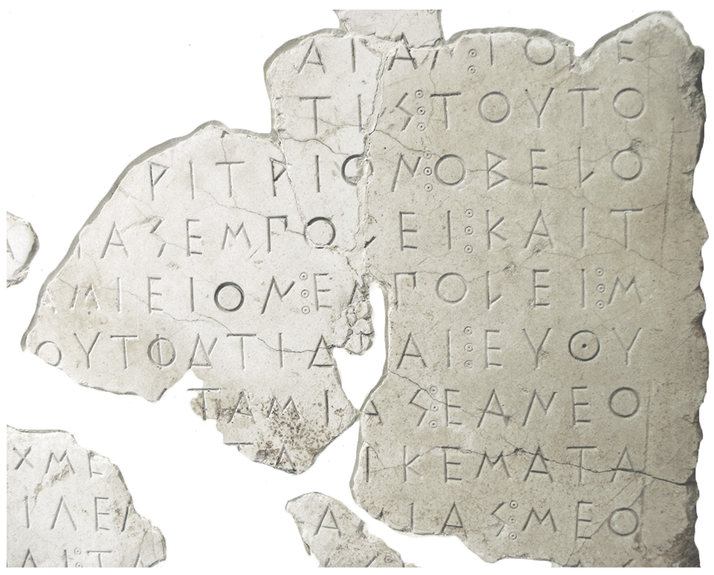
Restoring ancient text using deep learning: a case study on Greek epigraphy

Yannis Assael[1,2,*], Thea Sommerschield[1,*], Jonathan Prag[1]
University of Oxford[1], DeepMind[2]

1. Restoring ancient text

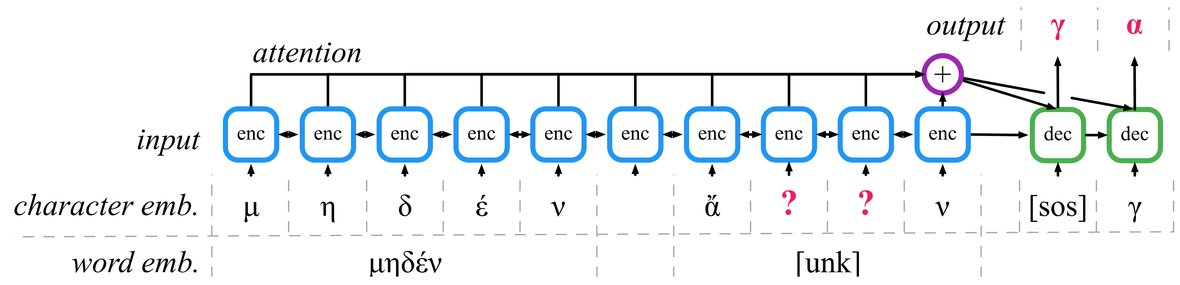
- Ancient History relies on Epigraphy, the study of ancient inscribed texts known as ‘inscriptions’, to reconstruct the thought, society and history of past civilisations.
- Few surviving inscriptions are fully legible and complete, as many have been damaged in time.
- The restoration of these documents is a complex and time consuming task for experts, known as epigraphists.
- Epigraphists find textual and contextual “parallels” within vast repositories to estimate the likelihood of different restoration hypotheses.



Damaged inscription: a decree concerning the Acropolis of Athens (485/4 BCE). *IG* I³ 4B. (CC/BY-SA 3.0, WikiMedia)

2. Pythia

- Pythia is the first ancient text restoration model that recovers missing characters from a damaged text input using deep learning.
- It brings together Ancient History and Deep Learning, providing historians with multiple textual restorations and their confidence level.
- seq2seq with attention using both character and word embeddings.
- *Training:* randomly corrupting the text and using it as ground truth.



Pythia-Bi-Word processing the phrase μηδέν ἄγαν (mēdén ágan) “nothing in excess”, a fabled maxim inscribed on Apollo’s temple in Delphi.

3. Generating the PHI-ML dataset

- The largest digital corpus of Greek inscriptions, PHI Greek Inscriptions, was converted to machine actionable text, which we call PHI-ML.
- Ancient Greek inscriptions date 7th century BCE - 5th CE century.
- *Text cleanup*: compute character frequencies, standardise alphabet, strip human annotations and inconsistencies.
- *Text processing:* ‘-’ for missing characters and ‘?’ for characters to predict; match the number of ‘-’ with those conjectured by epigraphists.
- *Result:* PHI-ML, which consists of more than 3.2 million words. Certain texts were held out as test and validation sets.

Split	Inscriptions	Words	Chars
Train	34,952	2,792k	16,300k
Valid	2,826	211k	1,230k
Test	2,949	223k	1,298k

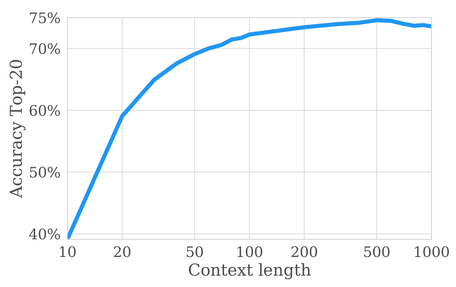
4. Experimental evaluation on PHI-ML

- Uni- and Bi-directional character and word embedding models.
- Pythia-Bi-Word’s predictions achieve a 30.1% character error rate, compared to the 57.3% of the evaluated human epigraphists.
- In 73.5% of the cases the target sequence was in Top-20 hypotheses.

