Ὅσοι ἄνθρωποι, τοσαῦται γνῶμαι

Harmonizing Guidelines for Handwritten Text Recognition of Ancient Greek

DH2025, Lisbon, July 14-18th, 2025

https://dhtr25.anthologiagraeca.org/

DH2025
ACCESSIBILITY
& CITIZENSHIP

Mathilde Verstraete, Maxime Guénette,
Matenia Vlachou, Marianne Reboul
& Marcello Vitali-Rosati

Université de Montréal

ENS DE LYON Institut de recherche et d'histoire des textes

cnrs

Motivation

- Growing number of **datasets** for Ancient Greek
- Expanding literature on HTR and **annotation** approaches
- Increasingly robust and generalizable **models**
- Clear need for **interoperable** practices and datasets ≠ silos

Textual Variability and Fluidity

Diacritics (accents & breathings)

Spelling

Script Types (bookhand, cursive)

Abbreviations, *nomina sacra*, ligatures

BROAD
Temporal,
Geographical,
Dialectal,
DISTRIBUTION

Capitalisation practices

Allographs

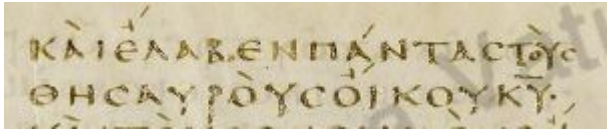
Visual polysemantism

Spacing, Punctuation & other Auxiliary Signs

Textual Variability and Fluidity

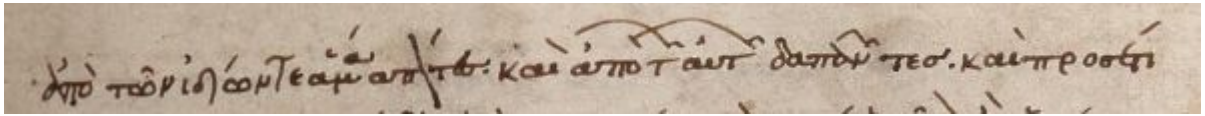
Script Types and Capitalisation practices

From Majuscule to Minuscule

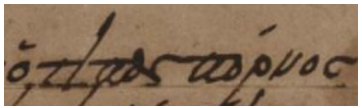


Vat. Gr. 1209 p. 426 (IVe s.)

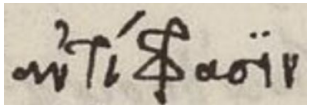
Vat. Gr. 126, f. 33^{r} (XIe s.)



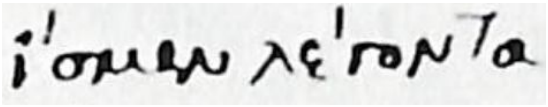
Various intermediate scripts



Vat. Gr. 2200, f. 53^{r} (VIIIe s.)



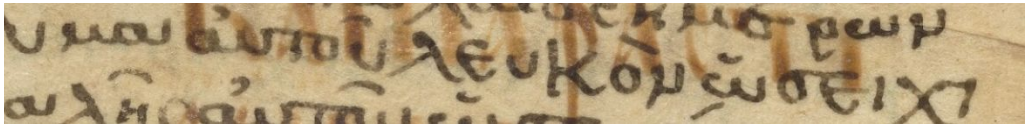
Barb. gr. 221, f. 56v (XV-XVIe s.)



Ut. Arch. 258, f. 5^{r} (XVe s.)

Textual Variability and Fluidity

Spacing, Punctuation & other Auxiliary Signs

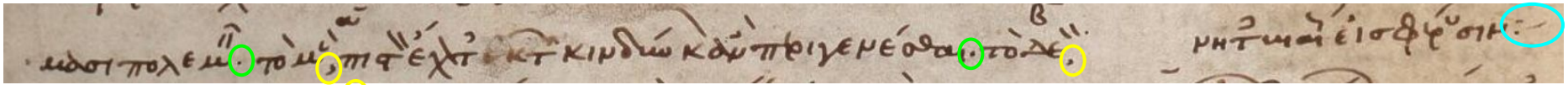


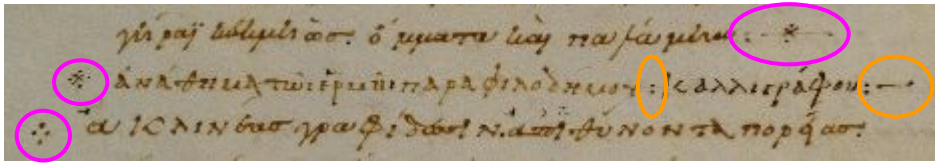
Bodl. Barocci 197*, 2^{v} (X-XIIIe s.)

