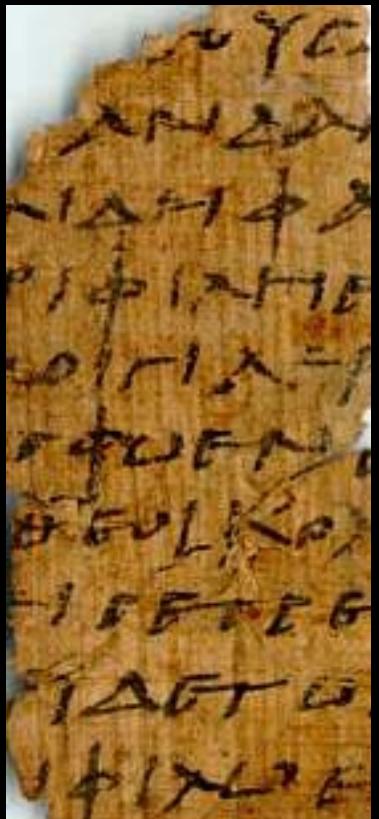
>[GRK-Papyri\\_97](#) is a selection of GRK-Papyri\_120 to count only 19 writers (97 images) is a study currently submitted to ICDAR 2023.

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

# Script characterization

Homer's *Iliad*:

1500 papyri across a millenium



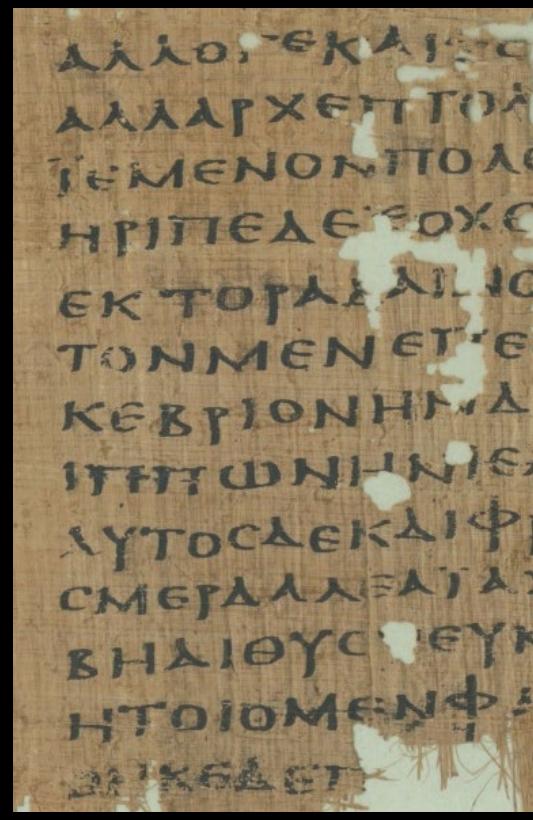
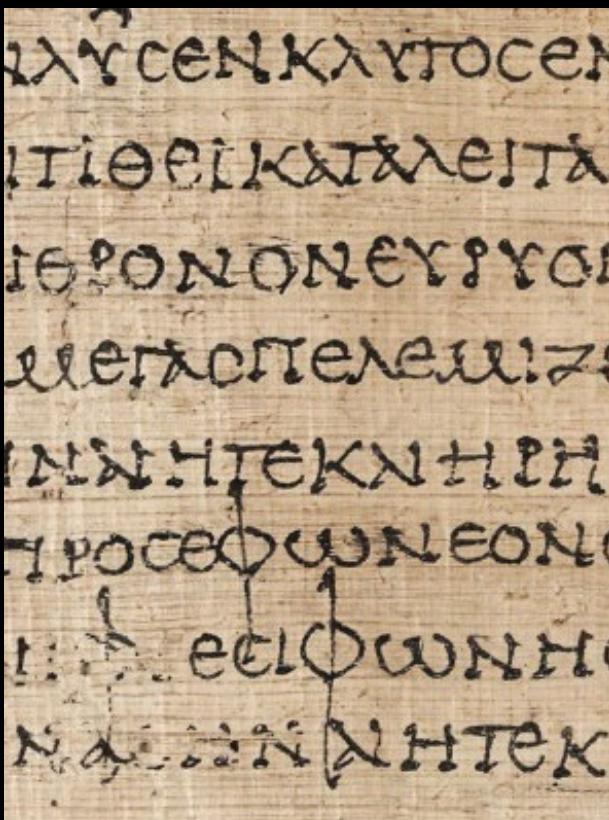
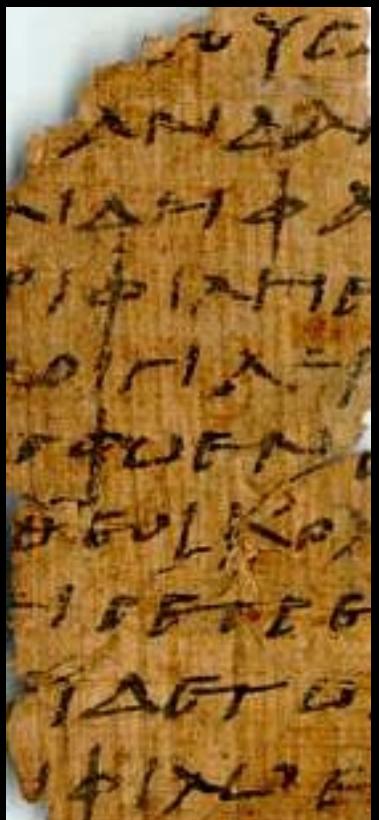
# Script characterization

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

**FIND**

Resources  
PP

Enter Search Here

60242 [3s] - P\_Laur\_IV\_128r

P\_Laur\_IV\_128r

Word List  
Structures  
Paleography  
Syntax  
Translation  
Chāyā

**EDIT**

Link segment  
Save polygon  
Replace polygon  
Delete segment  
Auto Link  
Link by number  
139 Seg. Number

Off  
Number segs.

View  
Numbers

**60242\_P\_Laur\_IV\_128r [epidoc]**

**Iliad.5.159** (ΕΝΩ') ΥΙΑΣ ΠΡΙΑΜΟΙΟ ΔΥΩ ΛΑΒΕ Δ(ΑΡΔΑΝΙΔΑΟ)  
**Iliad.5.160** (ΕΙΝ) ΕΝΙ ΔΙΦΡΩ ΕΟΝΤΑC ΕΧΕΜΜ(ΟΝΑ ΤΕ ΧΡΟΜΙΟΝ ΤΕ.)  
**Iliad.5.161** (ΩC ΔE) ΛΕΩΝ EN BOYCI ΘΟ(P)ΩΝ E(Ξ AΥΧΕΝΑ ΑΞΗ)  
**Iliad.5.162** (ΠΟΡΤΙ)ΟC ΗΕ ΒΟΟC ΞΥΛΟΧΟΝ ΚΑΤΑ (ΒΟΣΚΟΜΕΝΑΩΝ ,)  
**Iliad.5.163** (ΩC T)ΟYC ΑΜΦΟΤΕΡΟΥC ΕΞ (ΙΠΠΩΝ ΤΥΔΕΟC ΥΙΟC)  
**Iliad.5.164** (ΒΗΣΕ) ΚΑΚΩC ΑΕΚΟΝΤΑC (,) E(ΠΕΙΤΑ ΔΕ ΤΕΥΧΕ' ΕΣΥΛΑ ·)  
**Iliad.5.165** (ΙΠΠΟΥ)C Δ(') OIC ΕΤΑΡΟΙC ΔΙΔΟ(Y ΜΕΤΑ NHAC ΕΛΑΥΝΕΙN.)  
**Iliad.5.166** (ΤΟΝ Δ') ΙΔΕΝ ΑΙΝΕI(AC ΑΛΑΠΑΖΟΝΤΑ CTIXAC ΑΝΔΡΩΝ ,)  
**Iliad.5.167** (ΒΗ Δ' IM)EN (AN) TE M(ΑΧΗΝ KAI ANA ΚΛΟΝΟΝ ΕΓΧΕΙΑΩΝ)  
**Iliad.5.168** (ΠΑΝΔΑΡΟΝ) ANTI(ΘΕΟΝ ΔΙΖΗΜΕΝΟC EI ΠΟΥ ΕΦΕΥΡΟΙ ·)

**Isabelle**  
 E: DScriptorEditor  
 V: Users  
 A: Isabelle (Work in progress)

**FIND**

Enter Search Here

60220 [4.1] - P\_08440\_R\_001

**EDIT**  
 View  
 Edit mode

**VIEW**  
 Download  
 Properties Segments

**LAYOUT**  
 Change Off  
 Panel Layout Sync Scroll

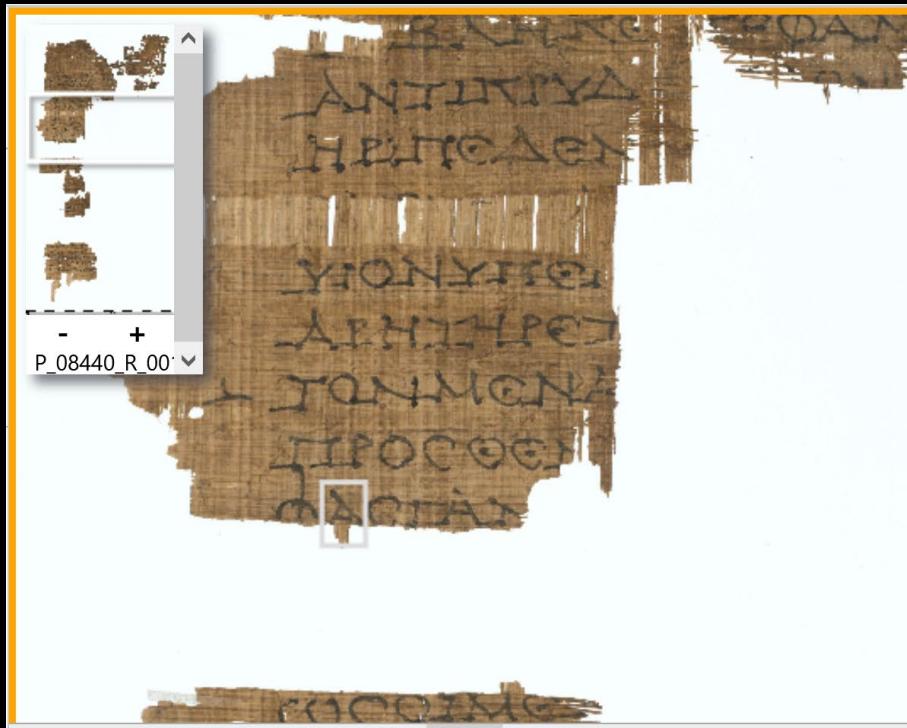


**60220\_P\_08440\_R\_001 [epidoc]**

**Iliad.5.69** ΠΗΔΑΙΟΝ Δ(') ΑΡ(') ΕΠ(ΕΦΝ)Ε ΜΕΓΗΣ Α(ΝΤΗΝΟΡΟΣ ΥΙΟΝ)  
**Iliad.5.70** ΟΣ ΡΑ ΝΟΘΟC MEN (ΕΗΝ ΠΥΚΑ Δ' ΕΤΡΕΦΕ ΔΙΑ ΘΕΑΝΟ)  
**Iliad.5.71** ΙΑ ΦΙΛΟΙC(I) ΤΕΚΕCCI Χ(ΑΡΙΖΟΜΕΝΗ ΠΟΣΕΙ Ω)  
**Iliad.5.72** ΤΟΝ MEN ΦΥΛΕΙΔΗ(C ΔΟΥΡΙ ΚΛΥΤΟΣ ΕΓΓΥΘΕ)N Ε(ΛΘΩΝ)  
**Iliad.5.73** (ΒΕ)ΒΛΗΚΕ(I) ΚΕΦΑΛ(HC) ΚΑ(TA INION ΟΞΕΙ ΔΟΥΡΙ)  
**Iliad.5.74** ΑΝΤΙΚΡΥ Δ(' ΑΝ' ΟΔΟΝΤΑς ΥΠΟ ΓΛΩΣΣΑΝ ΤΑΜΕ ΧΑΛΚΟC)  
**Iliad.5.75** ΗΡΙΠΕ Δ(') EN (ΚΟΝΙH ΨΥΧΡΟΝ Δ' ΕΛΕ ΧΑΛΚΟΝ ΟΔΟΥCIN)  
**Iliad.5.76** (ΕΥΡΥΠΥΛΟC Δ' ΕΥΑΙΜΟΝΙΔΗC ΥΨΗΝΟΡΑ ΔΙΟΝ)  
**Iliad.5.77** ΥΙΟΝ ΥΠΕΡ(ΘΥΜΟΥ ΔΟΛΟΠΙΟΝΟC ΟC ΡΑ ΣΚΑΜΑΝΔΡΟY)  
**Iliad.5.78** ΑΡΗΤΗΡ ΕΤ(ΕΤΥΚΤΟ ΘΕΟC Δ' ΖC ΤΙΕΤΟ ΔΗΜΩ)  
**Iliad.5.79** ΤΟΝ MEN Α(P' ΕΥΡΥΠΥΛΟC ΕΥΑΙΜΟΝΟC ΑΓΛΑΟC ΥΙΟC)  
**Iliad.5.80** ΠΡΟΣΘΕΝ (ΕΘΕΝ ΦΕΥΓΟΝΤΑ ΜΕΤΑΔΡΟΜΑΔΗN ΕΛΑC' ΖΜΟN)  
**Iliad.5.81** ΦΑΓΑΝ(O ΑΙΕΑC ΑΠΟ Δ' ΕΞΕC ΧΕΙΡΑ ΒΑΡΕΙAN)  
**Iliad.5.84** ΖC ΟI ΜE(N ΠΟΝΕΟΝΤO ΚΑΤΑ ΚΡΑΤΕΡΗN ΥCΜΙΝΗN)  
**Iliad.5.85** ΤΥΔΕΙΔ(HN Δ' ΖΥK ΑN ΓΝΟΙHC ΖΤΕΡΟΙC ΜΕΤΕΙH)  
**Iliad.5.86** ΗE (M)ΕΤA (ΤΡΩΕCCIN ΖΜΙΛΕΟI Η MET' ΑΧΑIΟIC)  
**Iliad.5.87** (ΘΥNE Γ)ΑΡ Α(M ΖΕΙDΙON ΖΤΑΜΩ ΖΛΗΘΟΝΤI ΕΟΙΚΩC)  
**Iliad.5.88** (ΧΕΙM)ΑΡΡΩ (ΟC T' ΖΚΑ ΖΕΩN ΕΚΕΔΑSCCE ΓΕΦΥΡΑC)

**Paleography for 60220\_P\_001**

A	B	C	D	E	F	G	H	I	J	K	L	M	N	Ξ	Ο	Π	Ρ	Σ	Τ



### 60220\_P\_08440\_R\_001 [epidoc]

- Iliad.5.69** ΠΗΔΑΙΟΝ Δ(') ΑΡ(') ΕΠ(ΕΦΝ)Ε ΜΕΓΗΣ Α(ΝΤΗΝΟΡΟΣ ΥΙΟΝ)
- Iliad.5.70** ΟC ΡΑ ΝΟΘΟC ΜΕN (ΕHN ΠΥΚΑ Δ' ΕΤΡΕΦΕ ΔΙΑ ΘΕΑΝΩ)
- Iliad.5.71** ICA ΦΙΛΟΙC(I) TEKECCI X(APIZOMENH ΠΟΣΕΙ Ω)
- Iliad.5.72** TON MEN ΦΥΛΕΙΔΗ(C ΔΟΥΡΙ ΚΛΥΤΟC ΕΓΓΥΘΕ)N E(ΛΘΩΝ)
- Iliad.5.73** (ΒΕ)ΒΛΗΚΕ(I) ΚΕΦΑΛ(HC) KA(TA INION ΟΞΕΙ ΔΟΥΡΙ)
- Iliad.5.74** ΑΝΤΙΚΡΥ Δ(') ΑΝ' ΟΔΟΝΤΑC ΥΠΟ ΓΛΩCCAN ΤΑΜΕ ΧΑΛΚΟC
- Iliad.5.75** ΗΡΙΠΕ Δ(') EN (ΚΟΝΙH ΨΥΧΡΟN Δ' ΕΛΕ ΧΑΛΚΟN ΟΔΟΥCIN)
- Iliad.5.76** (ΕΥΡΥΠΥΛΟC Δ' ΕΥΑΙΜΟΝΙΔΗC ΥΨΗΝΟΡΑ ΔΙΟN)
- Iliad.5.77** YION ΥΠΕР(ΘΥΜΟY ΔΟΛΟΠΙΟΝΟC ΟC ΡA ΣΚΑΜΑΝΔΡΟY)
- Iliad.5.78** ΑΡΗΤΗP ET(ΕΤΥΚΤΟ ΘΕΟC Δ' ΩC TIETO ΔΗMΩ)
- Iliad.5.79** TON MEN A(P' ΕΥΡΥΠΥΛΟC ΕΥΑΙΜΟΝΟC ΑΓΛΑΟC ΥΙΟC)
- Iliad.5.80** ΠΡΟΣΘΕΝ (ΕΘΕN ΦΕΥΓΟΝΤΑ ΜΕΤΑΔΡΟΜΑΔΗN ΕΛΑC' ΩMΟN)
- Iliad.5.81** ΦΑCΓΑN(Ω ΑΙΞΑC ΑΠO Δ' ΕΞECE ΧΕIΡΑ ΒΑΡΕΙΑN)

### Paleography for 60220\_P\_08440\_R\_001

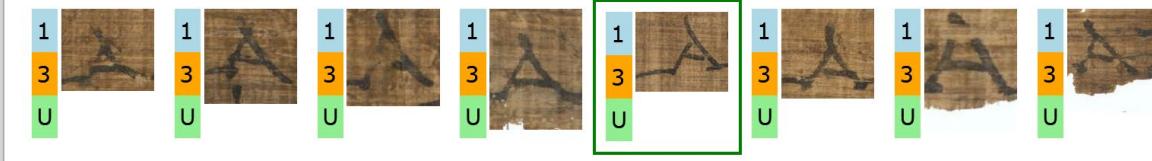
Sort: base

Tagging Orthographic Group A

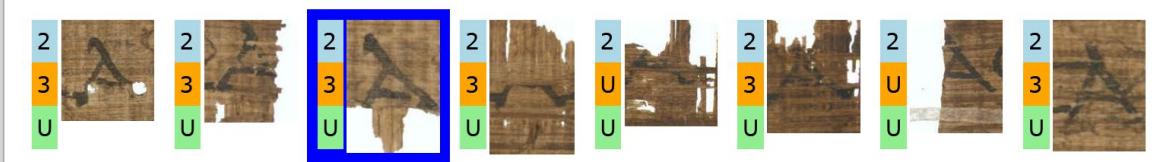
← →

Previous Next

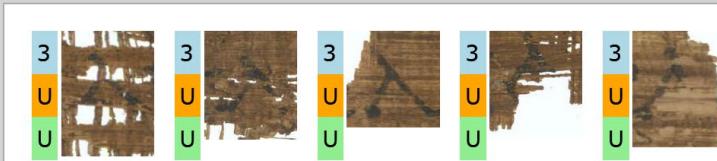
BT-1



BT-2



BT-3



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

## PalEx2

TM:  
60214 [6]

Select your image:  
P\_11645\_R\_001.jpg

Viewer   Explorer

Categories   Tags

Hide All ? A B Г Δ E Z H

BaseType  
 bt1  
 bt2  
 bt3  
 bt4  
 bt5

FootMarkType  
 ft1  
 ft2  
 ft3  
 ft4



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

## PalEx2

TM:  
60214 [6]

Select your image:  
P\_11645\_R\_001.jpg

← →

Click [HERE](#)for more info  
on the papyri!