Method	CER	Top-20
Ancient Historian	57.3%	–
LM Philology	68.1%	26.0%
LM Philology & Epigraphy	65.0%	28.8%
LM Epigraphy	52.7%	47.0%
PYTHIA-UNI	42.2%	60.6%
PYTHIA-BI	32.5%	71.1%
PYTHIA-BI-WORD	**30.1%**	**73.5%**

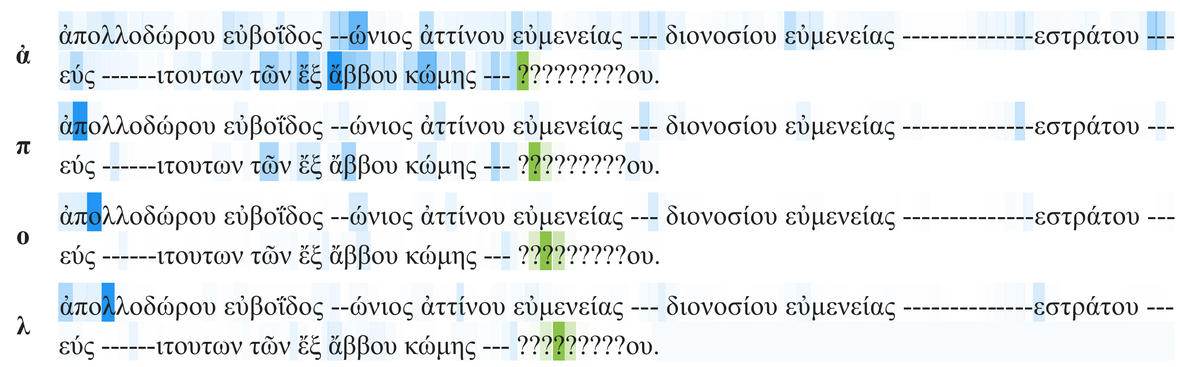
5. Importance of context

- Predictive performance of the Pythia-Bi-Word model under different context lengths.
- The performance peaks around 500 characters of context.



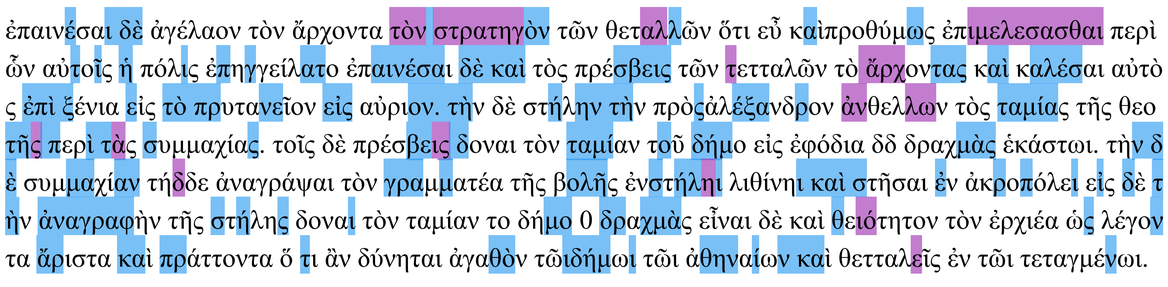
6. Pythia’s attention

- We show receptiveness to context visualising the attention weights.
- The personal name ‘*ἀπολλοδώρ*’ appears twice in the input text: the first 9 characters of the name’s second occurrence are hidden (‘?????????*ου*’).
- Pythia correctly predicts ‘*ἀπολλοδώρ*’ and the attention weights attend at each decoding step to the name’s first occurrence in the text.
- Substituting ‘*ἀπολλοδώρ*’ with a similar-length name ‘*ἀρτεμιδώρ*’ and repeating the experiment, Pythia’s prediction accordingly alters to ‘*ἀρτεμιδώρ*’.



7. Restoring full texts

- *Task:* iteratively apply Pythia to predict an inscription’s missing text and compare Pythia’s predictions with an edition of reference (Rhodes and Osborne, 2003).
- *Result:* correct restorations are in blue and erroneous ones in purple. In nearly all cases, the ground-truth of the erroneous predictions was in Pythia’s Top-20 restoration hypotheses.



8. Conclusions

- Try Pythia online and access PHI-ML:

  https://github.com/sommerschield/ancient-text-restoration/

- Our experimental results illuminate the ways Pythia can assist, guide and advance the ancient historian’s task.
- The combination of Machine Learning and Epigraphy can impact meaningfully the study of inscribed textual cultures, ancient and modern.