ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Detecting Erroneous Handwritten Byzantine Text Recognition

J. Pavlopoulos, V. Kougia, P. Platanou, H. Essler

universität wien

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Università Ca'Foscari Venezia

Detecting Erroneous HTR output

HTR output yields diverse error rates

⇒ manual (tedious, expensive) correction

⇒ delaying the preservation of manuscripts

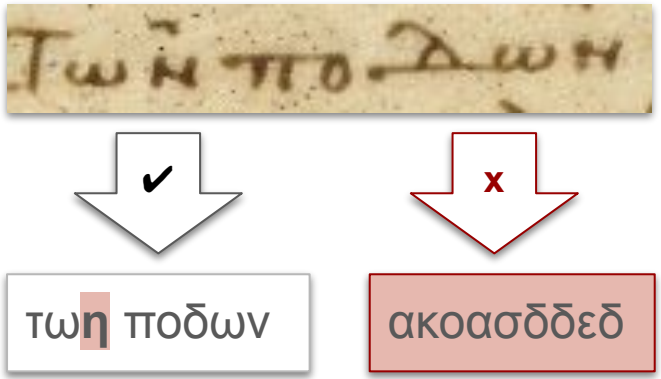
⇒ hindering the recognition of historical manuscripts

Detecting Erroneous HTR output

detecting erroneous/flawless HTR output

👉 easier for scrambled text

improved post-correction (e.g., when not to post-correct)



The Greek language: from antiquity to modern

10th to 16th c. CE

Contemporary: resembling what was considered as spoken language

Ancient: including Atticised Greek

NLP using both is beneficial

The Greek language: from antiquity to modern

Text from manuscripts and papyri in Byzantine Greek

Morphological categories gradually decreased or disappeared

Infinitives and participles present (≠ modern Greek)

Spelling conventions deviating from old orthographic rules (≠ ancient Greek)

The data: from the HTREC challenge

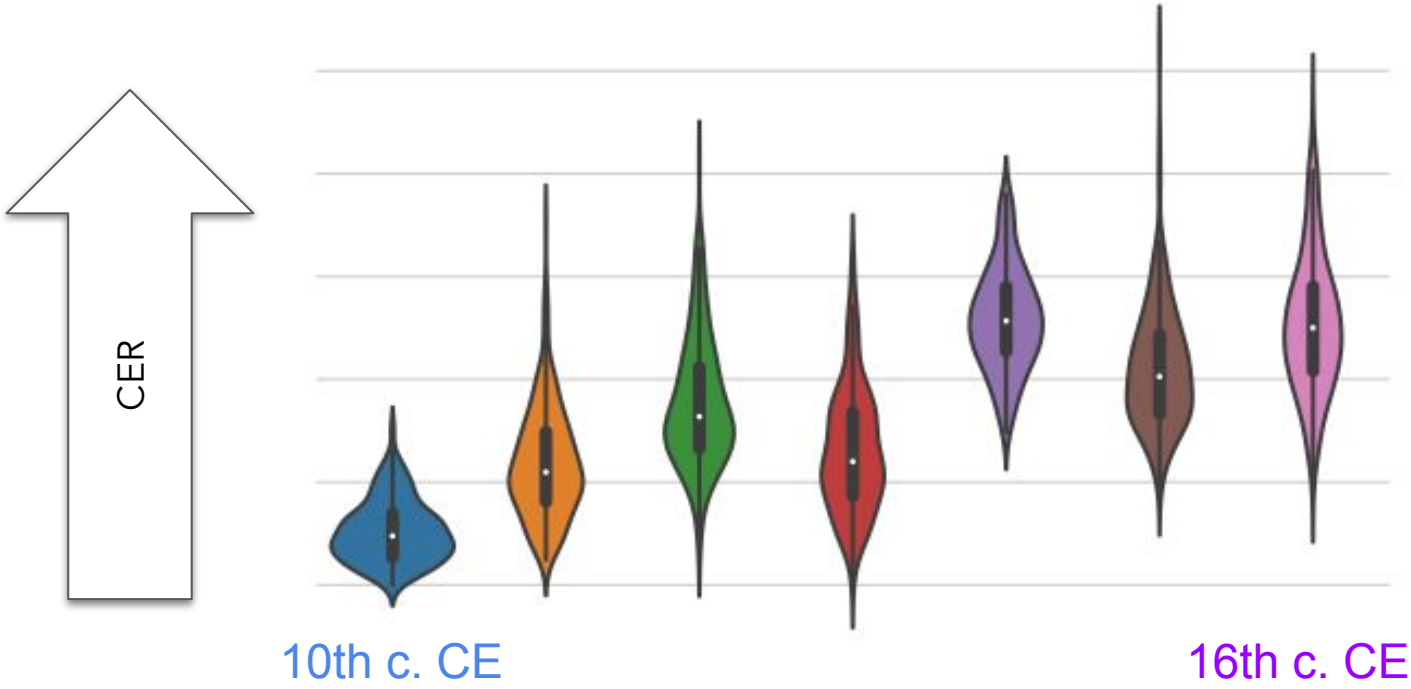
Transcriptions/Recognitions from the HTREC challenge

1,875 lines transcribed and HTRed

HTR model ⇒ Transkribus (under)trained on 7 images (1/cent.)

