

Abstract

The automated decipherment of lost languages has emerged as a critical frontier in digital humanities, yet the epistemological and technical limitations of AI models remain underexplored and misrepresented. This paper proposes to study to what extent can AI language models, such as NLPs trained in low-resource settings, help with the deciphering of lost languages, and what are the actual limitations relative to traditional methods used by linguists?. It argues that, despite general fear of AI replacement, while AI models accelerate decipherment workflows, they cannot operate independently, as they require expert validation and interpretation to generate reliable results, via a thorough analysis of traditional methods compared to AI-assisted models, backed up by successful case studies such as Linear B and Ugaritic. The results demonstrate how these models offer valuable advantages when working in low-data environments. The research deduces that AI is a tool with immense potential, which can be reached only through a human-machine collaboration that regards the ethical concerns of these processes.

The Role of AI in the Decoding of Lost/Ancient Texts

Sara-Alexandra Fluerasu

Introduction

- **The Issue:** Physical Degradation and the scarcity of data (low-resource settings) impede the deciphering of ancient languages
- **Research Question:** To what extent can AI language models (NLP) assist in deciphering lost languages, and what are their actual limitations compared to traditional methods?
- **Hypothesis:** AI cannot replace human expertise, but it significantly accelerates the decipherment workflow through human-machine collaboration

Methods

- **HTR Pipeline:** Overview of the pipeline stages—preprocessing, segmentation, feature extraction, classification, and post-processing.
- **Model Architectures:** Sequence-to-sequence models, Transformers (Self-Attention mechanisms), CNNs, and GANs for synthetic data augmentation.

Results

**Linear B & Ugaritic:** Luo et al.'s neural model achieved 67% accuracy on Linear B and a 5.5% improvement on Ugaritic without parallel data.  
**Etruscan:** Transfer learning from Latin and Ancient Greek consistently improved translation metrics despite sparse datasets.  
**Ithaca Project:** Autonomous restoration achieved 62% accuracy, which increased to 72% when integrated into historians' workflows.

Discussion

- **Data Dependency:** Overcoming low-resource limits through synthetic data generation and noise simulation.
- **Interpretability** ("Black-Box"): Deep learning models yield accurate predictions but lack transparent philological reasoning.
- **Trustworthy AI:** Integrating Explainable AI (XAI) frameworks like LIME and perturbation analysis (POMELO) to build trust and transparency.

Conclusion

**Summary:** AI serves as a powerful quantitative assistant rather than a replacement for linguists, offering major speed and pattern-recognition advantages.  
**Future Directions:** Expanding multilingual models, establishing ethical Open Science standards, and improving decision transparency.

References

[1] Sommerschild, T., Assael, Y., Pavlopoulos, J., et al. (2023). *Machine Learning for Ancient Languages: A Survey*.  
[2] Diao, Y., Xu, L., & Sun, W. (2025). *Ancient Script Image Recognition and Processing: A Review*.  
[3] Chen, D., Shi, F., Agarwal, A., et al. (2024). *Comparing Visual and Textual Representations of Ancient Logographic Writing Systems for NLP*.