Categories Tags

Hide	<input type="checkbox"/> BaseType
All	<input checked="" type="checkbox"/> bt1
.	<input type="checkbox"/> bt2
?	<input type="checkbox"/> bt3
<input checked="" type="checkbox"/> A	<input type="checkbox"/> bt4
B	<input type="checkbox"/> bt5
Γ	<input type="checkbox"/>
<input type="checkbox"/> Δ	FootMarkType
<input checked="" type="checkbox"/> E	<input type="checkbox"/> ft1
Z	<input type="checkbox"/> ft2
<input type="checkbox"/> H	<input type="checkbox"/> ft3
	<input type="checkbox"/> ft4

Viewer Explorer

Categories

BaseType

bt1

A

E

The image shows a digital interface for examining ancient manuscript fragments. On the left, there's a sidebar with a dropdown for 'TM' (Text Manuscript) set to '60214 [6]', a dropdown for 'Select your image' showing 'P\_11645\_R\_001.jpg', and a section for 'Categories' and 'Tags'. The 'Categories' section includes checkboxes for 'BaseType' (with 'bt1' checked), 'FootMarkType' (with 'E' checked), and other letters like 'A', 'B', 'Γ', 'Δ', 'Z', 'H'. The 'Tags' section also lists these categories. At the top, there are tabs for 'Viewer' and 'Explorer', with 'Viewer' currently selected. The main area displays a grid of handwritten cursive letters from a papyrus fragment. The letters are arranged in two rows. The first row contains a single letter 'A' and several 'E's. The second row contains several 'E's. The letters are written in dark ink on a light-colored, textured background. The interface is designed to facilitate the study and comparison of ancient script samples.



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](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, epsilon and mus
- AI: SimSiam network

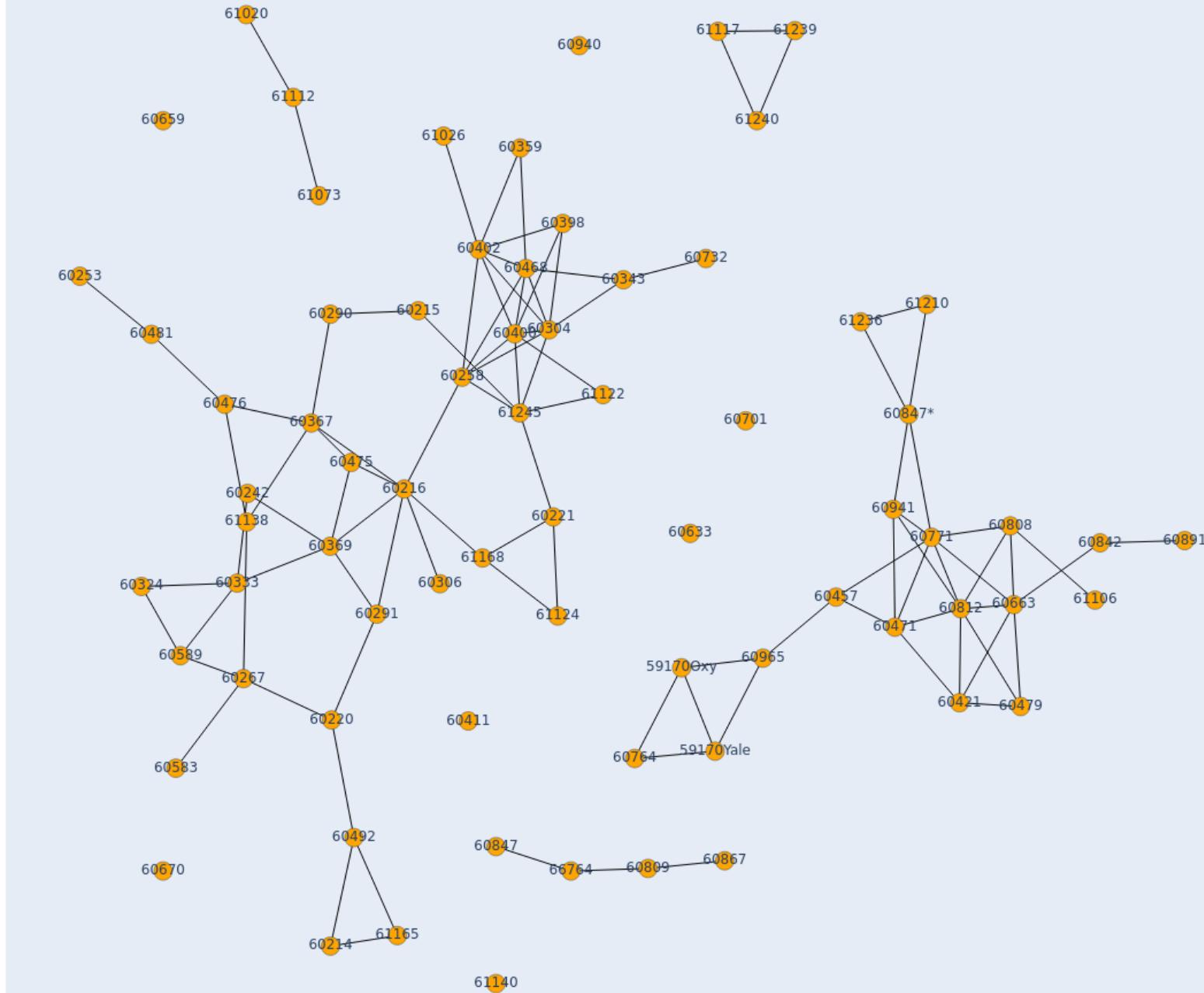


# Similarity estimation results

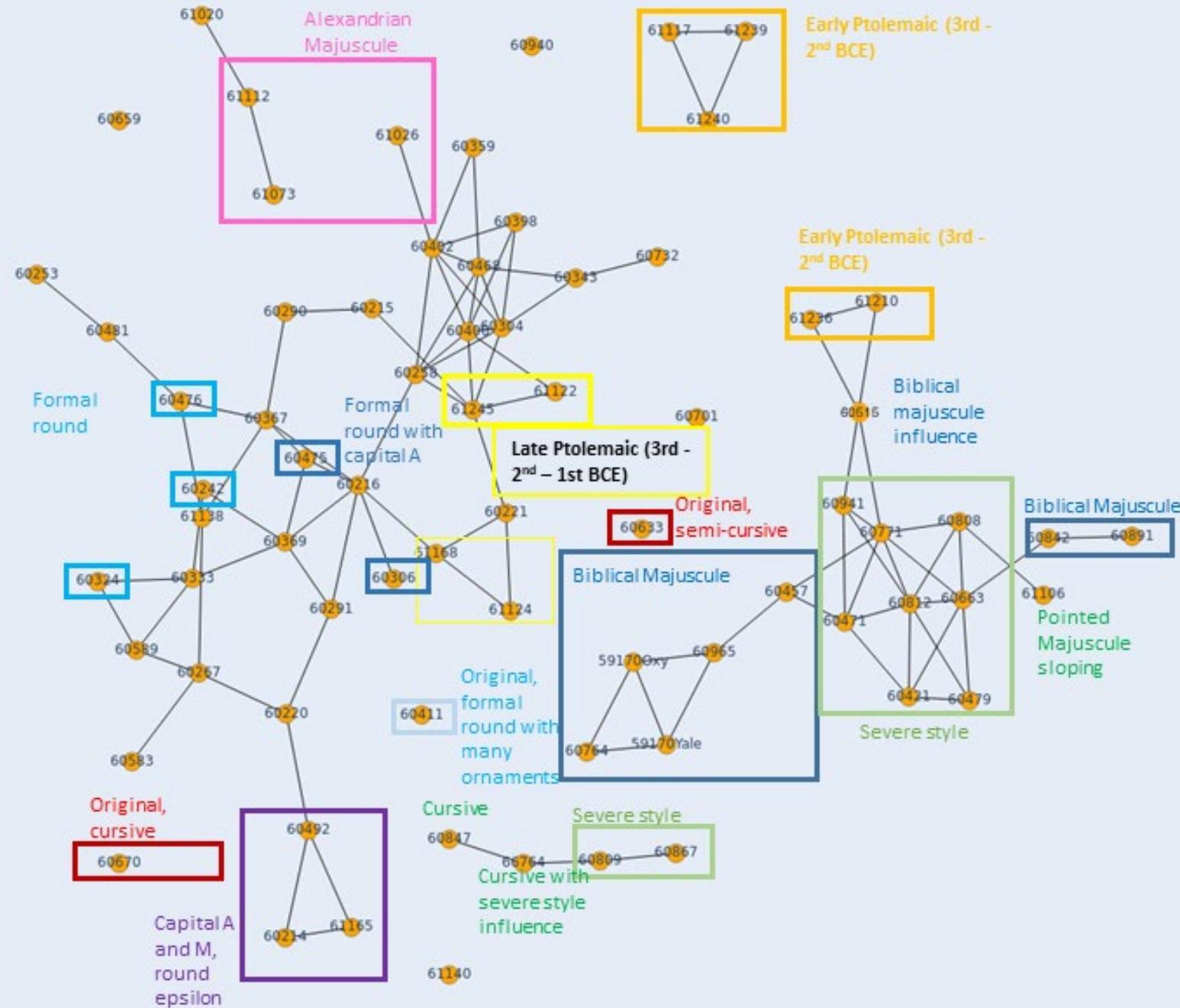
# Query image



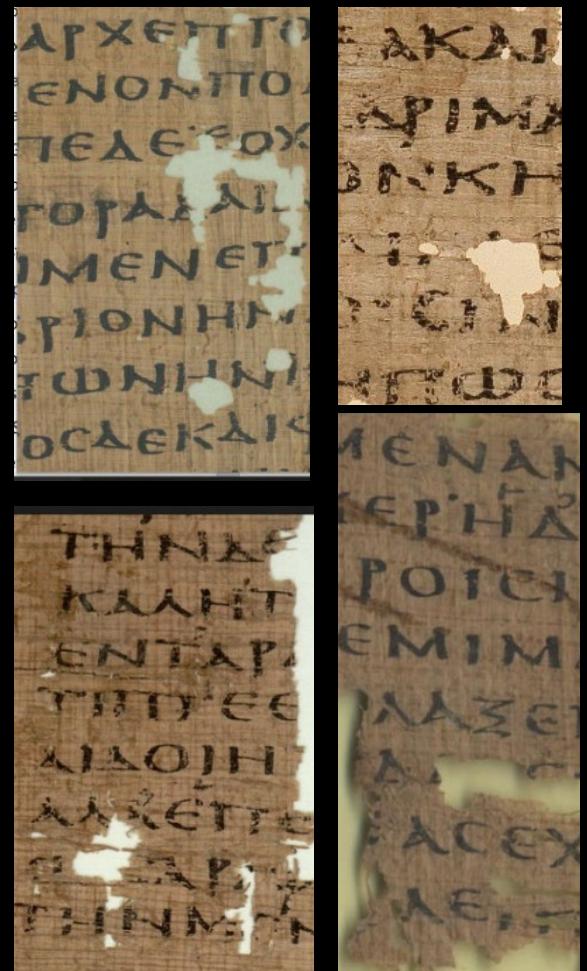
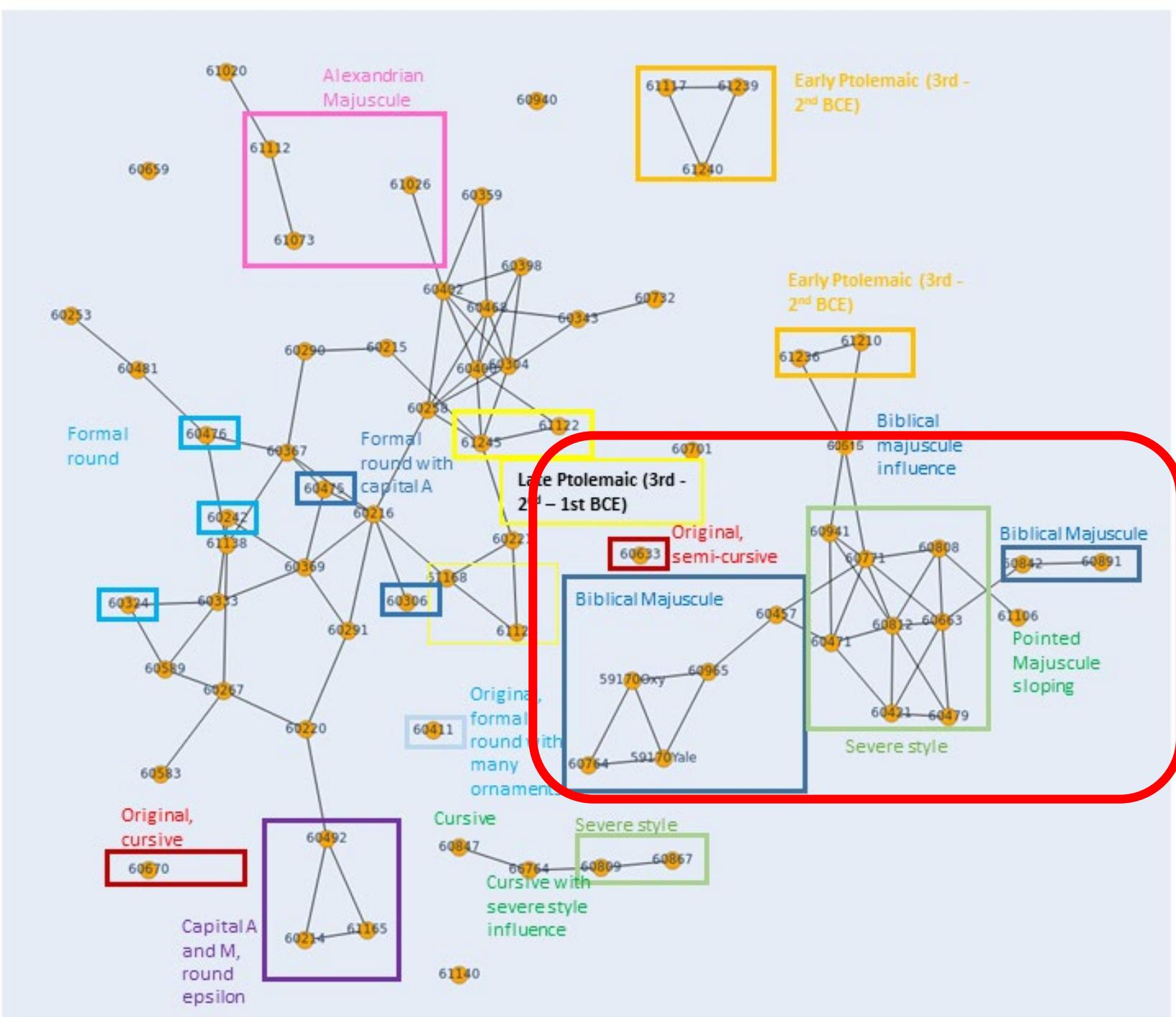
Similarity estimation results between papyri  
(combination of alpha, epsilon and mu scores)