Transcription	Recognition
ἐγγινομένα πάθη μὴ σβεννύντες ἀλλὰ τῇ 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

The data: and their character error rate



The data: organised for experiments

1,875 Transcriptions ⇒ labelled as flawless (0)

1,875 Recognitions ⇒ labelled as erroneous (1)

Perfectly balanced dataset of 3,744 text lines

80/10/10 train-dev-test split

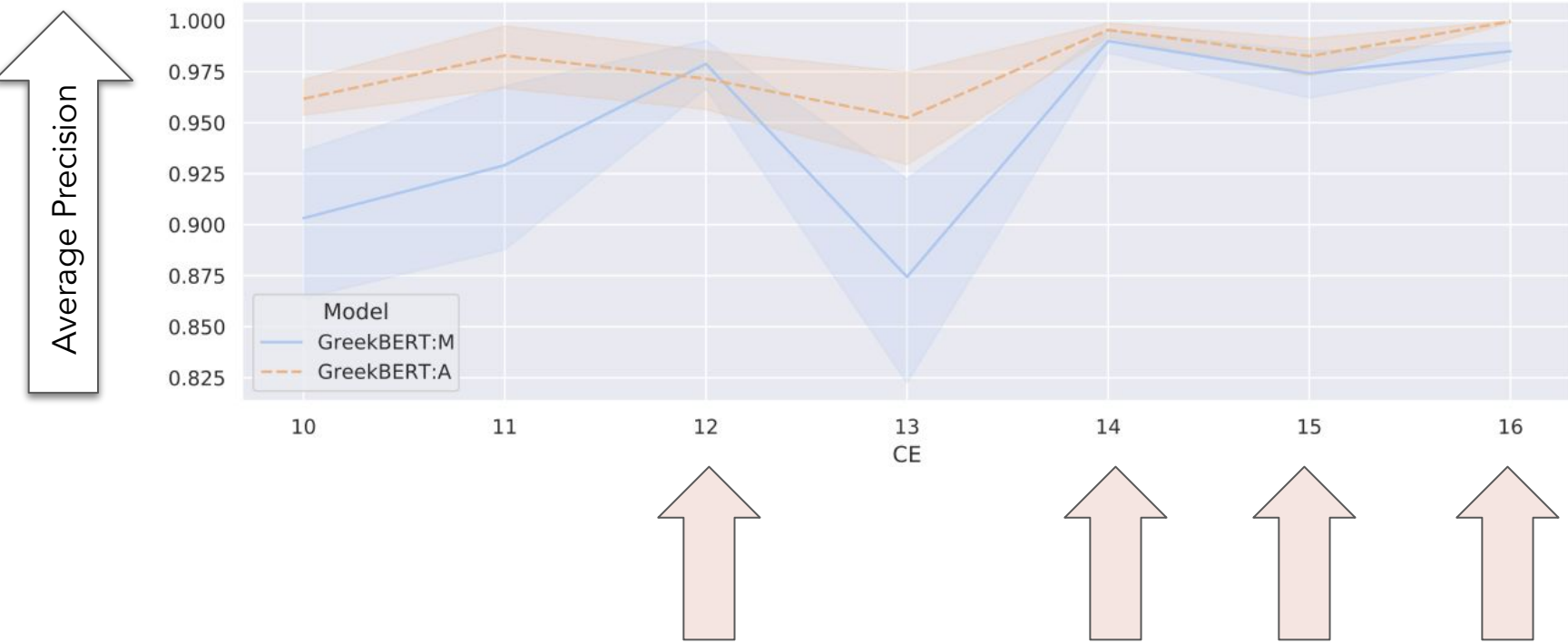
Experiments: machine learning

	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

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**

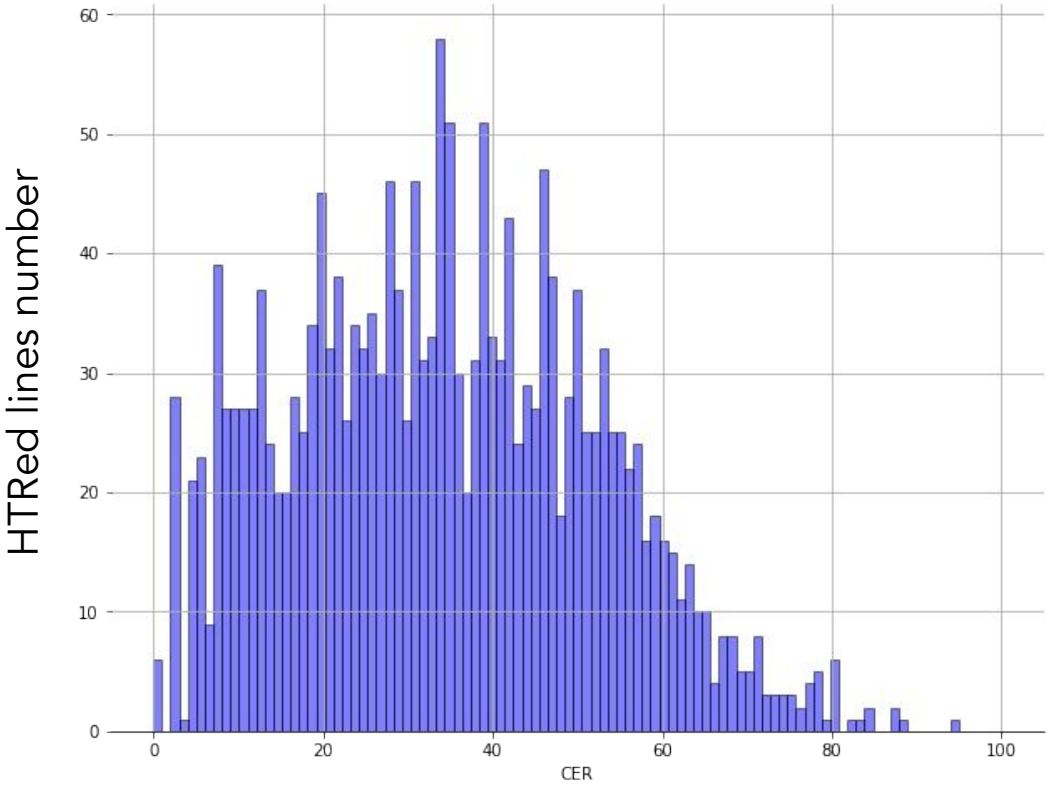
