A new machine-actionable corpus for ancient text restoration

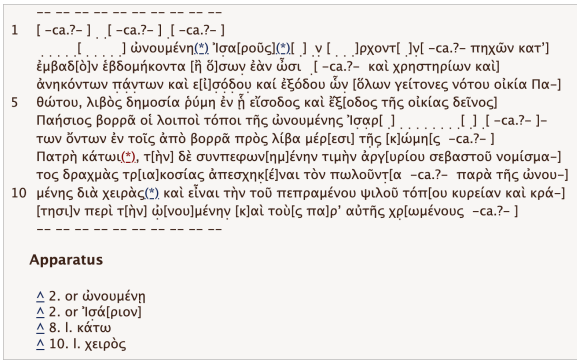
Will Fitzgerald and **Justin Barney**
Western Michigan University
{william.a.fitzgerald, justin.barney}@wmich.edu

Abstract

The Machine-Actionable Ancient Text (MAAT) Corpus is a new resource providing training and evaluation data for restoring lacunae in ancient Greek, Latin, and Coptic texts. Current text restoration systems require large amounts of data for training and task-relevant means for evaluation. The MAAT Corpus addresses this need by converting texts available in EpiDoc XML format into a machine-actionable format that preserves the most textually salient aspects needed for machine learning: the text itself, lacunae, and textual restorations. Structured test cases are generated from the corpus that align with the actual text restoration task performed by papyrologists and epigraphist, enabling more realistic evaluation than the synthetic tasks used previously. The initial 1.0 beta release contains approximately 134,000 text editions, 178,000 text blocks, and 750,000 individual restorations, with Greek and Latin predominating. This corpus aims to facilitate the development of computational methods to assist scholars in accurately restoring ancient texts.

1 Introduction

For the papyrologist and epigraphist, a fundamental task is the creation of an accurate transcription of the text under consideration. Often the physical medium supporting the text has undergone decay, leaving gaps, or "lacunae," in the text. Filling these gaps is a painstaking task. Kleve and Fonnes (1981) first recognized the potential of computer science for assisting with text restorations of this type, specifically by leveraging string-searching algorithms. Advances in computational approaches to text analysis, especially deep learning and large language models, may be able to aid scholars in the task of textual restoration. Developing such systems typically requires large amounts of data, both for training, and ideally for providing task-relevant means for evaluation.


-- -- -- -- -- -- -- -- -- --
1 [-ca.?-] . . [-ca.?-] . [-ca.?-]
. . . . [. . . .] ὠνουμένη(*) Ἰσα[ροῦς](*)[.] . ν [. . .]ρχοντ[.]ν[-ca.?- πηχῶν κατ']
ἐμβαδ[ὸ]ν ἑβδομήκοντα [ἣ ὅ]σων ἐὰν ὦσι . . [-ca.?- καὶ χρηστηρίων καὶ]
ἀνηκόντων πάντων καὶ ε[ἰ]σόδου καί ἐξόδου ὧν [ὅλων γείτονες νότου οἰκία Πα-]
5 θώτου, λιβὸς δημοσία ῥύμη ἐν ᾗ εἴσοδος καὶ ἔξ[οδος τῆς οἰκίας δεῖνος]
Παήσιος βορρᾶ οἱ λοιποὶ τόποι τῆς ὠνουμένης Ἰσαρ[.] . . . . . [.] . [-ca.?-]-
των ὄντων ἐν τοῖς ἀπὸ βορρᾶ πρὸς λίβα μέρ[εσι] τῆς [κ]ώμη[ς -ca.?-]
Πατρὴ κάτωι(*), τ[ἠν] δὲ συνπεφων[ημ]ένην τιμὴν ἀργ[υρίου σεβαστοῦ νομίσμα-]
τος δραχμὰς τρ[ια]κοσίας ἀπεσχηκ[έ]ναι τὸν πωλοῦντ[α -ca.?- παρὰ τῆς ὠνου-]
10 μένης διὰ χειρὸς(*) καὶ εἶναι τὴν τοῦ πεπραμένου ψιλοῦ τόπ[ου κυρείαν καὶ κρά-]
[τησι]ν περὶ τ[ἠν] ὠ[νου]μένην [κ]αὶ τοὺ[ς πα]ρ' αὐτῆς χρ[ωμένους -ca.?-]
-- -- -- -- -- -- -- -- -- --

Apparatus

^ 2. or ὠνουμένη
^ 2. or Ἰσά[ριον]
^ 8. l. κάτω
^ 10. l. χειρὸς

Figure 1: Leiden Transcription of P.Flor. 3 324, from Aegyptus.89.240, 2011.