# Stylistic interpretation



# Stylistic interpretation



# Similarity estimation results

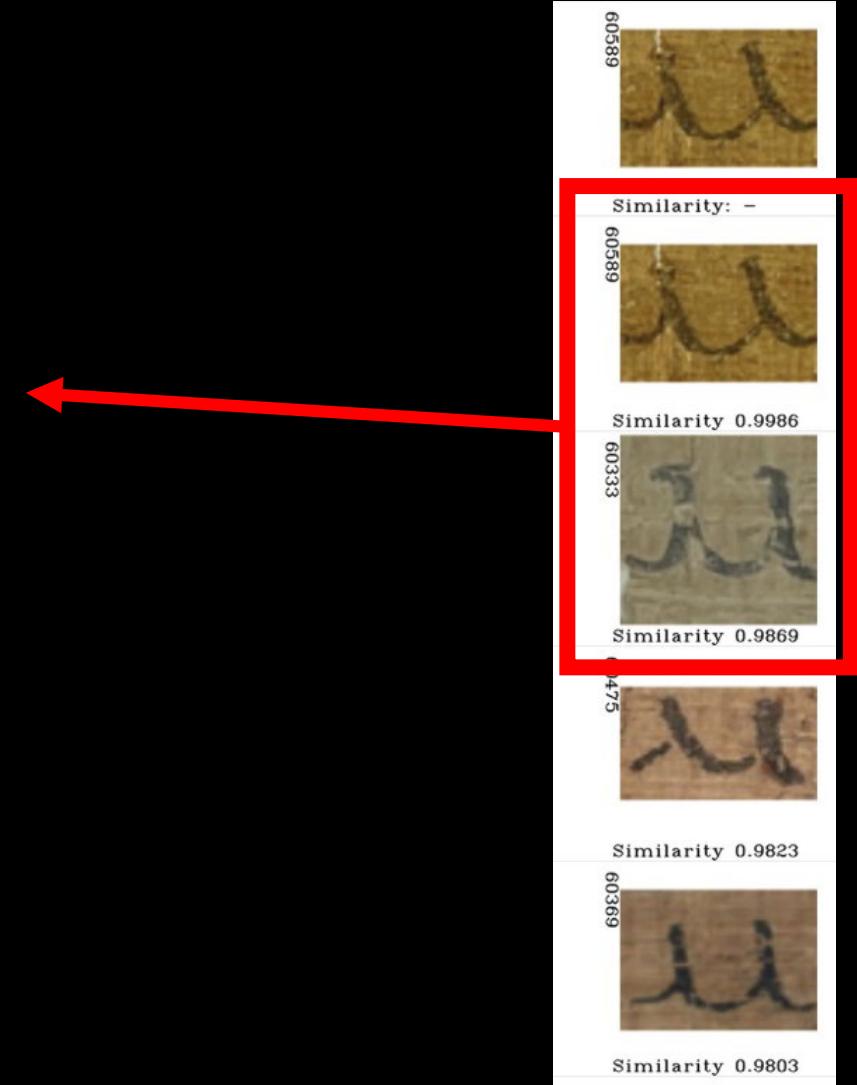
# Query image



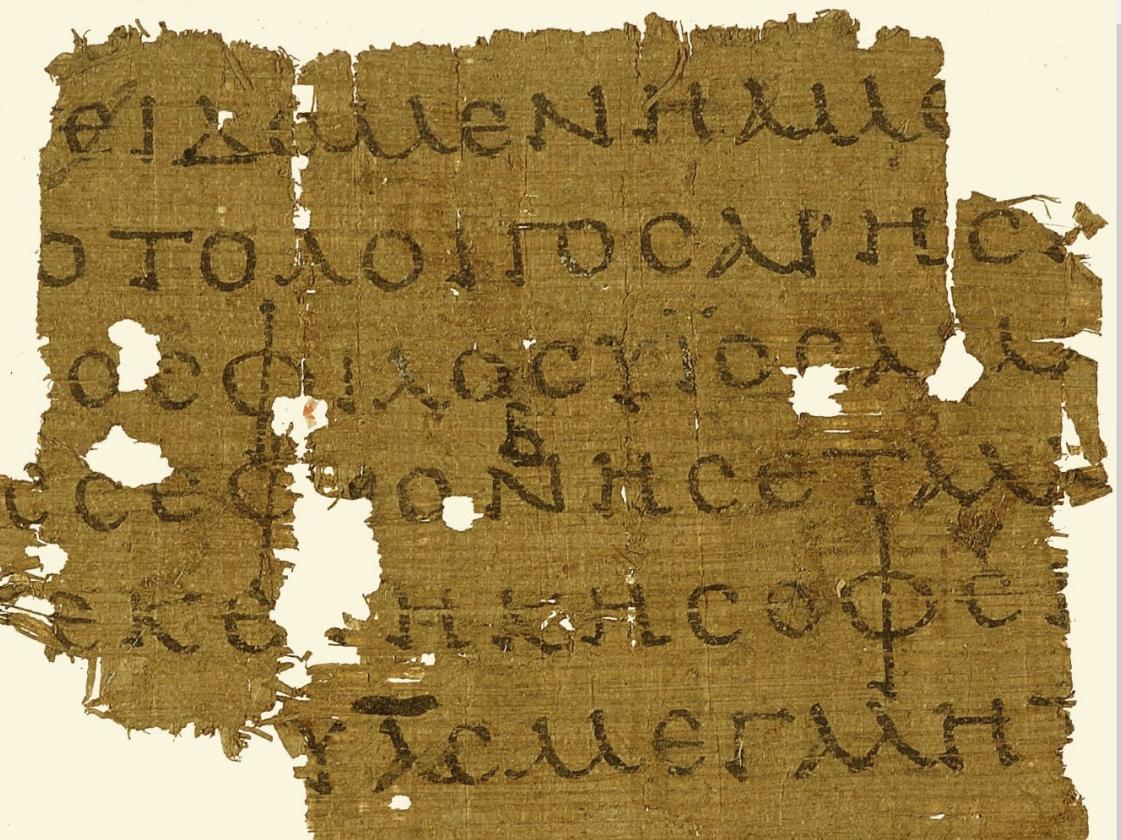
# Similarity estimation results

Query image		60589	60589	60589	60589	60589	60589
#1	591700xy	Similarity: -					
#2	591700xy	Similarity 0.9993	Similarity 0.9974	Similarity 0.9984	Similarity 0.9956	Similarity 0.9966	Similarity 0.9986
#3	59170Yale	Similarity 0.9879	Similarity 0.9945	Similarity 0.9948	Similarity 0.8613	Similarity 0.8836	Similarity 0.9869
#4	60965	Similarity 0.9515	Similarity 0.9360	Similarity 0.7645	Similarity 0.7638	Similarity 0.8090	Similarity 0.9823
	60764	Similarity 0.9261	Similarity 0.8283	Similarity 0.7635	Similarity 0.7414	Similarity 0.7940	Similarity 0.9803

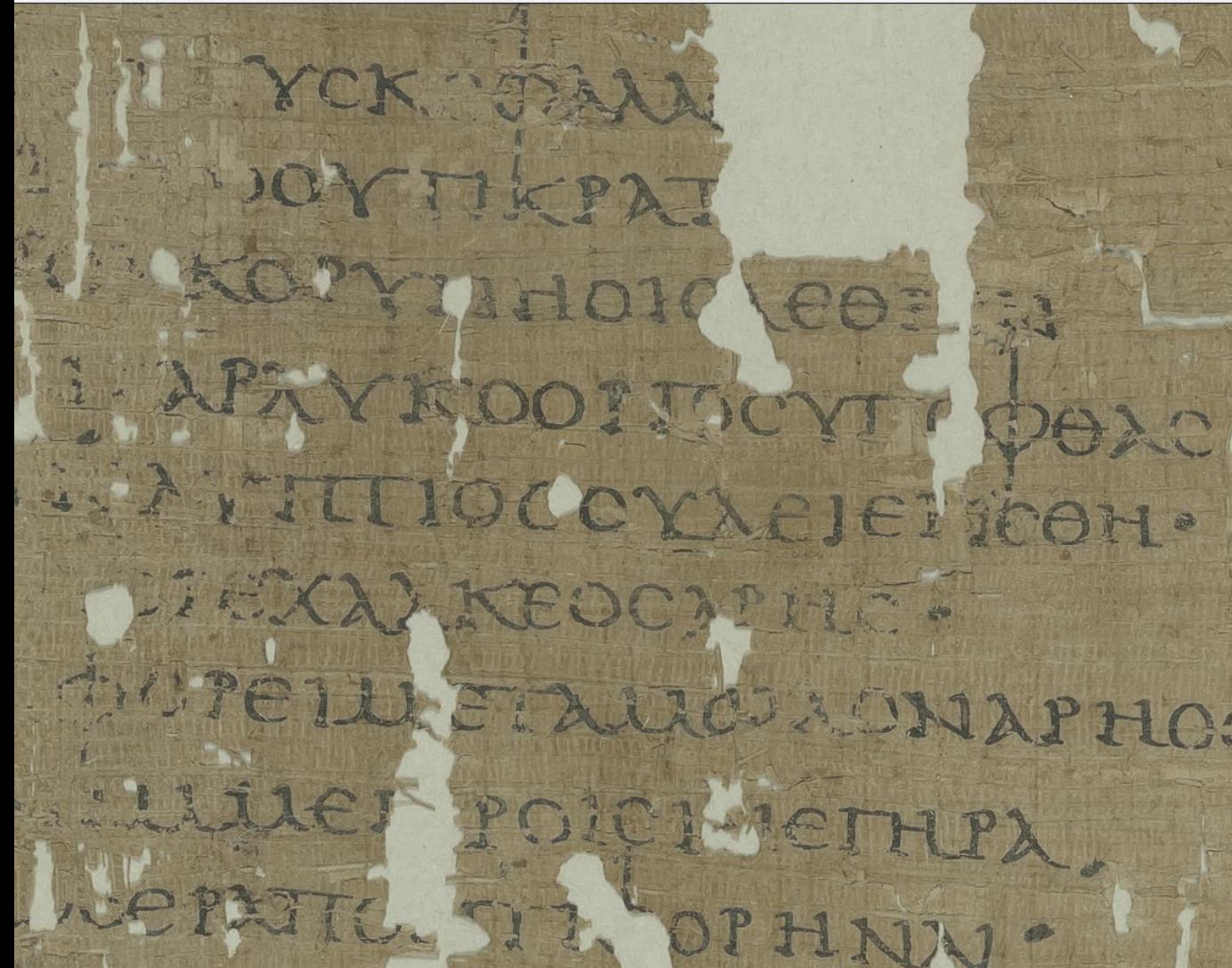
# Similarity estimation results



60589



60333



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

TM:  
60589 [6]

Select your image:  
BNU\_Pgr1876\_r.jpg

Categories Tags

Hide All

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

BaseType

- bt1
- bt2
- bt3
- bt4
- bt5

FootMarkType

- ft1
- ft2
- ft3
- ft4
- ft5
- ft6
- ft7
- ft8
- ft9

Viewer Explorer

Patlex

60589

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

IM:

60589 [6]

Select your image:

BNU\_Pgr1876\_r.jpg

Click [HERE](#)for more info on the papyri!

Categories

Tags

Hide All

.

?

A

B

Γ

Δ

E

Z

H

Θ

I

K

Λ

M

N

Ξ

Ο

Π

Ρ

C

Τ

Υ

Φ

X

Ψ

Ω

□BaseType

bt1bt2bt3bt4bt5

□FootMarkType

ft1ft2ft3ft4ft5ft6ft7ft8ft9

Viewer

Explorer

BaseType

Categories

bt1

.

?

A



B



Γ

Δ

Ε



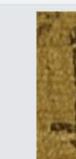
Ζ

Η



Θ

Ι



A



B



Г



Д



Е



Z



H



Θ

60589



60333

K



Λ



Μ



Ν



Ξ



Ο



Π

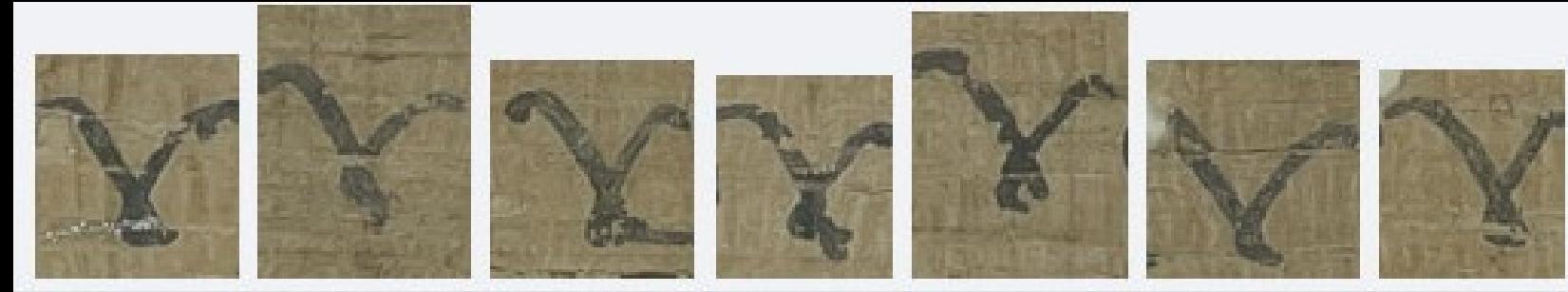


60589



60333

60333



60589



Home /

## ICDAR2023 Competition on Detection and Recognition of Greek Letters on Papyri



The 17th International Conference on Document Analysis and Recognition

PROGRAM ▾    IMPORTANT DATES    SUBMISSIONS ▾    ORGANIZING COMMITTEE    SPONSORS    AWARDS    CONTACT US

SAN JOSE, CALIFORNIA, USA 2023

August 21-26, 2023 – San José, California, USA



Click [HERE](#) for 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

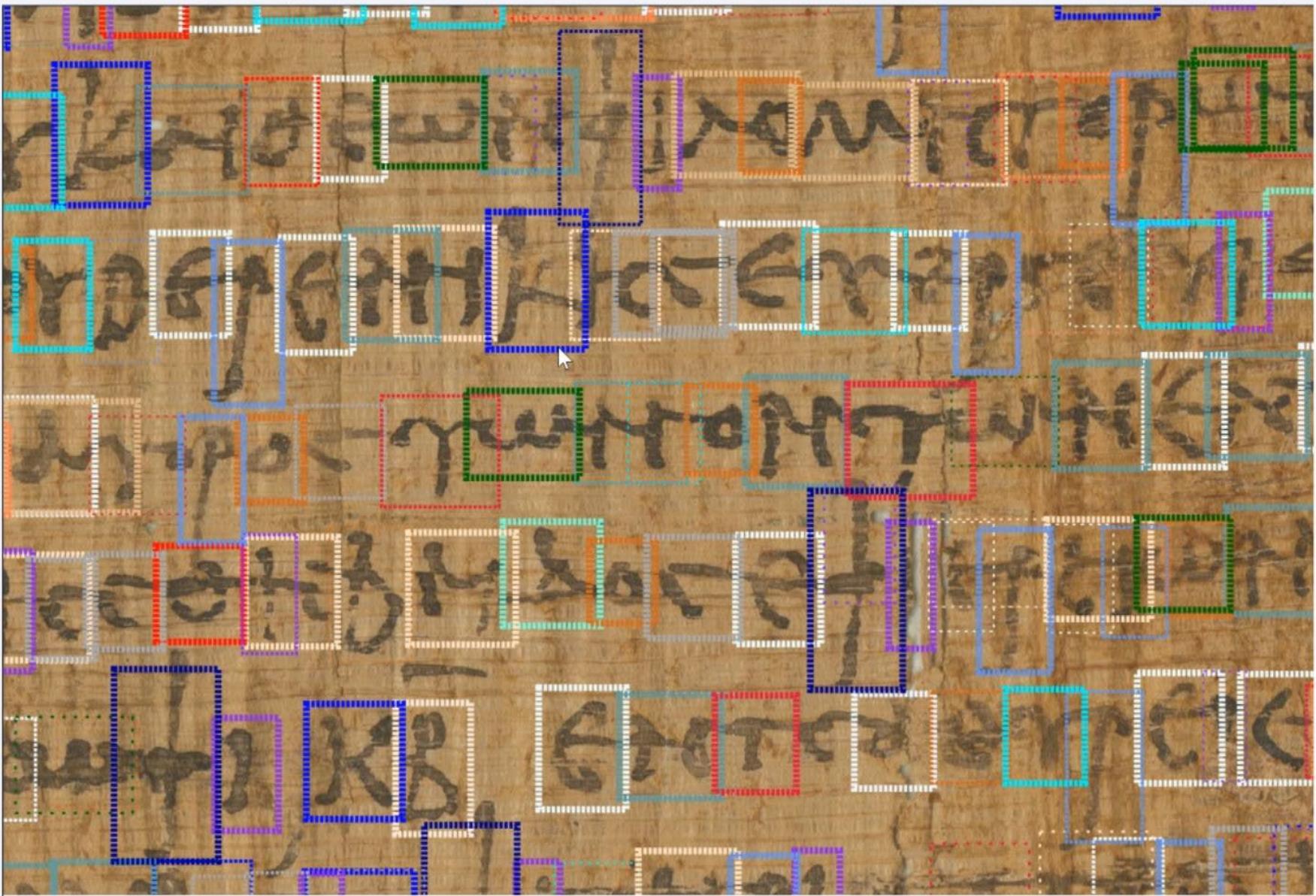
Τ

Y

Φ

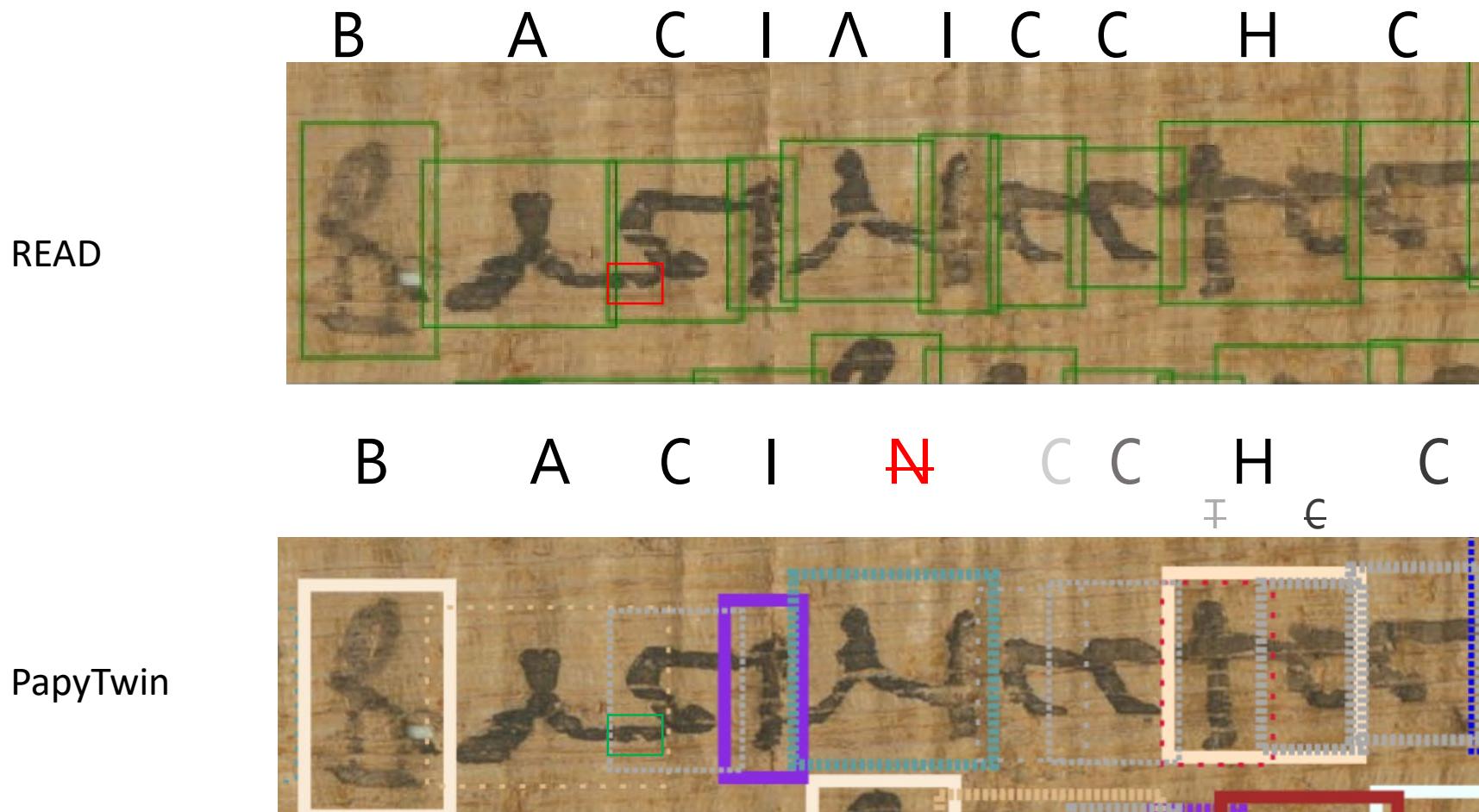
X

Η



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

# Human tool (READ) vs AI Model (PapyTwin)



Categories
<input type="checkbox"/> Hide All
<input type="checkbox"/> A
<input type="checkbox"/> B
<input type="checkbox"/> Γ
<input type="checkbox"/> Δ
<input type="checkbox"/> E
Z
<input type="checkbox"/> H
<input type="checkbox"/> Θ
<input type="checkbox"/> K
<input type="checkbox"/> I
<input type="checkbox"/> Λ
<input type="checkbox"/> M
<input type="checkbox"/> N
Ξ
<input type="checkbox"/> Ο
<input type="checkbox"/> Π
<input type="checkbox"/> P
Σ
<input type="checkbox"/> T
<input type="checkbox"/> Y
<input type="checkbox"/> Φ
<input type="checkbox"/> X
<input type="checkbox"/> Ψ
<input type="checkbox"/> C
<input type="checkbox"/> Ω

