Towards a Platform for AI-Assisted Papyrology

Matthew I. Swindall[1], Graham West[1], James H. Brusuelas[2], Alex C. Williams[3] and John F. Wallin[1]

[1]*Middle Tennessee State University, 1301 East Main St, Murfreesboro, TN 37132, USA*
[2]*University of Kentucky, 410 Administration Dr., Lexington, KY 40506, USA*
[3]*Amazon, AWS AI, 440 Terry Ave N, Seattle, WA 98109, USA*

Abstract

We propose an AI-powered platform to assist experts in transcribing, dating, identifying, and editing ancient manuscripts. In this paper, we discuss our ongoing work on AI-assisted Greek papyrology and our vision for a broader application that is intuitive for scholars of the ancient world. We envision this platform as an all-in-one system for AI-assisted papyrology that can be extended to additional languages and media.

Keywords

Digital Humanities, Machine Learning, Papyrology, Generative AI, Natural Language Processing, Transfer Learning, Handwritten Text Recognition, Blockchain & Smart Contracts

1. Introduction

A great deal of recent inter-disciplinary research has applied state-of-the-art computational methods, such as deep learning models, to the study of ancient texts. Efforts to map this field of research and to define the standards for machine learning on ancient languages, such as Sommerschield *et al.* 2023 [1], are opening doors for more collaboration between the machine learning and digital humanities communities. However, most existing AI tools are not approachable for experts in ancient languages and manuscripts due to the skill set required to utilize them. To contribute to this area of research, we propose our vision for an intuitive, AI-driven platform for analyzing ancient manuscripts, in particular, ancient Greek papyri.

1.1. The Ancient Lives Project

In 2011, a Zooniverse.org collaboration called the Ancient Lives project began crowdsourcing the transcription of papyrus fragments housed at the University of Oxford, such as the one shown in Figure 1. The project resulted in millions of annotations. These highly damaged fragments are challenging for most modern handwritten text recognition (HTR) methods.

Joint Proceedings of the ACM IUI Workshops 2024, March 18-21, 2024, Greenville, South Carolina, USA
mis2n@mtmail.mtsu.edu (M. I. Swindall); graham.west@mtsu.edu (G. West); james.brusuelas@uky.edu (J. H. Brusuelas); acwio@amazon.com (A. C. Williams); john.wallin@mtsu.edu (J. F. Wallin)
https://mis2n.github.io/ (M. I. Swindall); https://www.linkedin.com/in/graham-west-49b75a274/ (G. West); https://mcl.as.uky.edu/users/jbr454 (J. H. Brusuelas); https://www.cs.mtsu.edu/~jwallin/ (J. F. Wallin)
0000-0002-2507-6963 (M. I. Swindall); https://orcid.org/0000-0002-7095-1894 (G. West)

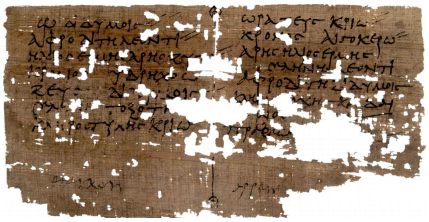

Figure 1: Example of papyrus fragment image used in the Ancient Lives Project.

Figure 2: Examples of character images from the AL-PUB dataset.

1.2. The AL-ALL and AL-PUB Datasets

The AL-ALL dataset, which was derived from the crowdsourced annotations collected during the Ancient Lives project, consists of 419,445 Greek characters, representing all 24 characters of the Greek alphabet, cropped from images of papyrus fragments. Due to ongoing papyrological research, only 205,797 character images from published papyri were made available as the AL-PUB dataset, shown in Figure 2 and available at https://data.cs.mtsu.edu/al-pub/. As demonstrated in Swindall *et al.* 2021 [2], this dataset has been instrumental in the development of deep learning methods for Greek character classification, especially for images of manuscripts that exhibit severe damage and decay.

1.3. Synthetic Characters with GAN's

One of the greatest challenges in crowdsourcing datasets is sampling bias. This was especially true for the AL-ALL and AL-PUB datasets. In Swindall *et al.* 2022 [3], StyleGAN2 was trained on samples from AL-ALL to generate synthetic images of Greek characters on papyrus. The two smallest samples in AL-ALL were doubled by adding these synthetic images. This created the AL-SYNTH dataset, which was used to train new classification models. The new models showed no change in overall accuracy, but demonstrated considerable increases in per-character accuracy for the augmented samples. This work demonstrates the usefulness of synthetically augmenting image datasets to reduce the effects of sampling bias. In addition, synthetic character images may be immensely useful for graphical reconstruction of papyri and stylistic comparisons.