ινουαὐτοῦλευκομῶσειχι (*scriptio continua*)

ἵνα οὖ αὐτοῦ λευκομῶσει χί (separated)

Vat. gr. 126, f. 33^{r} (XIe s.)





Pal. gr. 23, p. 153 (Xe s.)

Textual Variability and Fluidity

Spelling

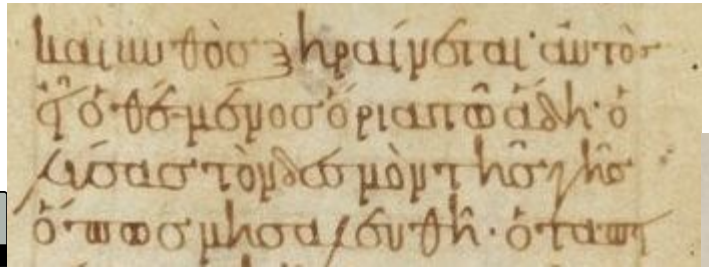
- [i] : ι η υ ει οι,
  - οἶκος or ἶκος,
  - χριστὸς or χρηστὸς,
- [o] : ο ου ω,
  - οἶκος or οἶκως,
  - λόγος or λώγος,
- [e] : ε or αι,
  - ἕτερος or ἑταῖρος,
  - κενός or καινός,
- Other morphological variants
  - crasis: ἐάω or ἐῶ,
  - apocope: λέγει or λέγ’

Textual Variability and Fluidity

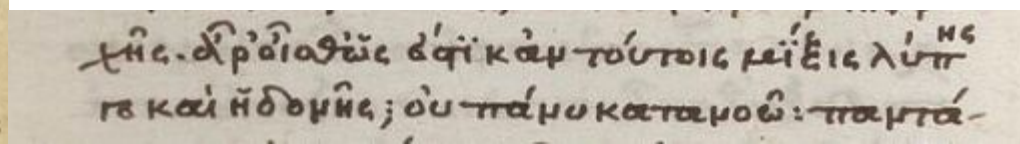
Diacritics

- τόνοι (combining or not)
  - rough breathing: ◌̔ (U+0314)
  - smooth breathing: ◌̓ (U+0313)
  - acute accent: ◌́ (U+1FFD or 00B4)
  - grave accent: ◌̀ (U+1FEF or 0060)
  - **perispomeni**: ◌͂ (U+1FC0 or 0020)
- χρόνοι (combining or not)
  - breve: ◌̆ (U+0306)
  - **macron**: ◌̄ (U+0304)

- διαστολή: , (U+002C) ?
- ὑφέν: ◌̵ (U+032E) ?
- ἀπόστροφος: ’ (U+02BC) ?
- iota subscript: ◌ͅ (U+0345)
- διαίρεσις: ◌̈ (U+00A8)



Par. gr. 1470, f. 135ᵛ (890).



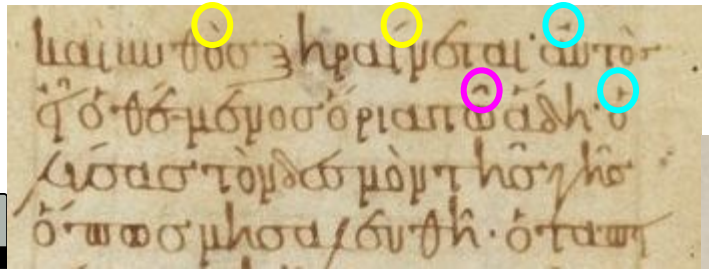
Vat. Urb. gr. 33, f. 31r (s. XV).