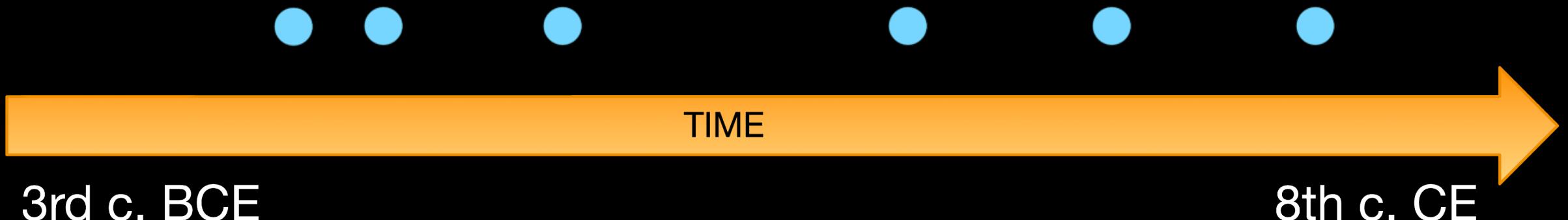
# EGRAPSA: Retracing the evolutions of handwritings in Greco-Roman Egypt thanks to digital palaeography

June 2023-May 2028  
SNSF, University of Basel

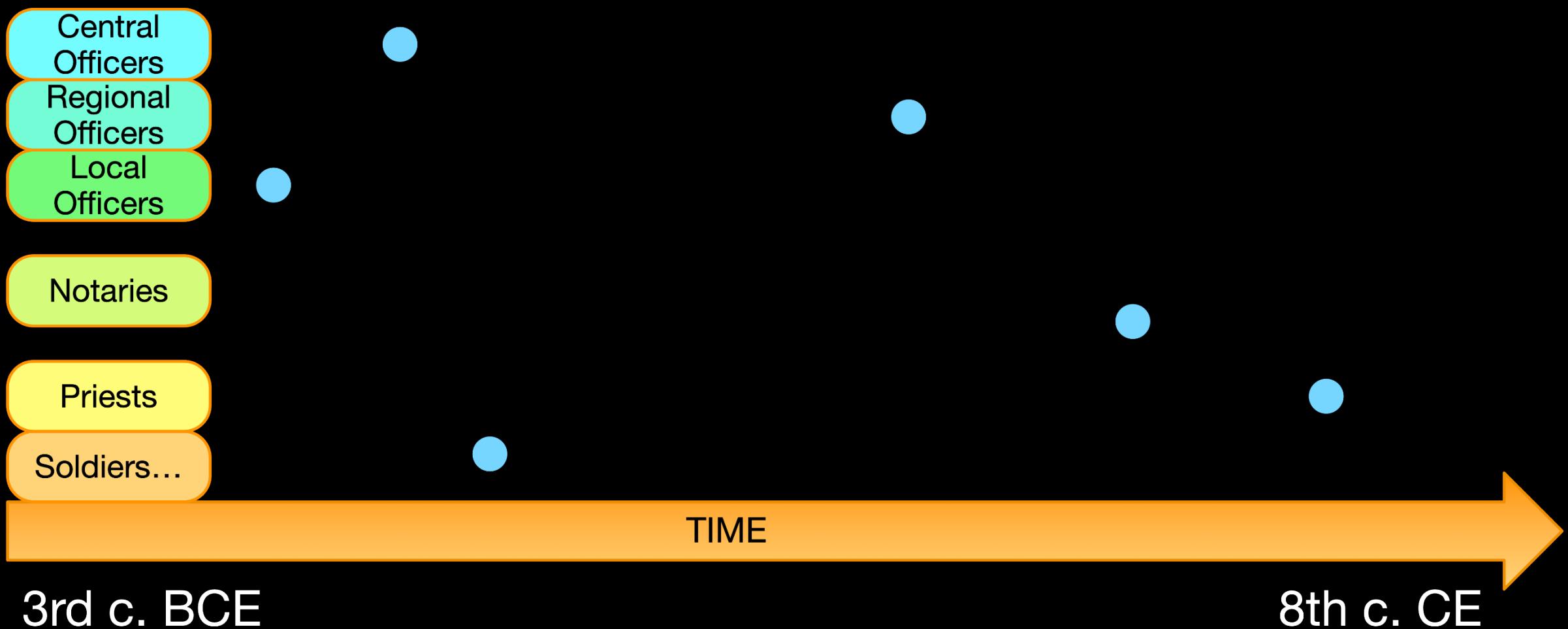
Project goal of EGRAPSA (“I have written”)

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

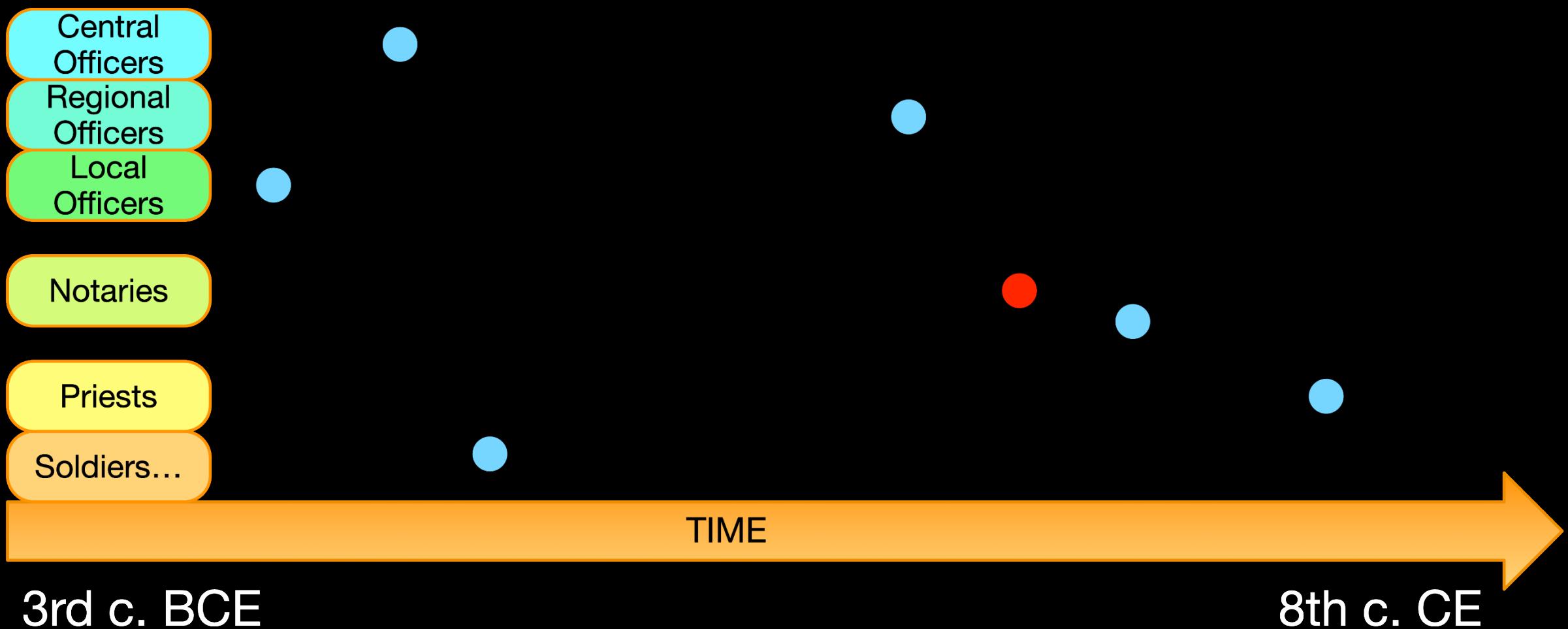
# Diachronic perspective: sound anchor points from dated texts



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

The screenshot shows the header of the d-scribes.org website. On the left, there's a logo for the University of Basel and the Department of Ancient Civilizations. In the center, the website's name "d-scribes.org" is displayed. On the right, there's a circular emblem. Below the header, a navigation bar has tabs for "Presentation", "Case studies", "Events", "Mailing list" (which is highlighted with a red underline), "Hierax enhancer", and "Publications". Underneath the navigation bar, a breadcrumb trail shows a house icon followed by "Mailing list". The main content area contains text about the D-Scribes List and a link to subscribe.

Ancient History | D-Scribes

Presentation Case studies Events Mailing list Hierax enhancer Publications

Home > Mailing list

D-Scribes List is a non-commercial mailing list aiming to connect people interested in Digital Humanities, especially in the fields of Digital Palaeography (Computerized Classification, letter/sign shape comparison), Writer Identification and Ancient Document Analysis (layout, alignment, shape of the fragments, annotating tools).

To subscribe, please follow [this link](#)

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”

Οτιός τονεσθεογι. Νομαλύ. ρύφου. τρόζου. γαλιν. λέδιο::  
Κείται ωρού φρού. Λευτός. οξουδηνάλη. Στούδην. Κεράμην. Βούδηνηπ::  
Τουλιάκην. ποταμού άπογονότην. ο ίη. υφουνγόνον. άρνη. Ταρέ<sup>η</sup>  
Αγαστήν. δένηνονια κράζο. ορίουν δικάλης Νεναστούνηρ.  
Γελέ καταναλώσι. αντηγαζότος. ~. 16th c. AD

παντοτε εντελεῖτον. προσεχή Καὶ μηδέ εἰς τὸν ποτηνόν ηγέρεται  
εῖτι. μετενθειαί είσαι μηχανούται τὸν ποτηνόν. ή ποτε είτι  
πρεσ. τοῦ ποτετηί μηχανούται Καὶ μηδέ επιχανούται. ή.  
προσει. τοῦ ποτετηί μηχανούται Καὶ μηδέ επιχανούται. ή.  
Δική φραστη. ποτε είτι μηχανούται Καὶ μηδέ επιχανούται. ή.  
14th c. AD

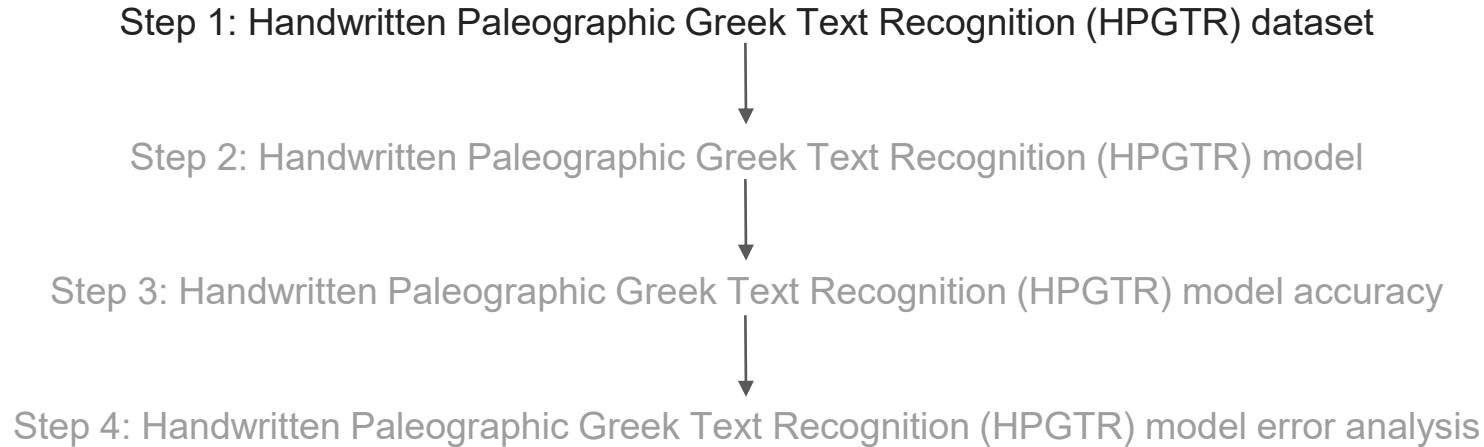
# Byzantine manuscripts and scripts

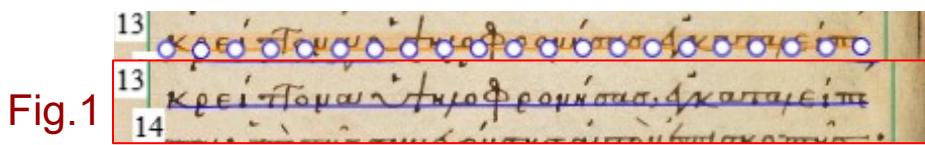
και πλάνατηνε. λαϊ βότων  
παγκαστλήγετομάν ομηρη  
άπορογρυπτοστηπαλέτην  
μέρη. λαϊ τούτοι μέμηνέ κεδαλ  
σαπτάδαδ. σεπτηνικέμπη  
κεκρήτηκεθεατρη. λαμπη

Πολλὴ μερὲπινη Δορή πηκαίας. πολλού  
λεγηπησιωνεγευταιόποροσ  
καιούμπρασ. λαϊ τοίασσοιχάρισκιώ.  
τίμεδημηέμαστμισθοράμηται  
10th c. AD

ΤΟ ΤΕ θωδειλεδο  
μετηνεσσον. λαγρομ  
αυτοίστασταρδην.  
περγστημκούμη.  
τηλινάσσορματτη  
μεσημβαίδητης

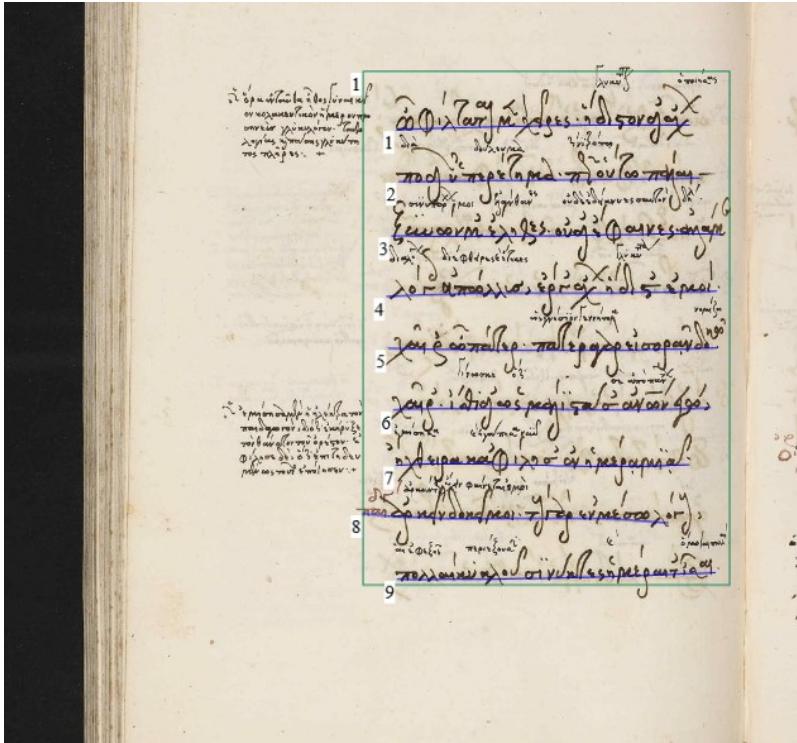
**Paraskevi Platanou, John Pavlopoulos, and Georgios Papaioannou. 2022.**  
**Handwritten Paleographic Greek Text Recognition: A Century-Based Approach.**  
In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6585–6589,  
Marseille, France. European Language Resources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





14 κρείττονα υψηλοφρονησας εγκαταλειπε

14 κρείττονα υψηλοφρονησας εγκαταλειπε

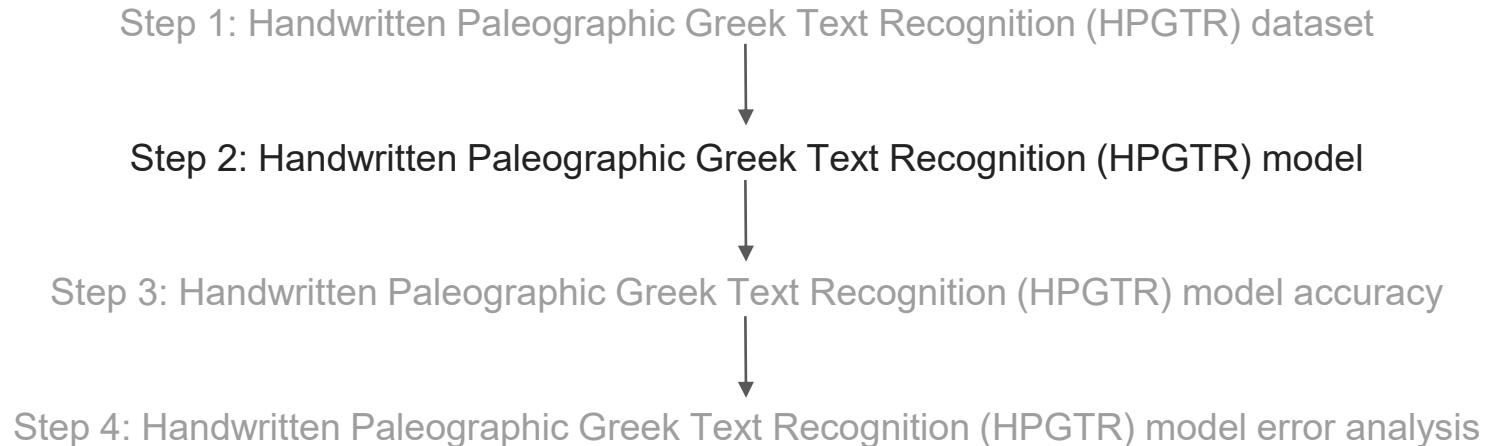


How do we effectively transcribe?