Here we introduce the 1.0 beta version of the Machine-Actionable Ancient Text Corpus (MAAT Corpus), which provides training and evaluation data for the development of machine learning models that aid in the restoration of ancient Greek, Latin, and Coptic texts.

2 Current text restoration corpora

There are several different corpora used in creating systems for text restoration of ancient text. Two existing systems, Pythia (Assael et al., 2019) and its successor Ithaca (Assael et al., 2022) use Greek inscription data from the Packard Humanities Institute (Packard Humanities Institute, 2023) that have been converted to a modified Leiden Convention (Wilcken, 1932) format. Papavassiliou et al. 2020 created a corpus of Mycenaean Linear B texts for the restoration of Linear B tablets. Background large-language models have been trained on corpora as well, such as Latin BERT (Bamman & Burns, 2020) and AristoBERTo (Myerston, 2022), GreBerta (Riemenschneider & Frank, 2023).

Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), pages 56–60
August 15, 2024 ©2024 Association for Computational Linguistics

```
{
   "corpus_id": "EDH",
   "file_id": "HD056774",
   "block_index": 1,
   "id": "EDH/HD056774/1",
   "title": "Epitaph from Municipium Claudium Virunum, bei – S. Andrä/Lavanttal
(Noricum)",
   "material": "gesteine",
   "language": "la",
   "training_text": " Ursuius vius sibi \nfecit et <gap/>\niurae uxo[ri]",
   "test_cases": [
     {
       "case_index": 1,
       "id": "EDH/HD056774/1/1",
       "test_case": " Ursuius vius sibi \nfecit et <gap />\niurae uxo[..]",
       "alternatives": [
         "ri"
       ]
     }
   ]
}
```


Figure 2: Example JSON representation of a single **ab** block with one test case; \n reflects a **lb** element.

3 Corpora of interest

Papyrologists and epigraphists have generally agreed upon using a specialized schema developed originally for epigraphy, EpiDoc (Elliott et al., 2006), based on the TEI format (*TEI*, 1994). The largest corpus of epigraphy stored in EpiDoc format is maintained by the Epigraphic Database Heidelberg (*Epigraphic Database Heidelberg*, 1993), which focuses primarily on Latin inscriptions from the Roman Empire. The largest corpus of papyrological texts is Papyri.info, a collaboration among several institutions that hosts papyrological data in Greek, Latin, Coptic, and Arabic (*Papyri.Info*, 2007).

The EpiDoc format provides extensive capabilities for describing metadata for inscriptions and papyri. It also has an XML-structured format as an alternative to the Leiden Conventions. Texts are described in **ab** blocks (originally standing for "anonymous block") and provide a richer description language for text editions than the Leiden Conventions. Because the Leiden Conventions format is more compact, we will use this format for examples printed in this paper.

4 Features of MAAT corpus

Unfortunately, for many machine learning and large language models, the structure of the **ab** blocks is too rich, since it provides internal structure for annotations, stylistic information and so on (the Leiden Conventions also communicate some of these features). With respect to building systems for text restoration, a simpler system is required. As Assael et al. 2022 note, these corpora need to be "machine-actionable." For this reason, they ought to be easy to feed into machine learning systems for learning and for evaluation.