Textual Variability and Fluidity

Diacritics

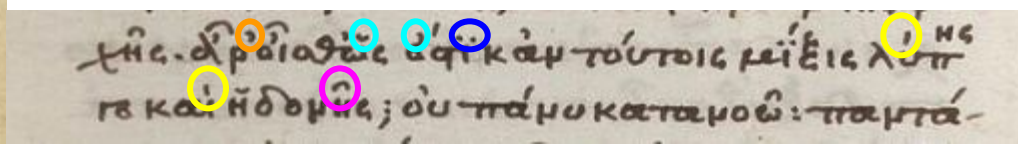
- τόνοι (combining or not)
  - rough breathing: ◌̔ (U+0314)
  - smooth breathing: ◌̓ (U+0313)
  - acute accent: ◌́ (U+1FFD or 00B4)
  - grave accent: ◌̀ (U+1FEF or 0060)
  - perispomeni: ◌͂ (U+1FC0 or 0020)
- χρόνοι (combining or not)
  - breve: ◌̆ (U+0306)
  - macron: ◌̄ (U+0304)

- διαστολή: , (U+002C) ?
- ὑφέν: ◌̵ (U+032E) ?
- ἀπόστροφος: ’ (U+02BC) ?
- iota subscript: ◌ͅ (U+0345)
- διαίρεσις: ◌̈ (U+00A8)



Par. gr. 1470, f. 135ᵛ (890).

Vat. Urb. gr. 33, f. 31r (s. XV).



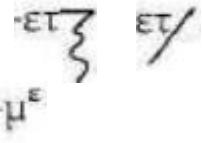
Textual Variability and Fluidity

Abbreviations, *nomina sacra*, ligatures

Abbr. by suspension:

Desinences:

- -εται
- -μενος



Monograms:

- μοναχός

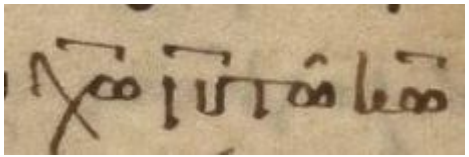
Numbers:

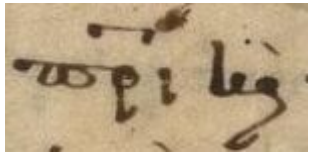
- X for χίλιοι
- H for ἑκατόν

Tachygraphic signs , e.g. ϗ for καί

Abbr. by contraction:

- (Nomina Sacra) Sacred words,
- Proper names
- Numbers





Vat. gr. 2079, f. 109^{v}

Ligatures (e.g):

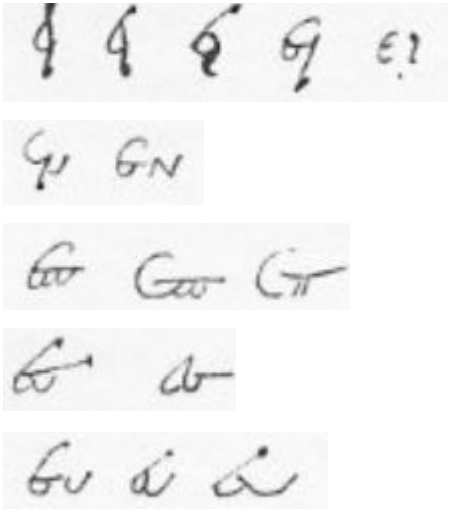
-ει

-εν

-επ

-εσ

-ευ

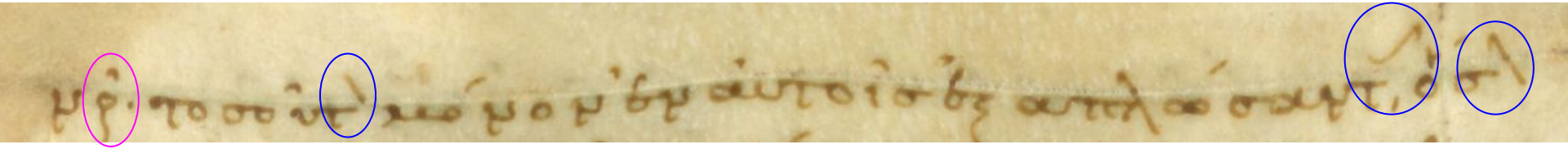


Agati (1984). Note paleografiche all' "Antologia Palatina". Boll.Class. 5, pp. 43-59 (fig.1l)

Textual Variability and Fluidity

Abbreviations, *nomina sacra*, ligatures

Vindob. Phil. G. 314, f. 110^{r} (925)