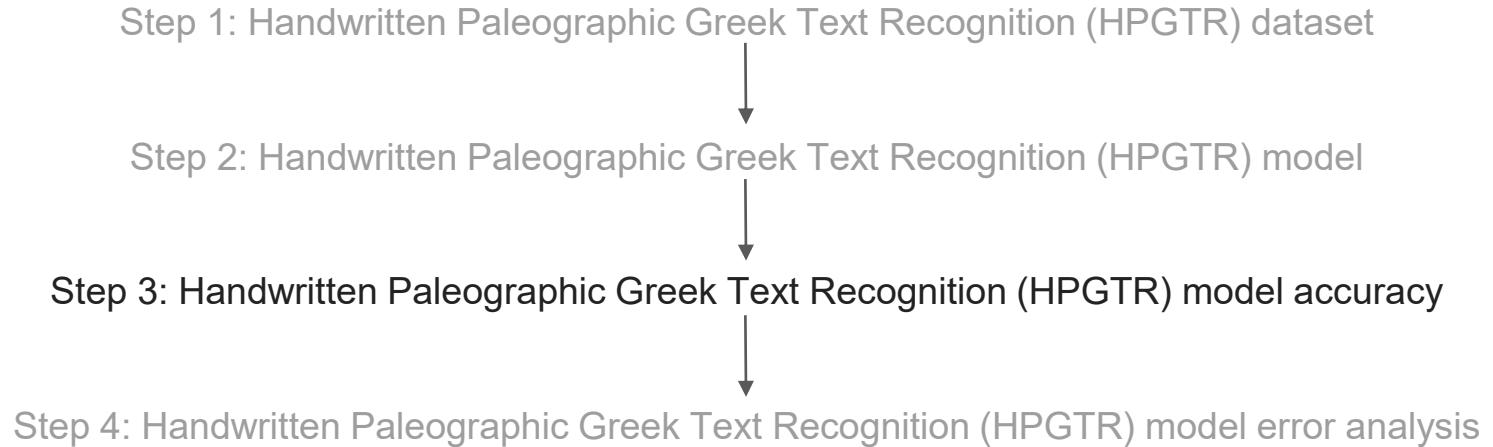
How do we manage effectively the layout?

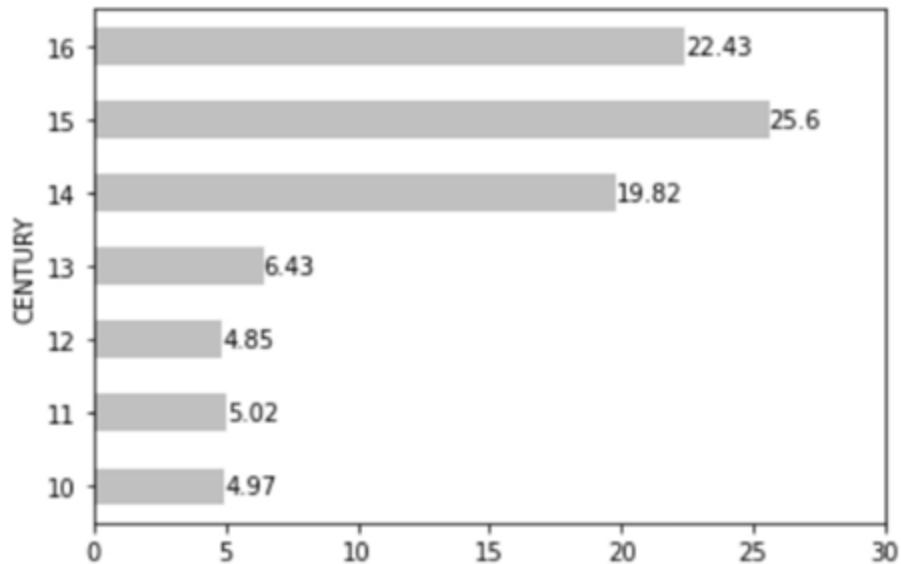
Fig.2

**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.

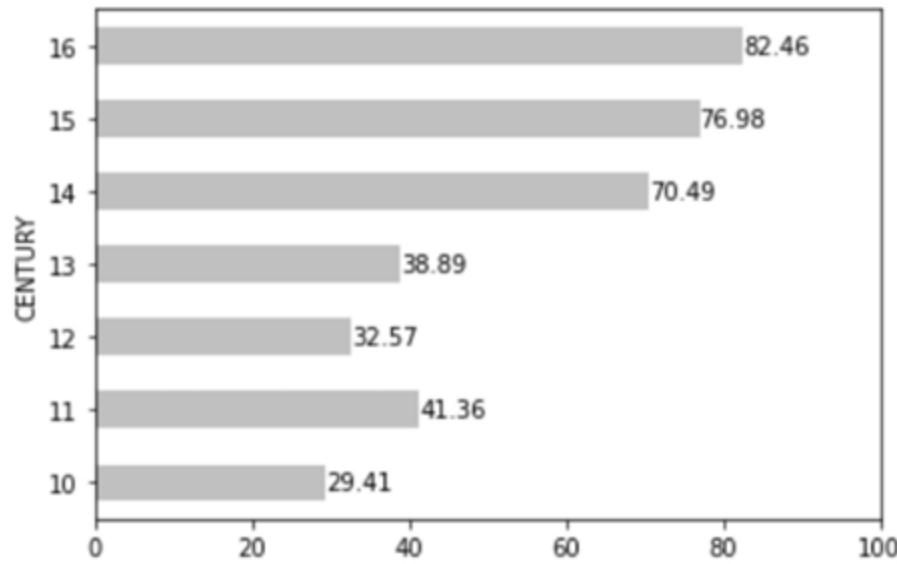


**Paraskevi Platanou, John Pavlopoulos, and Georgios Papaioannou. 2022.**  
**Handwritten Paleographic Greek Text Recognition: A Century-Based Approach.**  
In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6585–6589,  
Marseille, France. European Language Resources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



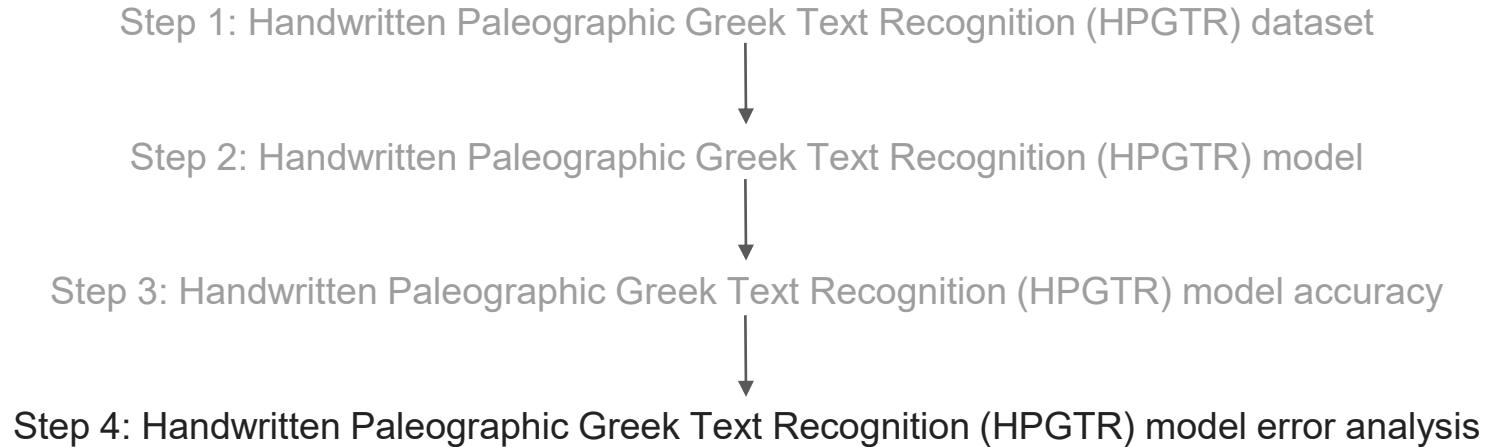


*Character error rate (percent shown horizontally) in HPGTR.N per century*



*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.



καὶ τὸν σελήνην πανταφέρειν· ὦ

σελήνην

ou

ἐσι· ὡς τῷ σόματισσον οὐδὲν τῇ καρδίᾳ



σελήνην

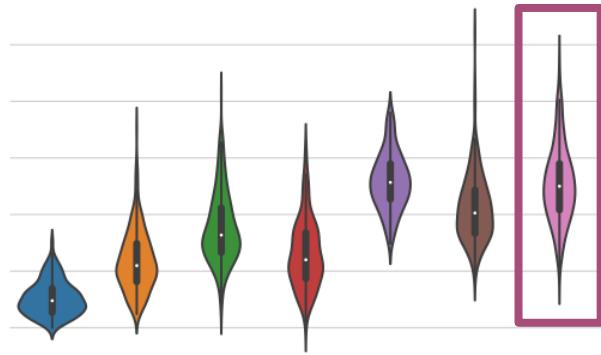
τῷ μαρτυρίῳ θράσυ γραφίοις / έγραψεν

ει

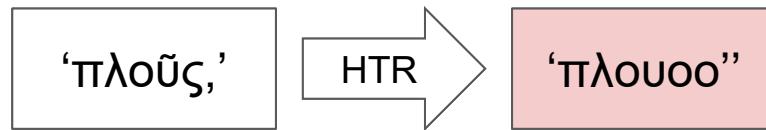
καὶ διώσκορος φύλακας τῆς παραγίδας τοῦ καλλίου

ευ

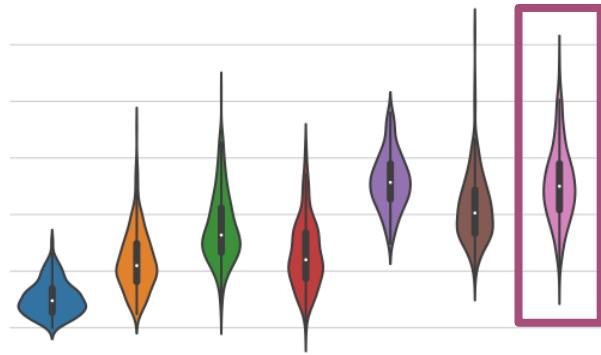
# Error analysis: the 16th c. CE



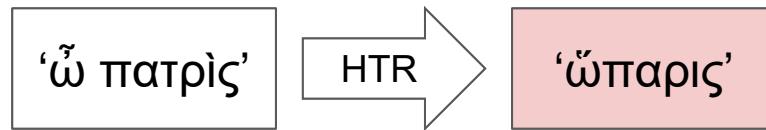
highest avg. CER



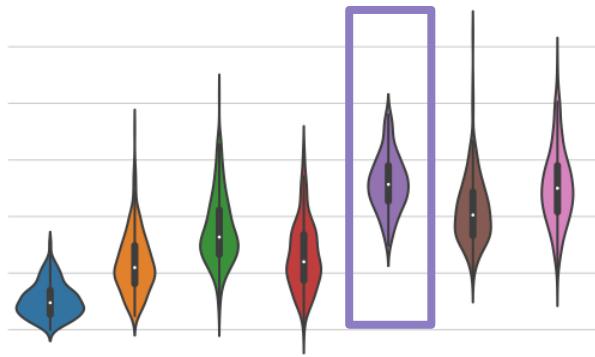
# Error analysis: the 16th c. CE



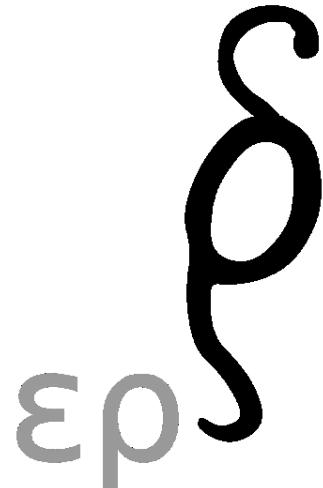
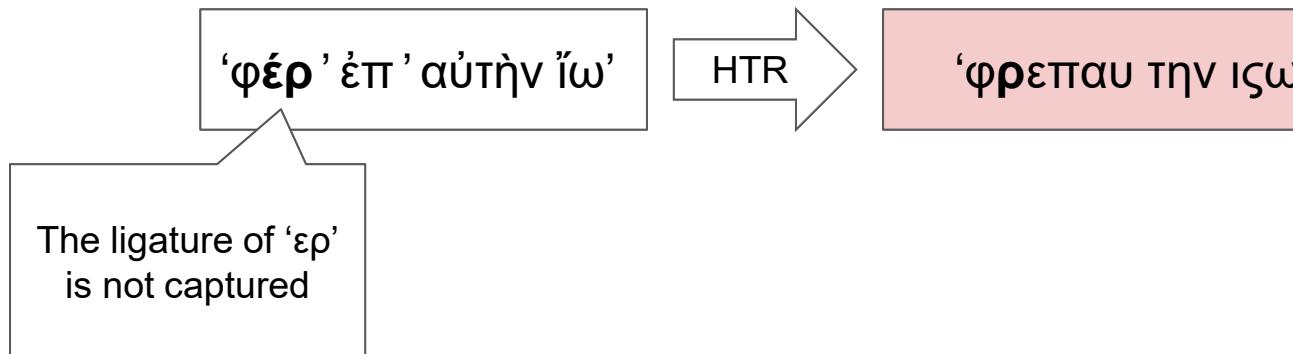
highest avg. CER



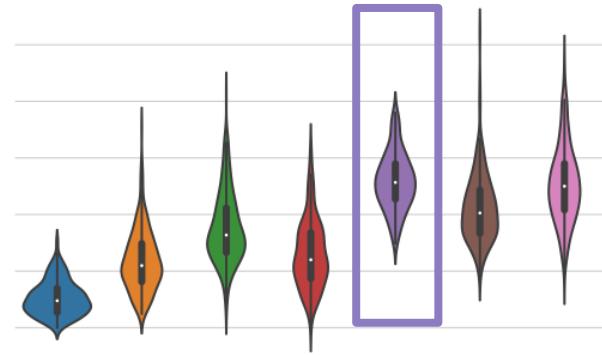
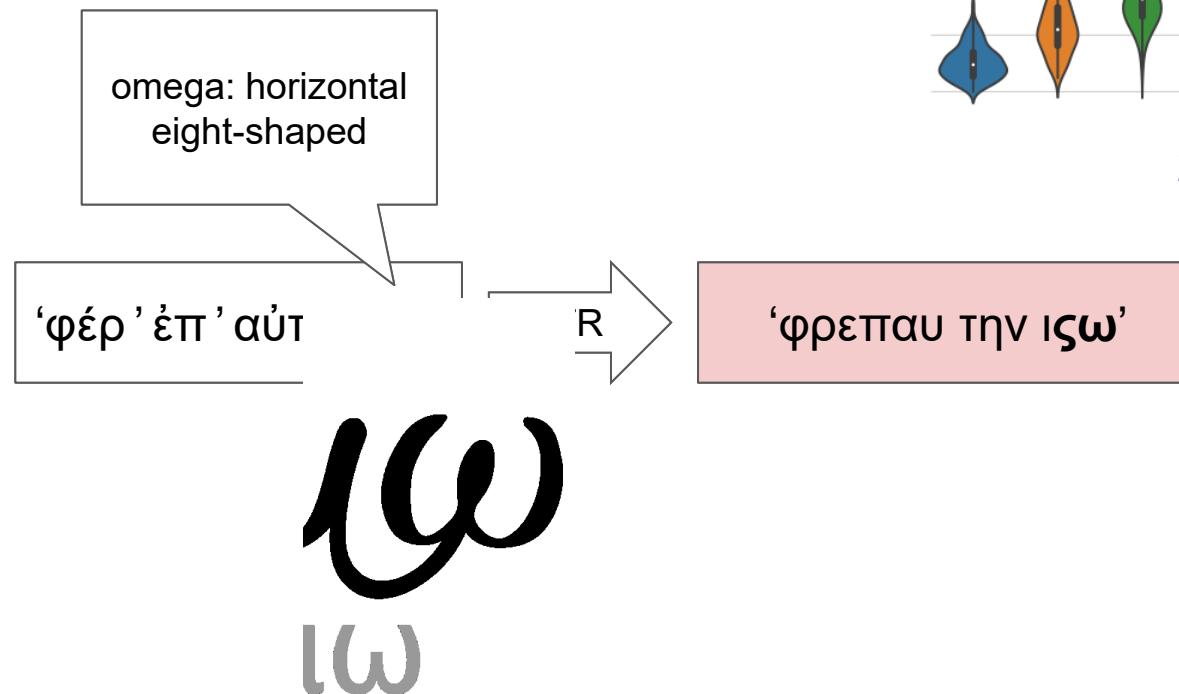
# Error analysis: the 14th c. CE