Figure 1 shows the text from a typical edition (P.Flor 3 324) from Papyri.info, a contract for the sale of property (*Aegyptus.89.240*, 2011). For this paper, three things should be noted. First, text restorations are provided in square brackets. For example, in line three, the brackets in the phrase [ἢ ὅ]σων indicate that "ἢ ὅ" has been supplied by the papyrologist and that the letter forms are not visible on the papyrus itself. Second, missing text that the editor has not restored is indicated by dots. One dot corresponds to one missing letter; therefore, the number of dots signifies the approximate number of letters known to be missing. The marking "-ca.?" or "- - -" indicates a gap of unknown extent. Third, alternate restorations of the text are sometimes given in the *apparatus criticus*. These alternate readings represent viable textual conjectures, which were not ultimately chosen by the editor as their preferred reading. While digital editions print alternative restorations less commonly than print editions, they are sometimes encoded in the XML

Corpus	Editions	Blocks	Restorations
DCLP (Digital Corpus of Literary Papyri in EpiDoc XML)	1,938	11,581	129,806
DDbDP (Duke Databank of Documentary Papyri)	59,693	85,626	507,985
EDH (Epigraphic Database Heidelberg)	72,353	80,753	113,944
Totals	133,984	177,960	751,735

Table 1: Counts of Editions, Blocks, and Restorations from the corpora represented in the Machine-Actionable Ancient Text Corpus

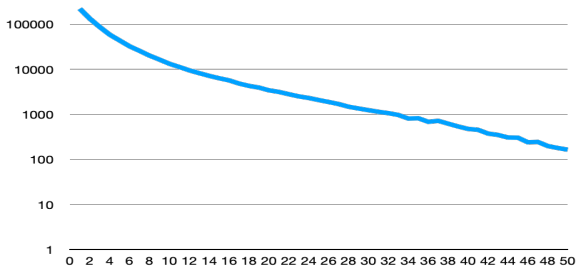


Figure 3: Distribution of gap lengths of text restorations in the MAAT Corpus, logarithmic scale

data. In our sample text from Figure 1, two apparatus notes appear for line two of the transcription.

To make a corpus machine-actionable for learning, especially for large language models, we stripped away all but the most textually salient aspects of the text, using Unicode UTF-8 encoding. Our corpus includes the preserved text, as well as unclear letters and restorations. Although typographical conventions such as casing, interlinear word space, punctuation, accents, breathing marks, and other diacritics are typically not found on the source material, such typography is retained. Line breaks (indicated by the **lb** element in EpiDoc XML) are also preserved. Unclear text, indicated by Leiden Conventions with a sublinear dot, is treated no differently than preserved text; text that has been restored by an editor is bracketed. For example, the text "καὶ ε[ἰ]ς̣ό̣δου̣" converted to "καὶ ε[ἰ]σόδου." Occasionally (as in Figure 1) there are alternative readings of a restored text, but since alternative readings are difficult to process, the first primary text restoration is chosen. Abbreviations, especially prevalent in Latin inscriptions, are not expanded.

Gaps in the text that have not been restored by an editor must also be indicated. There are, essentially, three types of gaps: gaps of known length, gaps of approximately known length, and gaps of unknown length. Gaps of known length are converted to a dot for each missing letter. Similarly, gaps of approximate length are treated as if the gap length is known. The EpiDoc XML tag **<gap/>** is used for gaps of unknown length. Gaps are sometimes indicated *within* a restored text, and such gaps are moved outside. For example, the text "τὸν πωλοῦν̣τ[α -ca.?- παρὰ]" is converted to "τὸν πωλοῦντ[α]<gap/>[παρὰ]".

In the end, all texts in the MAAT corpus are written in a simplified format for easier use by machine learning models. Texts from the **ab** blocks in EpiDoc XML format are converted to a light, Leiden-like format, but with a bare minimum of annotative markings: text and gaps of known and unknown length, with restored text in brackets. Figure 2 provides our data for a 1st-2nd century CE epitaph from the Roman province Noricum (*EDH HD056774*, 2014).

Typically, in machine learning tasks, a portion of a training corpus is set aside for evaluation. In the most successful system to date for inscription restoration, Ithaca (Assael et al., 2022), one to ten characters are artificially hidden during the testing phase, and the machine (or parallel human evaluator) is tasked with restoring these artificial lacunae. A similar text-masking evaluation method is used in Papavassileiou et al., 2023 for Mycenaean Linear B tablets, although they also ask the model to perform text restoration of some real lacunae.