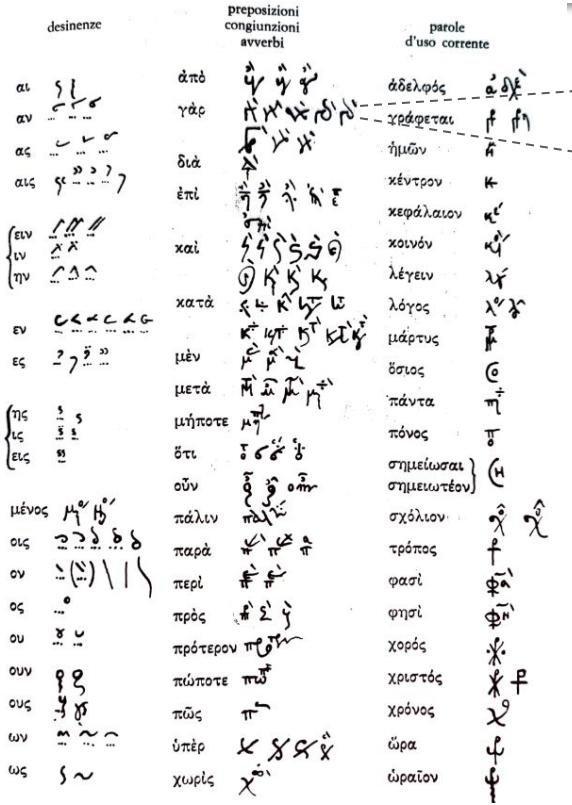
νοῦν · τοσοῦτὸν μόνον ἐν αὐτοῖς ἐξαπλώσαντάς , ὅσον

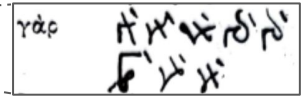
ν ⸗ · τοσοῦτ ⸗ μόνον ἐν αὐτοῖς ἐξαπλώσαντ , ὅσ ⸗