2nd highest



# Error analysis: the 14th c. CE



# Detecting Erroneous HTR output

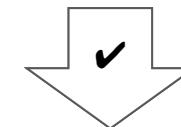
- HTR output yields diverse error rates
  - ⇒ manual (tedious, expensive) correction
  - ⇒ delaying the preservation of manuscripts
  - ⇒ hindering the recognition of historical manuscripts

# Detecting Erroneous HTR output

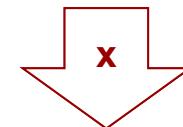
- HTR output yields diverse error rates
- Detecting erroneous/flawless HTR output
  - easier for **sbarecmld ttxe** v. non-scrambled text

# Detecting Erroneous HTR output

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



Τωη ποδων



ακοασδδεδ

# The Greek language: from antiquity to modern

10th CE to 16th c. CE

**Contemporary:** ~ spoken language

**Ancient:** including Atticised Greek

⇒ NLP with both wins

The data source →

## HTREC 2022

Improving the HTR output of Greek papyri and Byzantine manuscripts



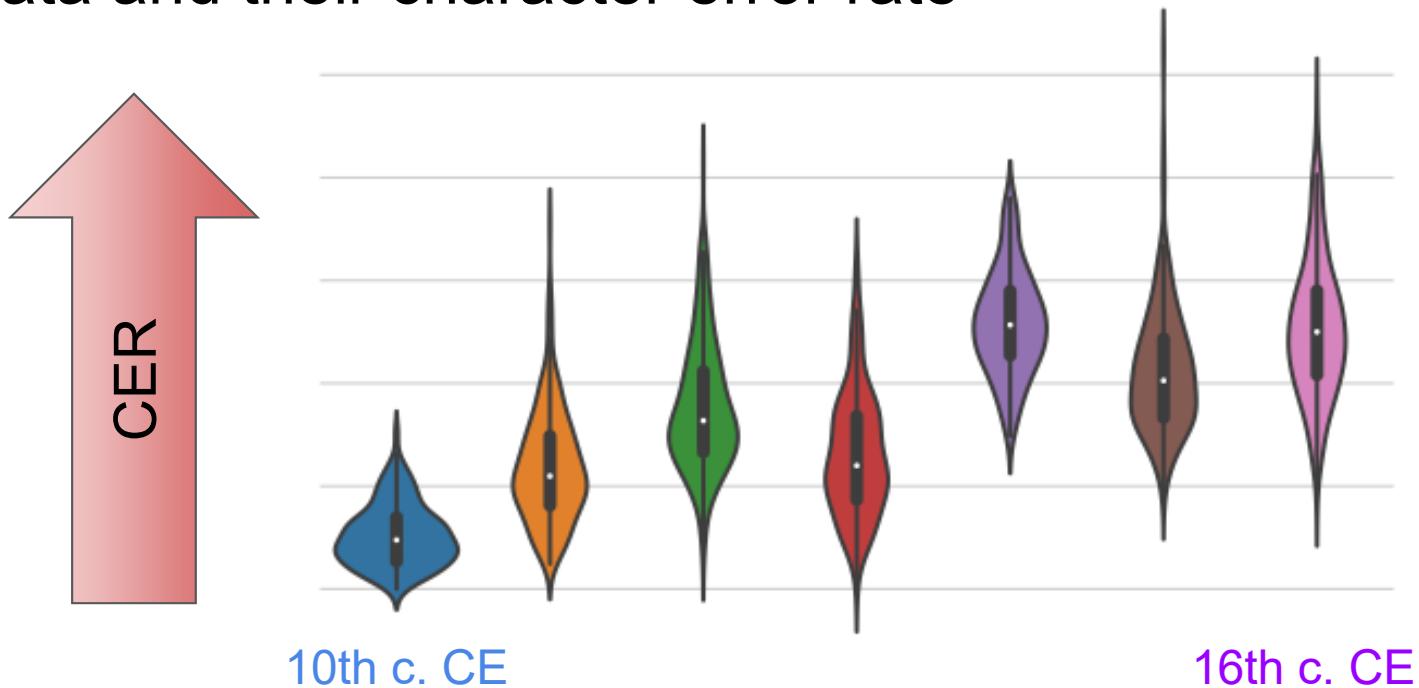
1 Travel Grant



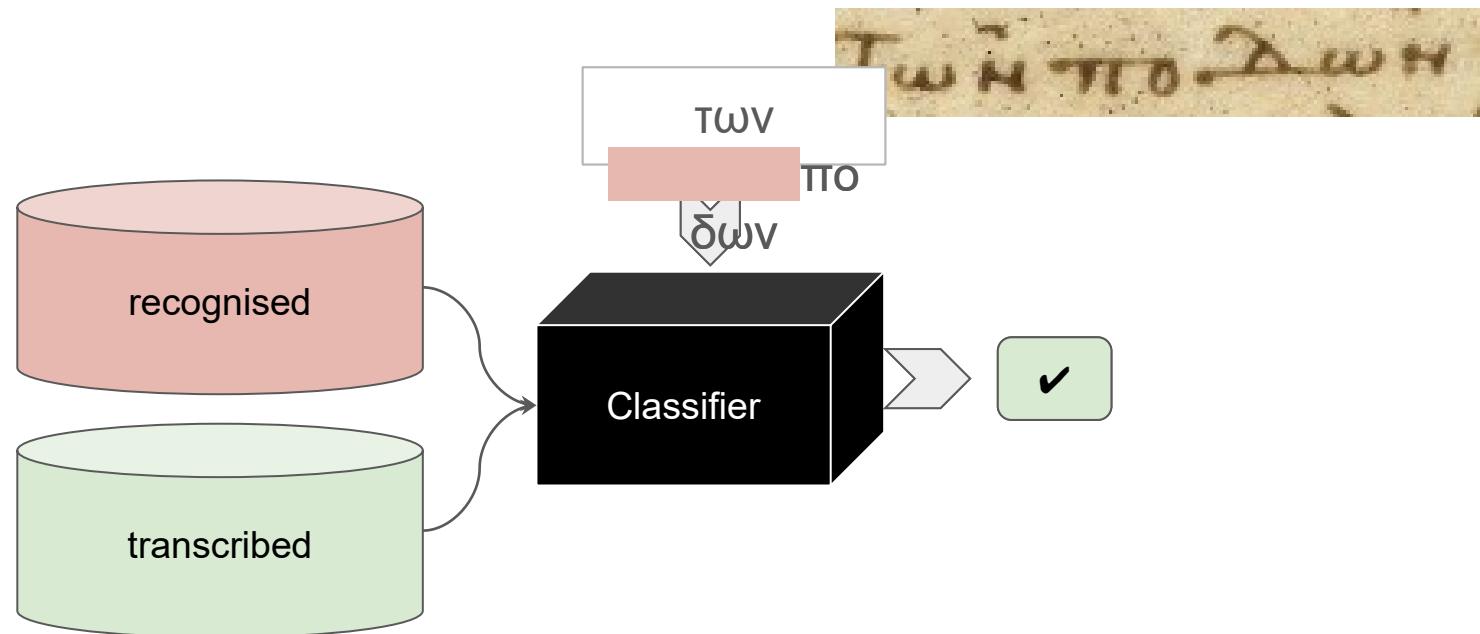
- under-trained Transkribus
- transcribed & recognised lines

Transcription	Recognition
έγγινομένα πάθη μὴ σβεννύντες ἀλλὰ τῇ εκλύσει (the born-in passions not extinguishing but the release)	έγγενομεναπαδημησμεννωτες ἀλλατήε 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 ἄμπελῶνα ἀλλὰ καὶ ὁ διὰ τες (- set on fire the vineyard but and the due to the)	εμπυριζου σιμαμπελῶνα ἀλλακαι ὅδξα

# The data and their character error rate



# Experiments: setup



# Experiments: machine learning

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

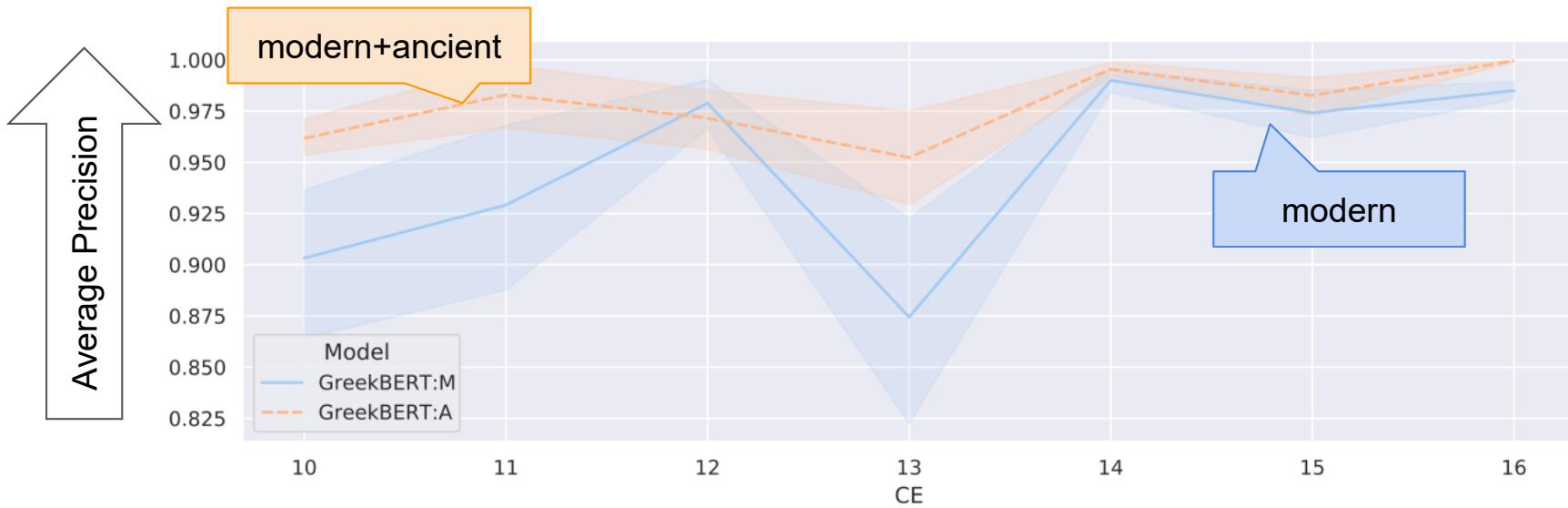
# Experiments: deep learning

	<b>AP</b>	<b>AUC</b>	<b>F1 (+)</b>	<b>F1 (-)</b>
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
<b>GRU</b>	0.79	0.79	0.68	0.71
<b>GreekBERT:M</b>	0.95	0.94	0.88	0.88
<b>GreekBERT:M+A</b>	<b>0.97</b>	<b>0.97</b>	<b>0.90</b>	<b>0.91</b>

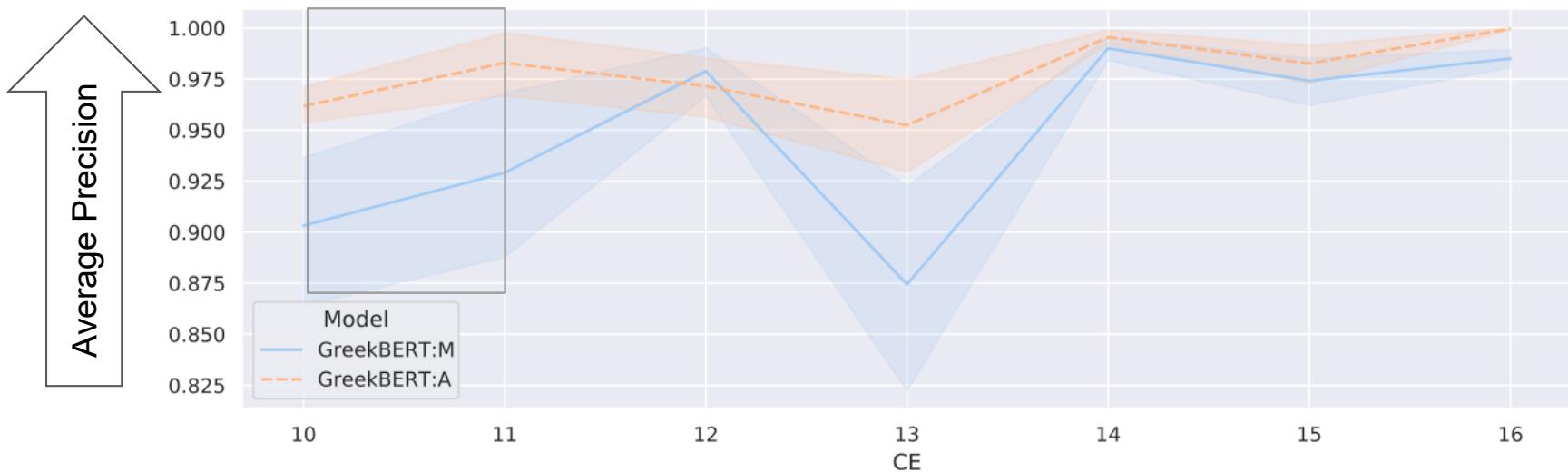
# Experiments: deep learning

	<b>AP</b>	<b>AUC</b>	<b>F1 (+)</b>	<b>F1 (-)</b>
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
<b>GreekBERT:M</b>	0.95	0.94	0.88	0.88
<b>GreekBERT:M+A</b>	<b>0.97</b>	<b>0.97</b>	<b>0.90</b>	<b>0.91</b>