The large number of restorations created by papyrologists and epigraphists found in the base corpora of Greek, Latin, and Coptic texts provide a rich opportunity to create evaluation data that are aligned with the actual text restoration task. Although it may be useful to train a system using artificial lacunae, it is more valuable to evaluate on the text restorations done by working papyrologists and epigraphists. These practitioners do not work with random lacunae, since lacunae *in situ* are not random: they follow a logarithmic distribution in length (see Figure 3), and tend to occur in certain locations. The immediate textual context of real lacunae also tends to be much deteriorated and uncertain, in comparison to the sites of artificial lacunae.

To that end, we can create test cases by using the actual lacunae and text restorations that are present in papyrological and epigraphic sources and use the (retained) training data with the restored text for evaluation. Because there are possible alternative readings for a restored gap, though, it is better to have a structured test case that retains those readings. This will slightly complicate the evaluation metrics. Rather than using, for example, character error count for a single restoration, we need to use the minimum character error count for a (possibly singleton) set of alternatives. Similarly, calculating the top-*n* rate will need to consider the presence of the proposed restoration in the set of alternatives.

Thus, a single test case needs a little more structure, containing at least the text with a gap to be filled, plus its alternatives. For example, for the text "ὠνουμένη Ἰσα[....]" the two alternative readings "ροῦς" and "ριον" are required. Note that, because letter forms of different types take up different amounts of space on the material substrate (and therefore calculations of the number of missing letters are approximate), alternatives might, in fact, have different character lengths. In these cases, the mask to be restored will comprise the mode of alternative lengths.

5 Format and distribution of data

Data in the MAAT Corpus is structured as a set of JavaScript object notation (JSON) records (Bray, 2014), one record for each **ab** block. Each record contains metadata about the block (an id field, source corpus, source file id, block index within the file, material, and language). It also has the training text, as described above. For each restored text, a test case is created, also containing an id, test case index within the text, the test case itself, and the set of alternatives. For statistical purposes, the number of alternatives, the number, mode, maximum, and minimum lengths of the alternatives are also described.

Currently, there are approximately 134,000 editions processed in the MAAT Corpus, representing approximately 178,000 **ab** blocks and 750,000 individual text restorations.

There is a small representation of Coptic texts in the MAAT Corpus (around 1% of the total, mostly papyri). Latin editions outnumber Greek editions (54% and 45%, respectively). Papyrological texts tend to be longer than inscriptions; papyrological texts tend to be written in Greek and inscriptions in Latin, so the number of Greek blocks is greater than the number of Latin ones (53% and 46%, respectively). The number of text restorations in Greek greatly outnumber Latin ones (83% and 16%, respectively).

The gap lengths of restored text created by papyrologists and epigraphists found in the MAAT Corpus vary widely, and follow an unsurprising logarithmic or Zipfian distribution. Gaps of length 1 (that is, one character) account for 30% of all gaps, and gaps of length 4 or less account for 67%. Gaps of length 10 or less account for 87% of all gaps. Figure 3 shows the distribution.

6 Data availability and next steps

We are now releasing the Machine-Actionable Ancient Text Corpus in a beta state at https://zenodo.org/records/12518435 (Fitzgerald & Barney, 2024). The corpus is not meant to compete with current systems, such as Papyri.info and EDH, whose use cases are different. Instead, we hope that the MAAT Corpus will aid the creation of software systems that can help working papyrologists and epigraphists accurately and efficiently hypothesize text restorations in new editions of current and newly recovered texts and inscriptions. Code for creating the corpus can be found at https://github.com/WMU-Herculaneum-Project/maat.

We welcome the collaboration of other scholars and institutions in the service of adding additional data to the MAAT corpus, including data from other ancient languages. Our specific interest is in text restoration of Greek papyrological texts, but we would like to expand this to Arabic and other non-western texts as well. Given the similarities of the text restoration task and its evaluation methodology among texts of different language traditions, such expansions promise to be fruitful.

In the future, we also intend to create a pathway by which any data made available in DSL-based formats (Del Grosso et al., 2023; Williams et al., 2015) can be converted for inclusion in future versions of the corpus.