οῦν ὸν άς ον

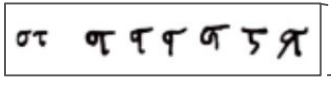
Textual Variability and Fluidity

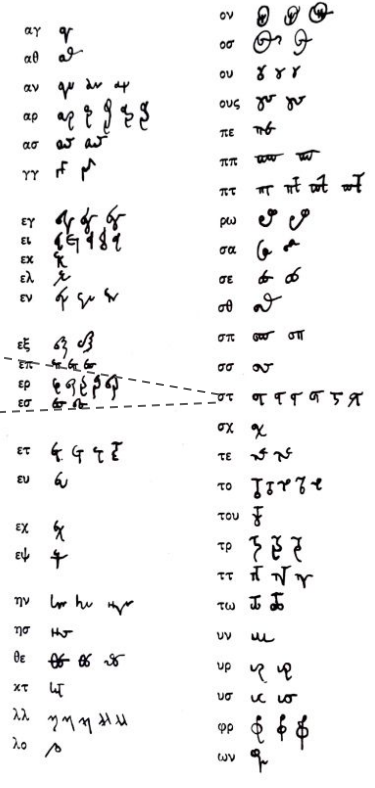
Abbreviations, ligatures





visible components → fully merged into ligature

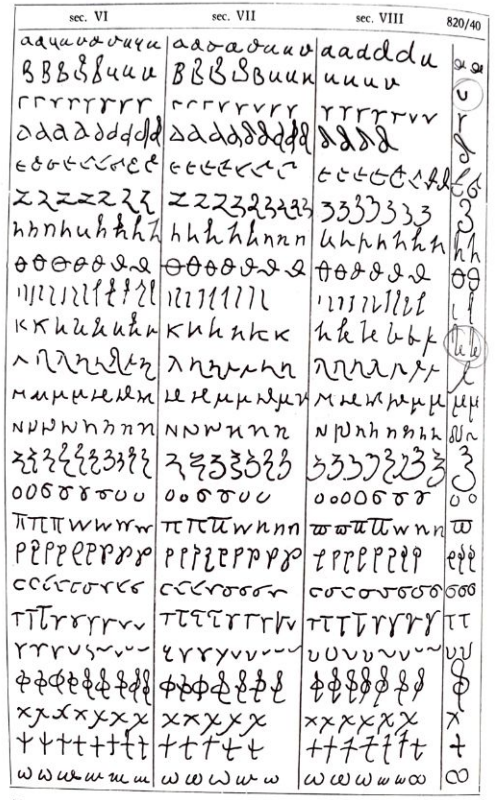




E. Mioni (1973) *Introduzione alla Paleografia Greca*, Padova, p. 97 - 98

Textual Variability and Fluidity

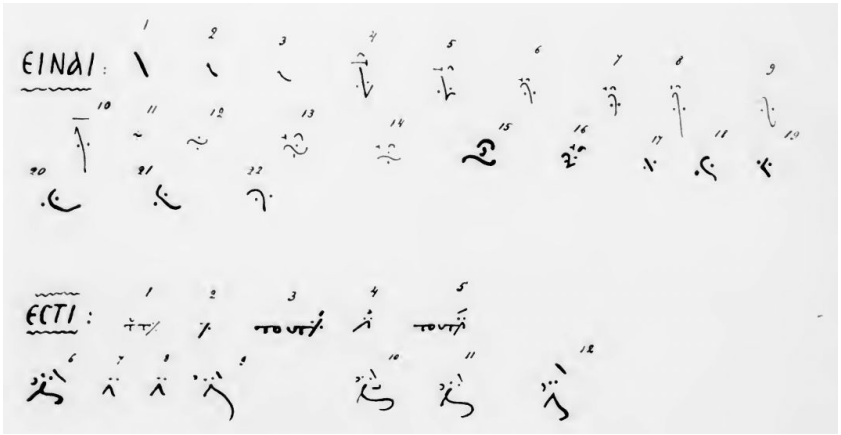
Allographs and Morphological Evolution



E. Mioni (1973) *Introduzione alla Paleografia Greca*, Padova, p. 62

Textual Variability and Fluidity

Visual polysemantism



 = -εισ, -ησ, -ισ

«κη» = 

G. F. Tsereteli (1904) *Sokrashcheniia v grecheskikh rukopisiakh*, St. Petersburg, tab. 7

Transcribing (in theory)

A non-universal notion:

- There is no single "correct" transcription
- Every transcription reflects methodological choices
- Transcription is always project-driven

Transcription = Interpretation:

→ Selection of what is deemed relevant

→ Normalization according to conventions

→ Contextualization within a framework of use

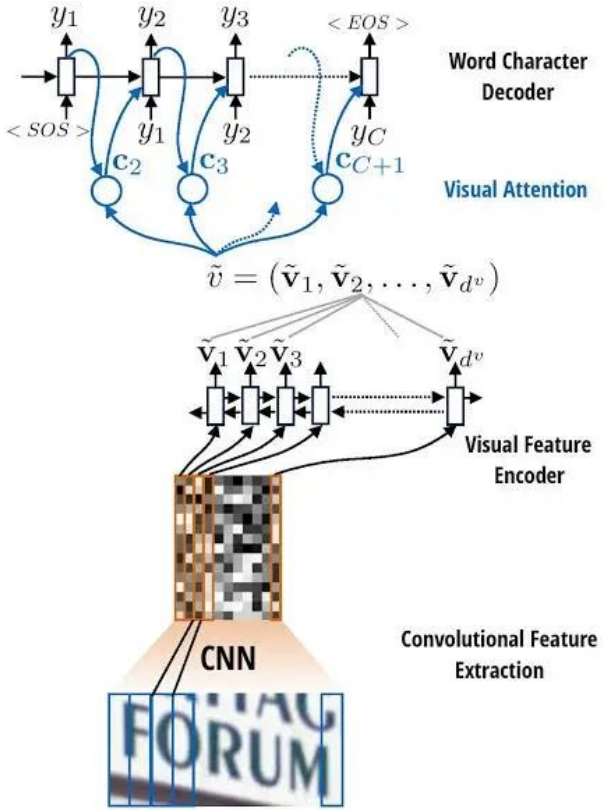
Transcribing (in practice)

Limitations of Unicode... :

- **Incomplete coverage** of:
  - some ligatures, abbreviations, diacritics combinations;
    - Should we always decompose them? (e.g.: ϗ or και?)
    - What do we do for those that don't exist in Unicode? Invent Markup?
  - some auxiliary signs;
- **Various encoding options** for a same form:
  - e.g.: ᾰ́ =
    - U+1F04 (precomposed)
    - or U+03B1 + U+0313 + U+0301 (decomposed)
  - oxia vs tonos, e.g.: ά = U-1F71 (Greek Small Letter Alpha with Oxia) or U-03AC (... with Tonos) ?

... Result as a **consistency nightmare** for HTR !

Classic HTR (RNN/LSTM): Visual Mapping



What it does...

- Visual "recognition": converts visual patterns (**glyphs**) into character sequences → close proximity between visual and semantic
- "Remembers" indirectly immediate context (preceding/following tokens)

What it doesn't do...