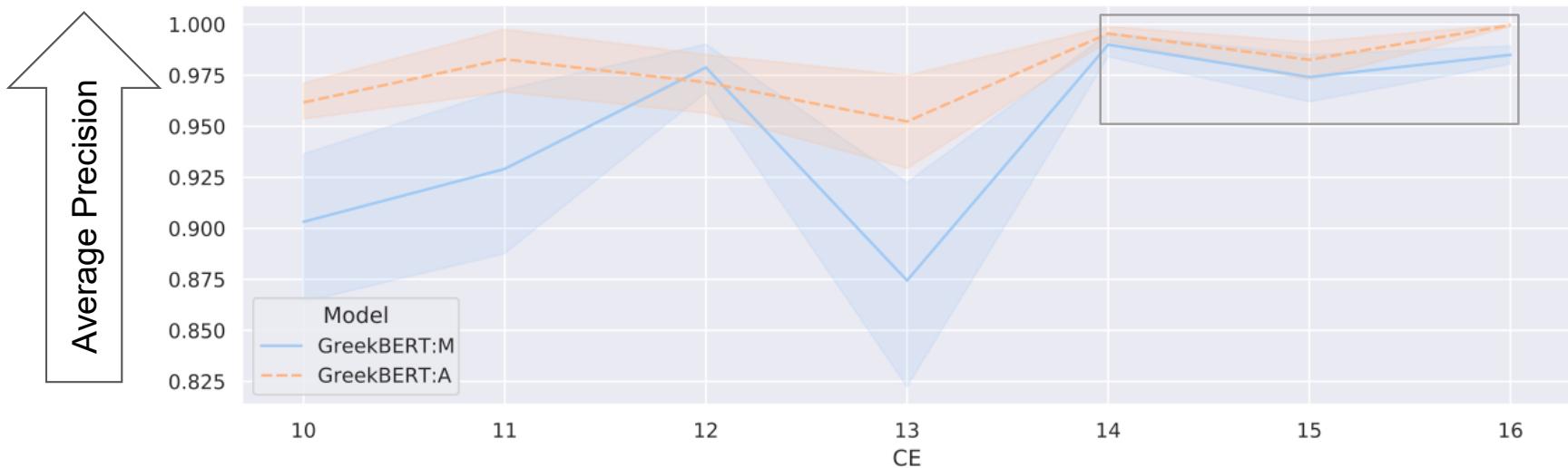
# Experiments: evaluation per century



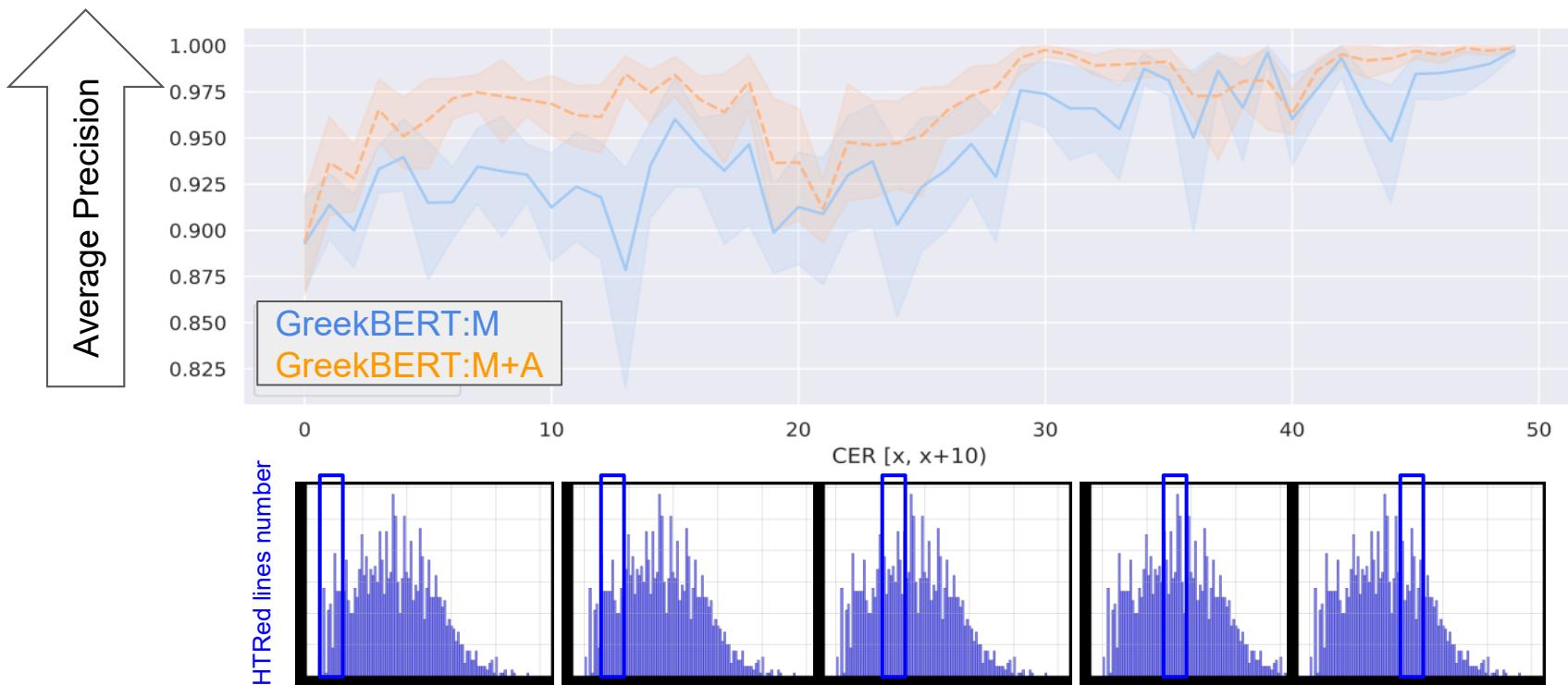
# Experiments: evaluation per century



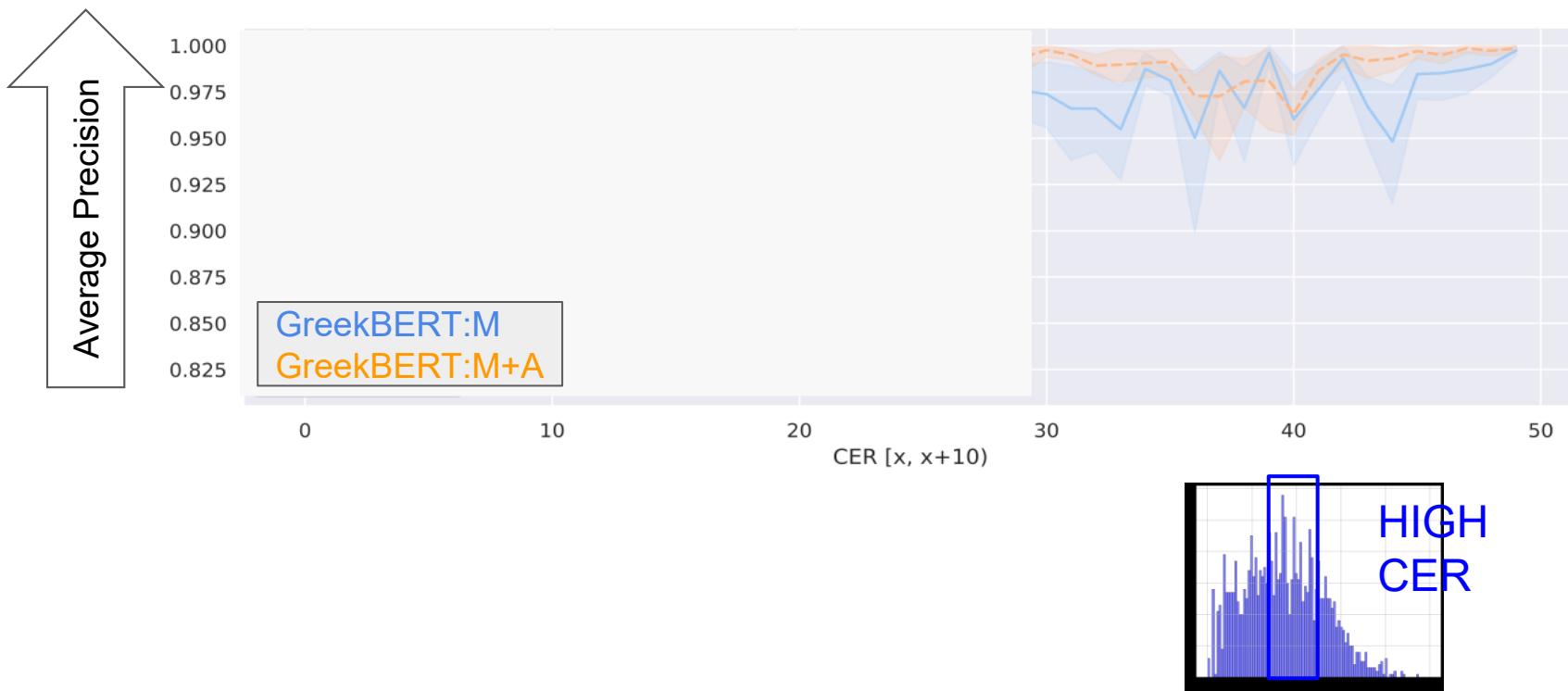
# Experiments: evaluation per century



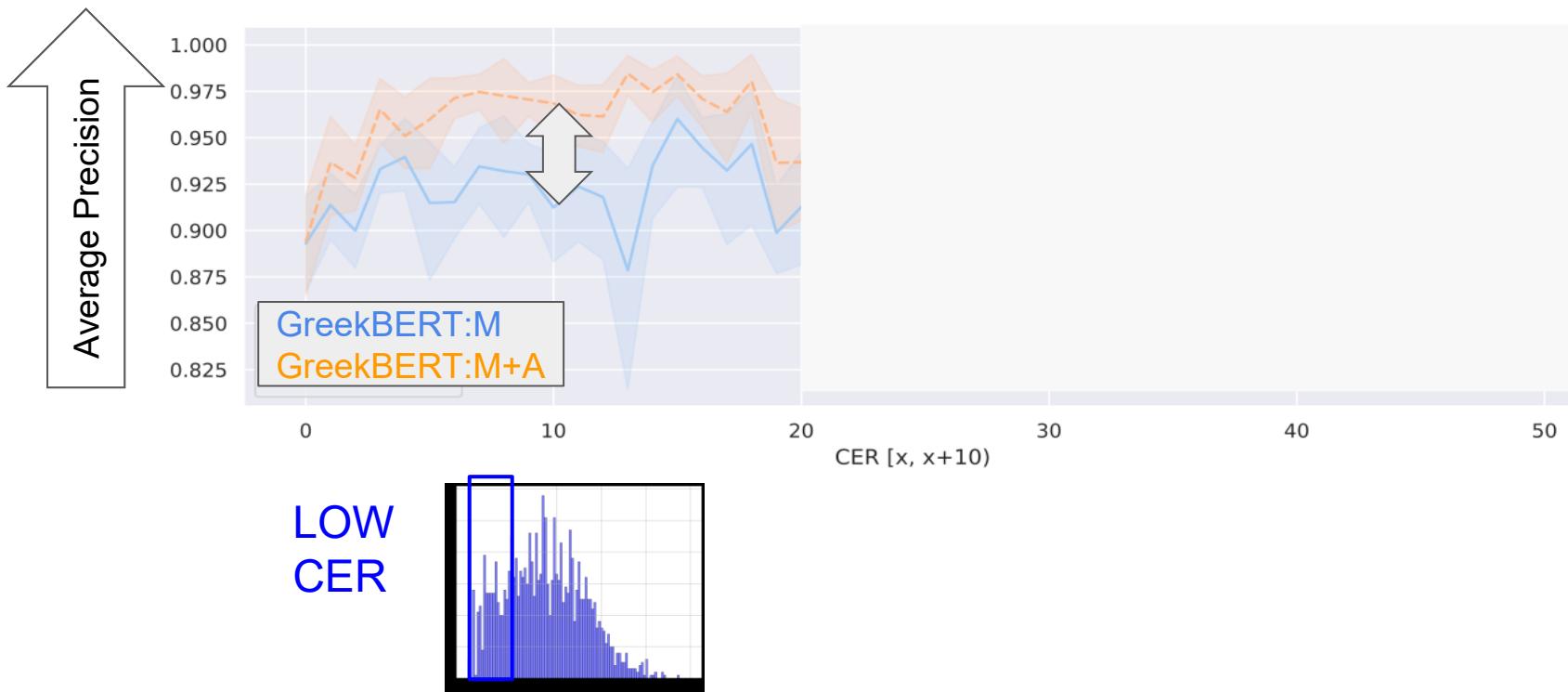
# Experiments: evaluation per error zone



# Experiments: evaluation per error zone



# Experiments: evaluation per error zone



# Error analysis

έγω δ ἀεί πως φιλακόλουθός είμι  
(I am always readily following)



# Error analysis

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

λα	και	γε	##δε	##ων	εκ	των	σκυλων	των	ισ	##μα	##ηλ	##η	##τι
----	-----	----	------	------	----	-----	--------	-----	----	------	------	-----	------

[...] ισμαηλιτι  
-κων [...]



# 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](mailto:maxime.guenette@umontreal.ca)

Université de Montréal

29/04/2024

Sunoikisis Session. HTR and OCR from papyrus to codex

# Codex Palatinus graecus

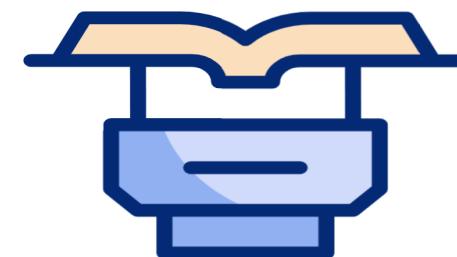
## 23

- 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

Sunoikisis Session. HTR and OCR from papyrus to codex

# 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

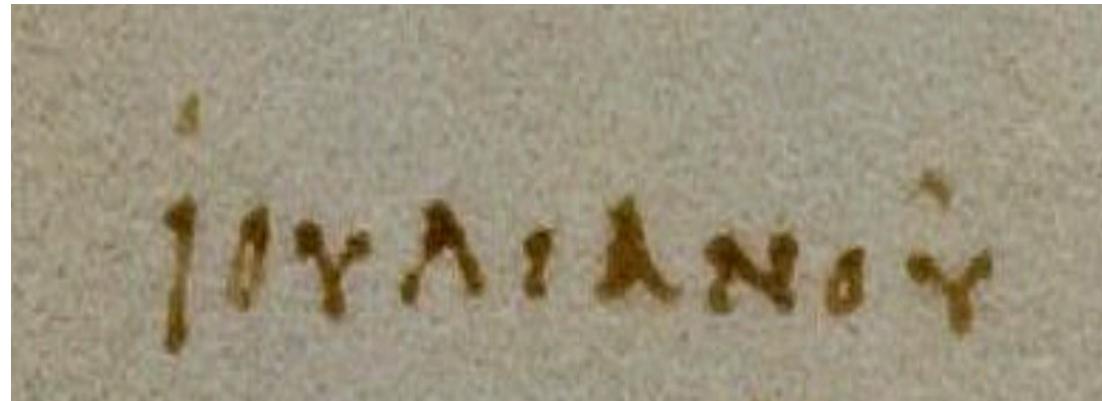
Handwritten Text Recognition



Sunoikisis Session. HTR and OCR from papyrus to codex

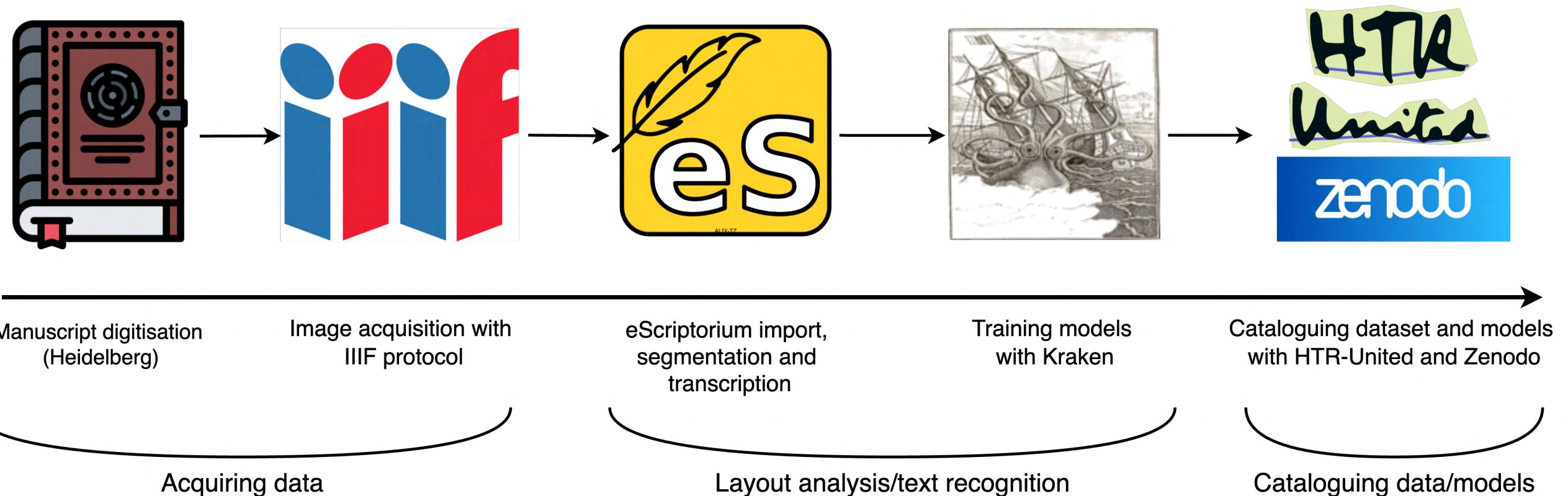
# State of the art

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



= ιουλιανοῦ

# Palatine Anthology pipeline



## 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 in and on the Byzantine Buildings of Constantinople
- ▶ Scholarly Annotations of the Digital Facsimile
- ▶ Bibliotheca Palatina