Evaluation per cent.



Evaluation per cent. ⇒ findings

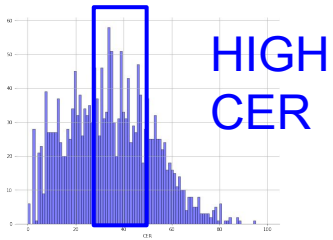
- The classification performance ~ the chronology of the manuscript
- Older manuscripts are benefit from further pre-training on ancient Greek
- Lines from recent manuscripts pose an easier classification challenge

Evaluation per error zone



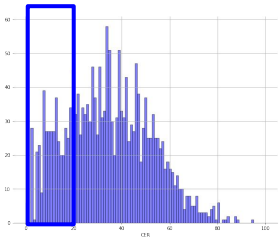
Evaluation per error zone

Evaluation per error zone



Evaluation per error zone

LOW CER

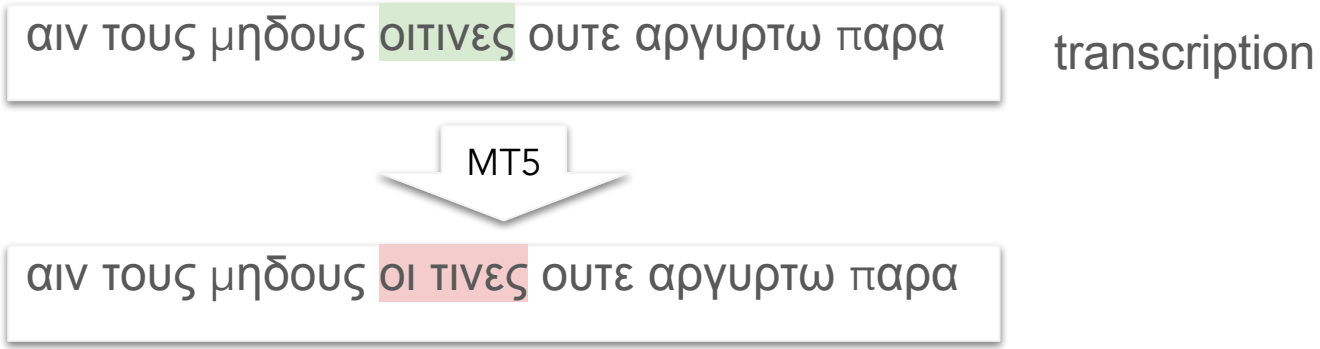


Substantiality assessment

Fine-tuned MT5 for the post-correction task (recognition ⇒ transcription)