7 Conclusion

This paper introduces and announces the publication of the MAAT Corpus, which provides an easily accessible, versioned corpus of machine-actionable ancient texts that can be used in machine learning. It also makes available evaluation data, via its test cases, that closely track the task of text restoration

as done by working papyrologists and epigraphists. The MAAT Corpus currently includes approximately 60 Mb of ancient text, making it the largest corpus available for evaluating text restoration tasks. It is also the only dataset that uses actual lacunae and text restorations as test cases for evaluation.

Acknowledgements

The authors wish to thank the reviewers for their valuable comments on an earlier draft of this paper.

References

Aegyptus.89.240 = HGV P.Flor. 3 324 = Trismegistos 25457 = p.flor.3.324. (2011). Duke Databank of Documentary Papyri. http://papyri.info/ddbdp/aegyptus;89;240

Assael, Y., Sommerschield, T., & Prag, J. (2019). Restoring ancient text using deep learning: A case study on Greek epigraphy. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6367–6374. https://doi.org/10.18653/v1/D19-1668

Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., & De Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. *Nature*, *603*(7900), 280–283. https://doi.org/10.1038/s41586-022-04448-z

Bamman, D., & Burns, P. J. (2020). *Latin BERT: A Contextual Language Model for Classical Philology*. https://doi.org/10.48550/ARXIV.2009.10053

Bray, T. (2014). *The JavaScript object notation (JSON) data interchange format* (7159). RFC Editor. https://doi.org/10.17487/RFC7159

Del Grosso, A. M., Zenzaro, S., Boschetti, F., & Ranocchia, G. (2023). GreekSchools: Making Traditional Papyrology Machine Actionable through Domain-Driven Design. *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, 621–626. https://doi.org/10.1109/CiSt56084.2023.10409929

EDH HD056774. (2014, February 18). Epigraphic Database Heidelberg. https://edh.ub.uni-heidelberg.de/edh/inschrift/HD056774

Elliott, T., Bodard, G., & Cayless, H. (2006, 2024). *EpiDoc: Epigraphic Documents in TEI XML*. https://epidoc.stoa.org/

Epigraphic Database Heidelberg. (1993, 2021). https://edh.ub.uni-heidelberg.de/

Fitzgerald, W., & Barney, J. (2024). *The Machine-Actionable Ancient Text (MAAT) Corpus, Beta 1 (1.0.0)* . (MAAT Corpus) [dataset]. https://doi.org/10.5281/zenodo.12518435

Kleve, K., & Fonnes, I. (1981). Lacunology: On the use of computer methods in Papyrology. *Symbolae Osloenses*, *56*(1), 157–170. https://doi.org/10.1080/00397678108590755

Myerston, J. (2022, January 27). *aristoBERTo*. https://huggingface.co/Jacobo/aristoBERTo/blob/main/README.md

Packard Humanities Institute. (2023, July 12). *PHI Greek Inscriptions*. https://inscriptions.packhum.org/

Papavassileiou, K., Kosmopoulos, D. I., & Owens, G. (2023). A Generative Model for the Mycenaean Linear B Script and Its Application in Infilling Text from Ancient Tablets. *Journal on Computing and Cultural Heritage*, *16*(3), 1–25. https://doi.org/10.1145/3593431

Papyri.info. (2007, 2024). https://papyri.info/docs/about

Riemenschneider, F., & Frank, A. (2023). Exploring Large Language Models for Classical Philology. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15181–15199. https://doi.org/10.18653/v1/2023.acl-long.846

TEI: Text Encoding Initiative. (1994, 2024). https://tei-c.org/

Wilcken, U. (1932). *Das Leydener Klammersystem*. B.G. Teubner Verlagsgesellschaft.

Williams, A. C., Santarsiero, A., Meccariello, C., Verhasselt, G., Carroll, H. D., Wallin, J. F., Obbink, D., & Brusuelas, J. H. (2015). Proteus: A platform for born digital critical editions of literary and subliterary papyri. *2015 Digital Heritage*, 453–456. https://doi.org/10.1109/DigitalHeritage.2015.7419546