2. Our AI Tools

2.1. Automated Transcription

Utilizing models trained on AL-ALL, several machine learning tools have been developed that form a handwritten text recognition (HTR) pipeline. This pipeline expedites the process of producing a diplomatic transcription, which constitutes an un-edited typescript of the text visible in a given manuscript. The first tool, a character segmentation model, used transfer learning to re-task YOLOv5 with locating characters within papyrus images. The second tool is a character classification model, with a validation accuracy over 94%, which is based on a ResNet architecture used in Swindall *et al.* 2021 [2] and trained on AL-ALL. The third tool

Figure 3: An AI-assisted transcription of Greek text on papyrus. The color-coding (right) denotes line association. Each character is accompanied by a classification probability.

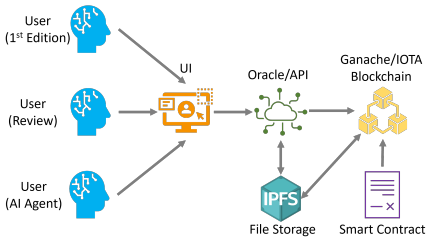

Figure 4: Proposed blockchain and smart contract system for management of digital editions.

is an unsupervised line-sequencing algorithm, which utilizes mean-shift clustering to group characters into lines based on their vertical coordinates.

2.2. Manuscript Dating

A routine task in papyrology is accurately dating manuscripts. In the case of documentary papyri (accounts, letters, leases, etc.), the scribe usually dates the manuscript; though the date is often lost due to damage. Literary papyri (ancient books) never contain a date, unless portions of it were reused for documentary purposes. In the absence of a date, papyrologists must infer it by comparing the handwriting with other dated manuscripts. To automate this process, a pipeline of models was created that can classify a fragment according to classes representing a period of two centuries (i.e., 400 BCE - 201 BCE, 200 BCE - 1 BCE, etc.), with a range of dates spanning from 400 BCE to 600 CE. To create this pipeline, images of documentary papyri with known dates were run through our HTR pipeline, thus obtaining dating information at the level of individual characters. Models were then trained via transfer learning on the ResNet classification model to attribute dates to individual characters. However, due to the high variability of handwriting styles, individual character dates can be unreliable. To address this, a Gaussian Process model was created which assigns a date to an entire fragment based on the predicted dates of its constituent characters. When trained on fragments with 25 or more characters, this model achieves a precision and recall of 75%-80%. Currently, we are investigating possible ways of increasing the temporal resolution without diminishing prediction quality.

3. Future Work

3.1. Natural Language Processing

Digital Epigraphy, which produced digital editions of ancient inscriptions, continues to be a promising area of natural language processing (NLP) research. Efforts such as Pythia [4] and masked language modeling [5] have demonstrated that human-level proficiency is probable for future NLP models. Additional challenges posed by Greek papyri include the lack of word division and punctuation, as well as the physical damage to the fragment resulting in missing characters. To combat these issues, a multi-phase approach may be necessary, including

identifying where characters are missing, predicting how many characters may be missing, and predicting what the missing characters are likely to be.

Beyond textual reconstruction, we believe it may be possible to use computational and deep learning methods for tasks including document identification, provenance, and detection of classification errors for existing digital editions. For example, Williams *et al.* 2014 [6] demonstrated the ability of genetic sequencing algorithms (especially for fragmented texts and texts with a history of textual variation) to compare transcriptions to a corpus of known texts for identification (author, work, etc.). This approach, paired with additional tools, may be invaluable for the AI-assisted study of ancient texts.

3.2. Born-Digital Edition Management with Blockchain & Proteus

With the increasing development of AI tools to assist in ancient manuscript research, it will be necessary to modernize the existing infrastructure for creating and managing born-digital editions of ancient manuscripts. Our Proteus platform, Williams *et al.* 2015[7] and Brusuelas and Meccariello 2023 [8], is a unique environment dedicated to creating, peer-reviewing, and managing born-digital editions of papyri. One of the challenges experienced with Proteus was the complexity of not only managing large volumes of editions using a database system, but also the different scholarly reconstructions (or versions) of the same papyrus fragment.