Reduces CER in 11% of the (flawed) recognitions 👍 but:

Alters 80% of the (flawless) transcriptions 👎



Error analysis

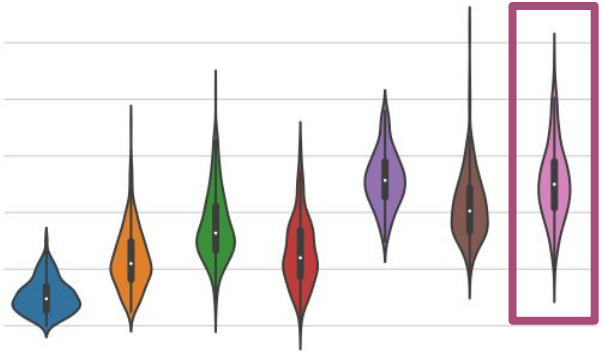
ἐγὼ δ ἀεί πως φιλακόλουθός εἰμι
(I am always readily following)

Error analysis

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

[...] ισμαηλιτι
-κων [...]

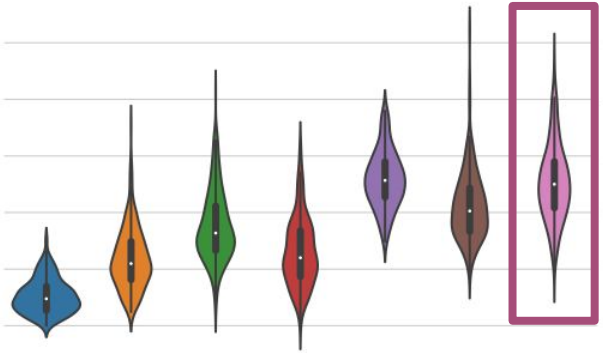
Error analysis: the 16th c. CE



highest avg. CER

‘πλοῦς,’ → HTR → ‘πλουοο’’

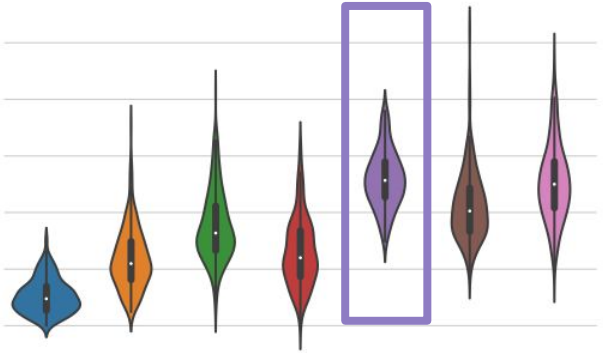
Error analysis: the 16th c. CE



highest avg. CER

‘ῶ πατρὶς’ → HTR → ‘ὤπαρις’

Error analysis: the 14th c. CE

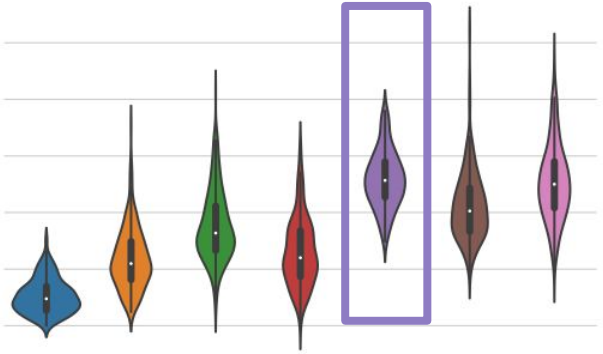


2nd highest

‘φ**έρ** ’ ἐπ ’ αὐτὴν ἴω’ → HTR → ‘φ**ρ**επαυ την ιςω’

The ligature of ‘ερ’ is not captured

Error analysis: the 14th c. CE



2nd highest

omega: horizontal eight-shaped

‘φέρ᾽ ἐπ᾽ αὐτὴν ἴω’ → HTR → ‘φρεπαυ την ιϛω’

Language and script generalisation

- Old, Middle, and Modern English and Japanese
- Common scripts in Greek and Latin manuscripts
- Common handwritten character shapes (E) in Latin scripts

Takeaways

- Detecting erroneous recognised lines ⇒ improves HTR in Byzantine Greek
- Fine-tuned PLMs perform well for the task, by contrast to baselines
- PLM pre-trained on modern and ancient Greek, strong for older centuries
- an application with T5

Future work:

- Larger dataset ⇒ more conclusions
- Calibrated evaluation sets ⇒ to focus on specific error types