DOI / Citation link: <https://doi.org/10.11588/digit.3449> ⓘ

URN: urn:nbn:de:bsz:16-diglit-34495 ⓘ

Metadata: METS

IIIF Manifest: v2.1, v3.0

License: Public Domain Mark

Use/Order

Feedback

Download ▾

Overview

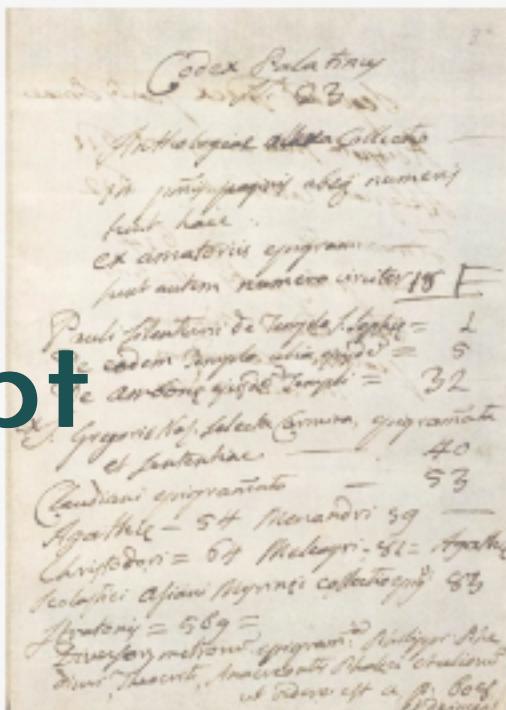
Quire structure



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... ⓘ



# 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](#)

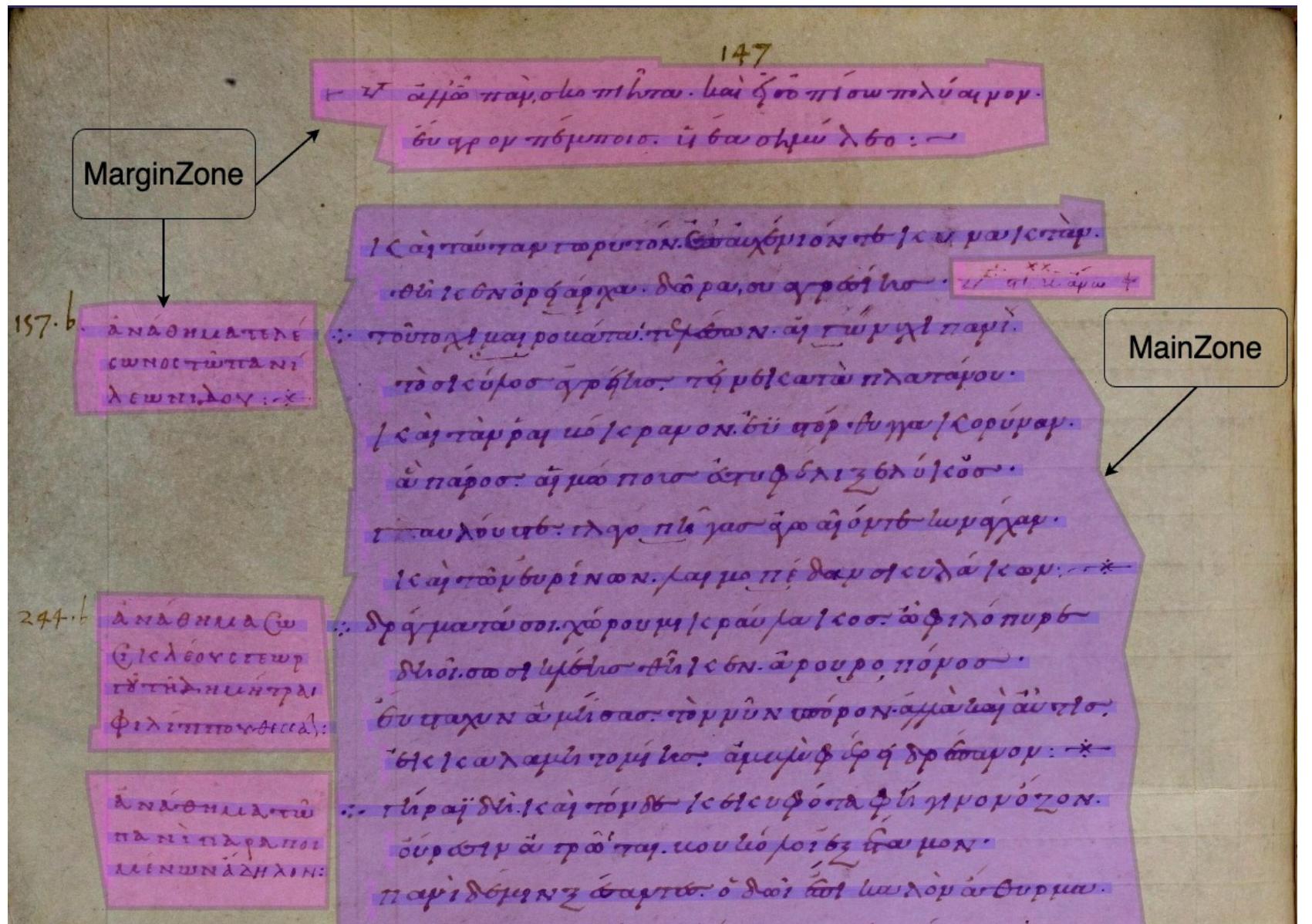


# eScriptorium

- 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



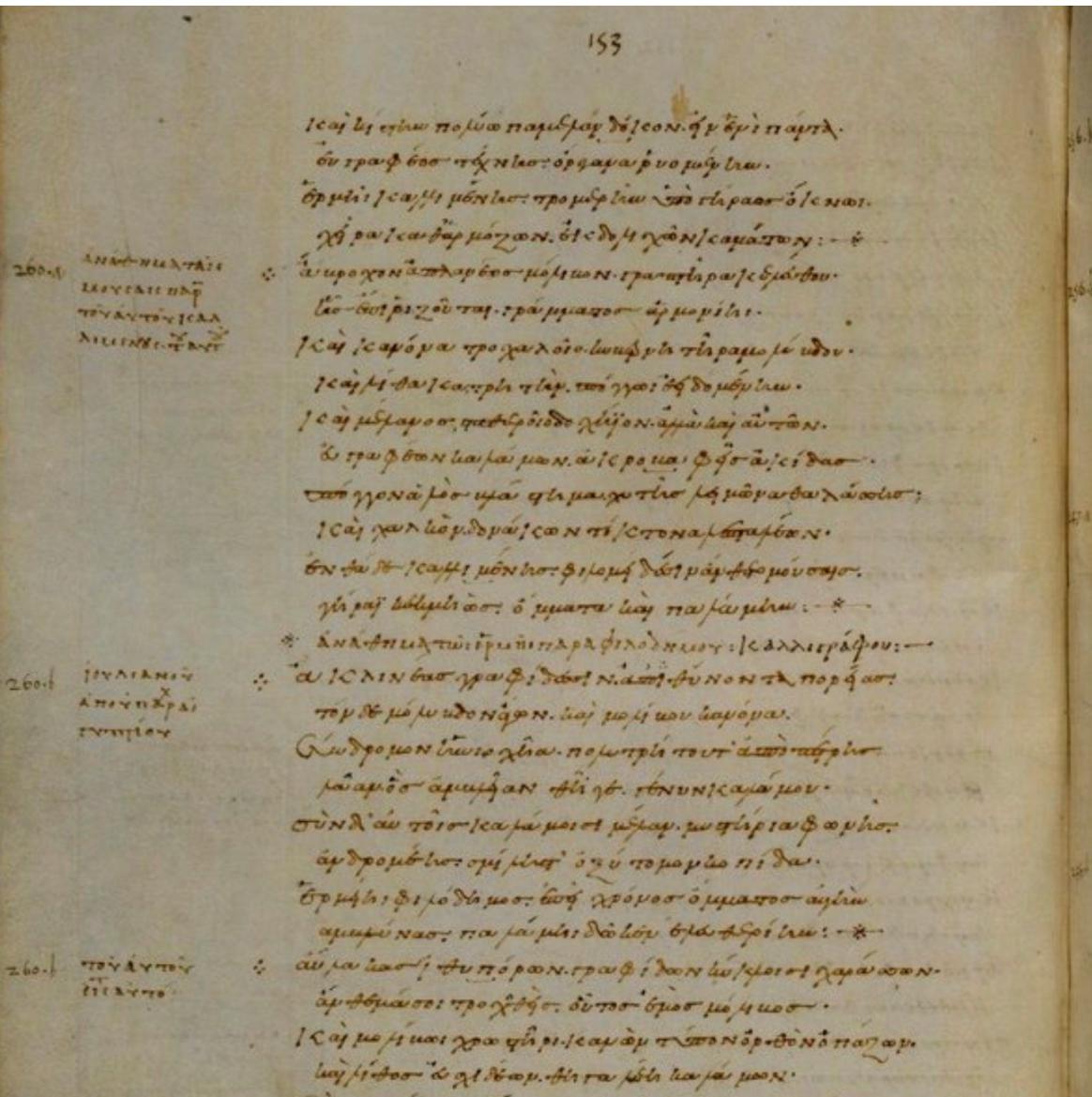
# Layout Analysis



Sunoikisis Session. HTR and OCR from papyrus to codex

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

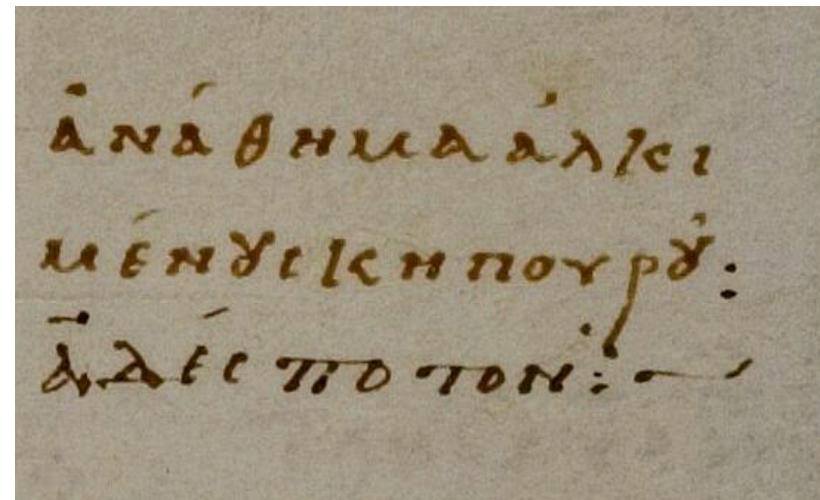
# Transcription



- καὶ κίστην πολύωπα μελανδόκον· εἰν ἐνὶ πάντα·  
εύγραφέος τέχνης· ὄργανα ρύομένην·  
έρμηι καλλιμένησ· τρομερὴν ὑπὸ γήραος ὅκνωι  
χεῖρα καθαρμόζων· ἐκ δολιχῶν καμάτων: \*
- ἀνάθημα ταῖς  
μούσαις παρὰ  
τοῦ αὐτοῦ καλ  
λιψενουσ· του αυτου
- ∴ ἀβροχον ἀπλανέοσ μόλιβον· γραπτῆρα κελεύθου·  
ἥσ ἔπι βίζουται· γράμματοσ ἀρμονίηι·  
καὶ κανόνα τροχαλοῖο· κυβερνητῆρα μολίβδου·  
καὶ λίθακα τρητήν· σπόγγωι ἐειδομένην·  
καὶ μέλανοσ σταθεροῖο δοχῆιον· ἀλλὰ καὶ αὐτῶν·  
εύγραφέων καλάμων· ἀκροβαφεῖσ ἀκίδασ·  
σπόγγον ἀλὸσ βλάστημα· χυτῆσ λειμῶνα θαλάσσησ·  
καὶ χαλκὸν δονάκων τέκτονα λεπταλέων·  
ἐνθάδε καλλιμένησ φιλομειδέσιν ἄνθετο μούσαισ·  
γήραϊ κεκμητῶσ· ὅμματα καὶ παλάμην: \*
- \* ἀνάθημα τῷ ἔρμῃ παρὰ φιλοδήμου: καλλιγράφου: ~
- ίουλιανοῦ  
ἀπὸ ὑπαρ αὶ  
γυπτίου
- ∴ ἀκλινέας γραφίδεσσιν ἀπιθύνοντα πορείασ·  
τόνδε μόλυβδον ἄγων· καὶ μολίβου κανόνα·  
σύνδρομον ἡνιοχῆα· πολυτρήτου τ' ἀπὸ πέτρησ  
λᾶαν· δσ ἀμβλεῖαν θῆγε· γένυν καλάμου·  
σὺν δ' αὐτοῖσι καλάμοισι μέλαν· μυστήρια φωνῆσ·  
ἀνδρομέησ· σμίλησ τ' ὄξυτόμον κοπίδα·  
έρμείηι φιλόδημοσ· ἐπεὶ χρόνοσ ὅμματοσ αύγην  
ἀμβλύνασ· παλάμηι δῶκεν ἐλευθερίην: \*
- τοῦ αὐτοῦ  
εἰσ τ αὐτό
- ∴ αύλακασ ίθυπόρων· γραφίδων κύκλοισι χαράσσων·  
ἄνθεμά σοι τροχθείσ ούτοσ ἐμὸσ μόλιβοσ·  
καὶ μολέβωι χρωστῆρι· κανὼν τύπον ὄρθὸν ὀπάζων·  
καὶ λίθοσ εύσχιδέων θηγαλέη καλάμων.

# Transcription guidelines

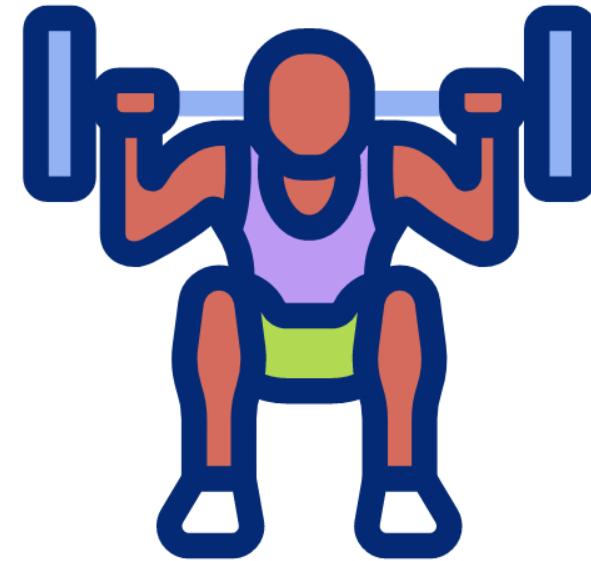
- Standardisation of sigma into “σ”
- Abbreviations are developed
- Significant symbols are transcribed
- Standardised punctuation (· and :)



άνάθημα ἀλκι  
μένουσ κηπουροῦ:  
ἀδέσποτον: ~

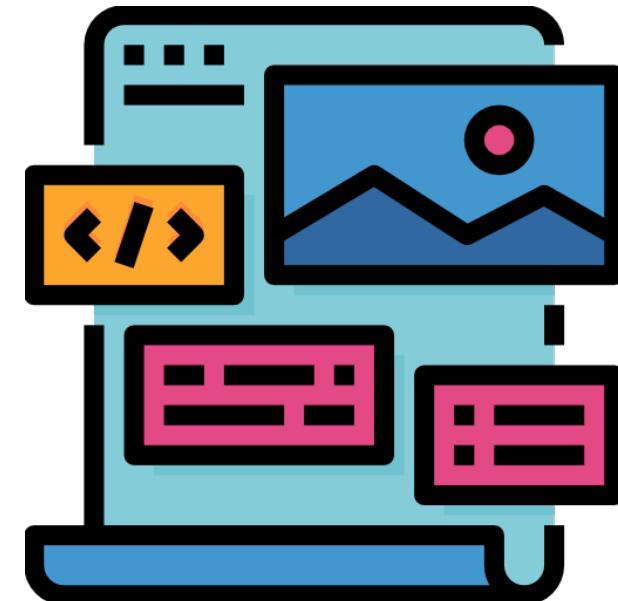
# Model training #1

- pp. 143-195, 50 pages<sup>1</sup>
- Segmenter model
- Recognizer model



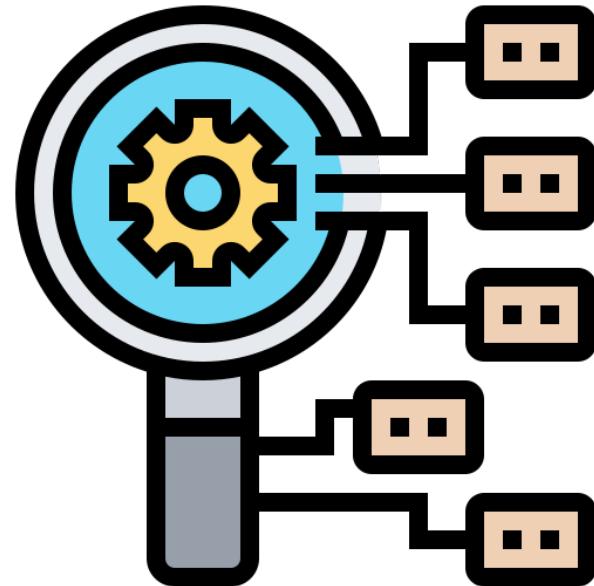
# *Segmenter*

- 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

```
1 from unicodedata import normalize
2 lettre = "ŵ"
3 len(normalize("NFC", lettre))
```

1

$\hat{\omega}$  =  $\hat{\omega}$  [U+1F66]

NFD = Normalization Form Canonical Decomposition

```
1 from unicodedata import normalize
2 lettre = "ŵ"
3 len(normalize("NFD", lettre))
```

3

$\hat{\omega}$  =  $\omega$  [U+03C9] + ' [U+0313] + ^ [U+0342]

Sunoikisis Session. HTR and OCR from papyrus to codex

# 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

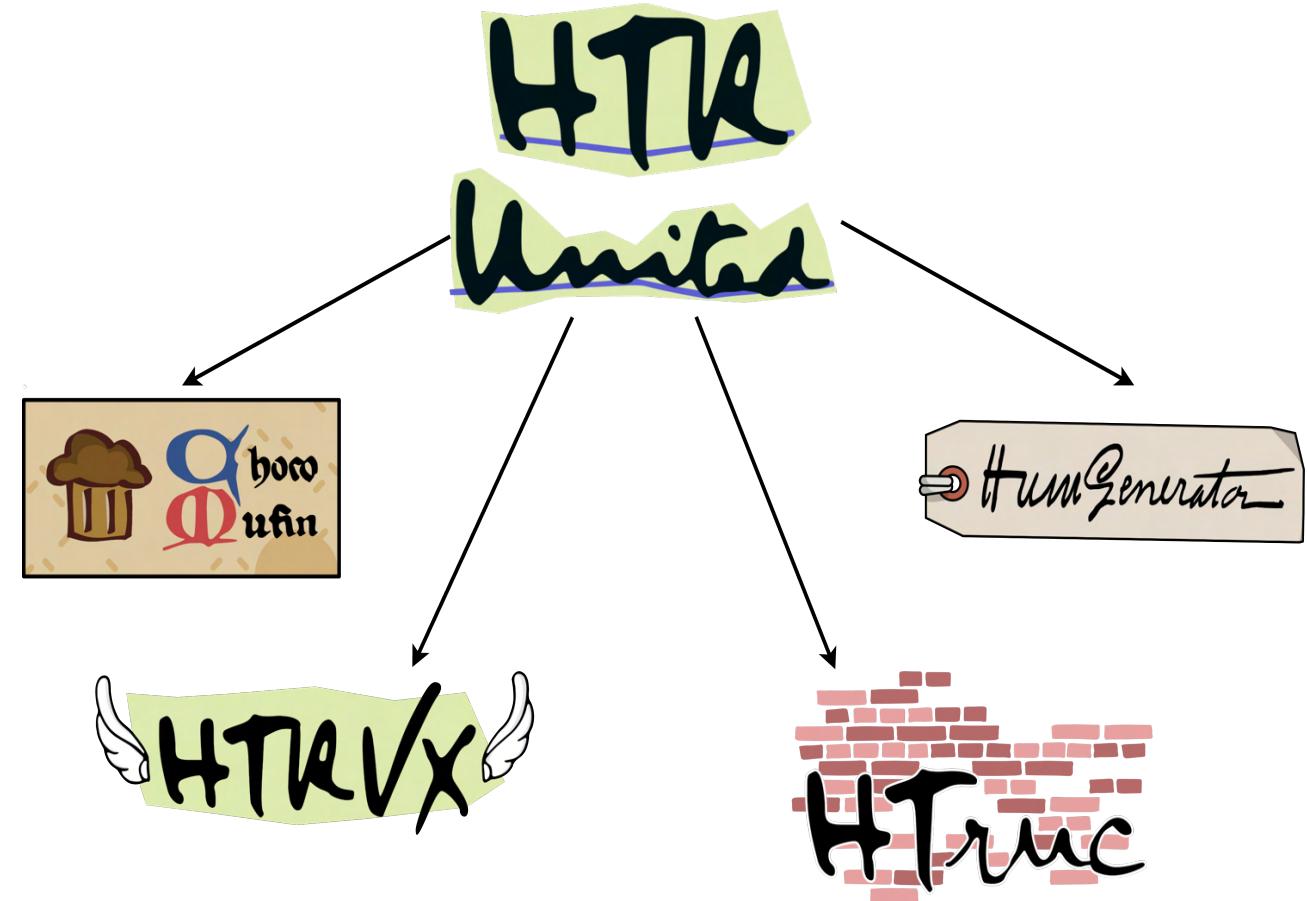


# Final results (v.1)

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

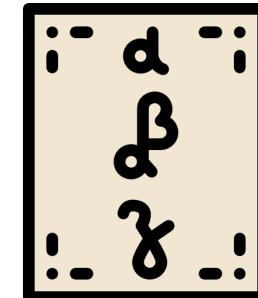
# FAIR data and standardised metadata

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



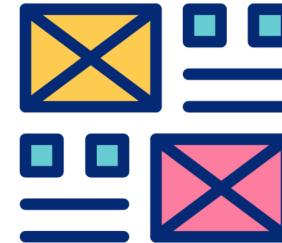
# Conclusion

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 handwriting,  
X<sup>th</sup> Byzantine minuscule



# Data repository

[https://gitlab.huma-num.fr/ecrinum/anthologia/htr\\_cpgr23](https://gitlab.huma-num.fr/ecrinum/anthologia/htr_cpgr23)