Currently a solution is in development which will utilize blockchain and smart contract technologies for the management and storage of digital editions. In this proposed system, illustrated in Figure 4, smart contracts are created for new, original editions. Rather than storing all data in complex databases, this smart contract stores only the location of the data on the blockchain itself. Editors of critical editions can submit their edition to the smart contract, which then stores the location of the critical edition's data. The editions themselves can be stored in a number of ways: on a local server where the blockchain is hosted, on a public blockchain, or in a distributed file storage platform such as the InterPlanetary File System (https://ipfs.tech/). Beyond offering a less complex method of edition management, blockchain and smart contracts offer an avenue to a more transparent and decentralized peer-review ecosystem, as discussed in Tenorio-Fornés *et al.* 2021 [9].

4. An AI-Driven Platform for Papyrology

Although we have developed a suite of AI-enabled methods to study the papyrology as a proof of concept application, these tools remain out of reach for many scholars in the field. We envision the creation of a holistic platform which incorporates a host of tools that assist in transcribing, dating, identifying, and editing manuscripts. Figure 3 shows an example transcription. Our approach is likely transferable to other kinds of manuscripts and languages. Instead of a platform limited to Greek papyrology, we envision one that can be interoperable with other language and manuscript datasets from the ancient world.

References

[1] T. Sommerschield, Y. Assael, J. Pavlopoulos, V. Stefanak, A. Senior, C. Dyer, J. Bodel, J. Prag, I. Androutsopoulos, N. de Freitas, Machine Learning for Ancient Languages: A Survey, Computational Linguistics 49 (2023) 703–747. URL: https://doi.org/10.1162/coli_a_00481. doi:10.1162/coli_a_00481.

[2] M. I. Swindall, G. Croisdale, C. C. Hunter, B. Keener, A. C. Williams, J. H. Brusuelas, N. Krevans, M. Sellew, L. Fortson, J. F. Wallin, Exploring learning approaches for ancient greek character recognition with citizen science data, in: 2021 17th International Conference on eScience (eScience), IEEE, 2021, pp. 128–137.

[3] M. Swindall, T. Player, B. Keener, A. Williams, J. Brusuelas, F. Nicolardi, M. D'Angelo, C. Vergara, M. McOsker, J. Wallin, Dataset augmentation in papyrology with generative models: A study of synthetic ancient greek character images, 2022, pp. 4948–4954. doi:10.24963/ijcai.2022/687.

[4] Y. Assael, T. Sommerschield, J. Prag, Restoring ancient text using deep learning: a case study on Greek epigraphy, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6368–6375. URL: https://aclanthology.org/D19-1668. doi:10.18653/v1/D19-1668.

[5] K. Lazar, B. Saret, A. Yehudai, W. Horowitz, N. Wasserman, G. Stanovsky, Filling the gaps in Ancient Akkadian texts: A masked language modelling approach, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 4682–4691. URL: https://aclanthology.org/2021.emnlp-main.384. doi:10.18653/v1/2021.emnlp-main.384.

[6] A. C. Williams, H. D. Carroll, J. F. Wallin, J. Brusuelas, L. Fortson, A.-F. Lamblin, H. Yu, Identification of ancient greek papyrus fragments using genetic sequence alignment algorithms, in: 2014 IEEE 10th International Conference on e-Science, volume 2, 2014, pp. 5–10. doi:10.1109/eScience.2014.14.

[7] A. C. Williams, A. Santarsiero, C. Meccariello, G. Verhasselt, H. D. Carroll, J. F. Wallin, D. Obbink, J. H. Brusuelas, Proteus: A platform for born digital critical editions of literary and subliterary papyri, in: 2015 Digital Heritage, volume 2, 2015, pp. 453–456. doi:10.1109/DigitalHeritage.2015.7419546.

[8] M. C. Brusuelas, J. H., Proteus: A platform for born-digital, critical editions of literary and subliterary papyri, Textual History of the Bible, Volume 3D: A Companion to Textual Criticism, Brill, 507-512. (2023).

[9] Ámbar Tenorio-Fornés, E. P. Tirador, A. A. Sánchez-Ruiz, S. Hassan, Decentralizing science: Towards an interoperable open peer review ecosystem using blockchain, Information Processing Management 58 (2021) 102724. URL: https://www.sciencedirect.com/science/article/pii/S0306457321002089. doi:https://doi.org/10.1016/j.ipm.2021.102724.