- Semantic interpretation
- Automatic abbreviation expansion (should this be part of HTR?)
- Spelling correction or normalization (spacing, punctuation, capitalization)

https://github.com/da03/Attention-OCR

Limitations for Historical Material

What HTR can do...

- Large-scale textual analysis
- Content discovery
- First-draft transcriptions

What HTR cannot do...

- Perfect transcription
- Precise paleographical analysis

→ Need for post-correction

Representativity → Allographs

- Rare variants underrepresented
- HTR struggles with unseen forms

New age HTR (ViT/LMM): Beyond recognition

Potential capabilities:

- Leverage external knowledge → contextualize and normalize
- Resolve abbreviations automatically
- Work with fewer examples (few-shot learning)

Base unit: ~~Glyph/Character~~ → Word-level or subword tokens

Limitations & challenges: Hallucinations (how can they be controlled?)

💡 **Key questions:**

- What should training data look like?
- Which tasks belong in HTR vs. post-processing?
- What task do we entrust the machine with?

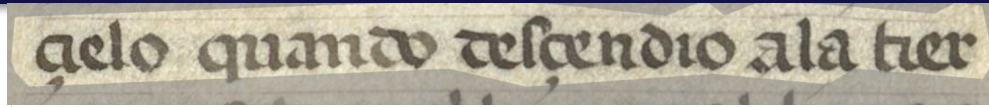
Transcription choices

Three* main transcription approaches:

1. Graphetic
2. “Editorial”/”Regularized”
3. Graphemic
4. * Mixed

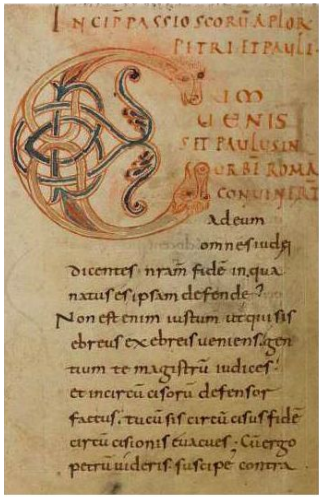
→ The choice depends on the desired granularity of information

1.



çielo quando deſçendio ala tier

2.



Passio sanctorum apostolorum Petri et Pauli.

1 Cum uenisset Paulus Romam, conuenerunt ad eum omnes Iudaei dicentes: Nostram fidem, in qua natus es, ipsam defende. non est enim iustum, ut cum sis Hebraeus ex Hebraeis ueniens, gentium te magistrum iudices, et incircumcisorum defensor factus tu cum sis circumcisus, fidem circumcisionis euacues. cum ergo Petrum uideris, suscipe contra eius doctrinam, quia

3.

cielo quando descendio a la tier

Existing Datasets

Dataset	Century	Script	Size (char.)	Approach	Abbr.	Diacritics	HTR Model
Eparchos	16th CE	Late byz. minuscule	116 894	Graphemic	Resolved	✅	❌
Méléagre	10th CE	Byzantine minuscule	114 273	Graphemic	Unresolved	✅	✅
Zenon Papyri	3rd BCE	Cursive	5 850	Graphemic	Unresolved	❌	❌
Stavronikita	14-16th CE	Mixed minuscule	103 080	Graphemic	Resolved	✅	❌
HPGTR	8-17th CE	Mixed minuscule	64 952	Graphemic	Resolved	✅	❌
Chrysostomicus I	10-14th CE	Mixed minuscule	56 791	Graphemic	Resolved	✅	✅

This workshop

Today's Goals

- Build an interdisciplinary community around HTR for Ancient Greek, bringing together philologists, paleographers and HTR users *(aka, you)*;
- **Identify the challenges** and needs of the community;
- Encourage participatory, use case-driven and **collaborative decision-making**;
- Develop **shared guidelines** for the transcription and encoding of Ancient Greek.

Today's Overview

- [09h-09h20 : Introduction]
- 09:20-11:00 : Case Studies
  - [#1 : M. Verstraete & M. Guénette]
  - #2 : E. Perdiki
  - #3 : A. Jambé
  - #4 : C. Vidal-Gorène (online)
  - #5 : I. Marthot-Santaniello
- 11:00-11:15 : Coffee break
- 11:15-12:30 : Discussions around guidelines

Website: https://dhtr25.anthologiagraeca.org/
Shared notes:
https://docs.google.com/document/d/1S0xRZb1ImEBfKL6xPu7KmB9dSr15MSYZlFP_S8ZwR2A/edit?usp=sharing