

# Machine Learning for Ancient Languages: A Survey

Thea Sommerschild\*

Ca' Foscari University of Venice

Department of Humanities

thea.sommerschild@unive.it

Yannis Assael\*

Google DeepMind

yannisassael@google.com

John Pavlopoulos\*

Athens University of

Economics and Business

annis@aueb.gr

Vanessa Stefanak

Google DeepMind

Andrew Senior

Google DeepMind

andrewsenior@google.com

Chris Dyer

Google DeepMind

cdyer@google.com

John Bodel

Brown University

Classics Faculty

John.Bodel@brown.edu

Jonathan Prag

University of Oxford

Faculty of Classics

jonathan.prag@merton.ox.ac.uk

---

\* Thea Sommerschild, Yannis Assael, and John Pavlopoulos contributed equally to this work. To whom correspondence should be addressed. E-mail: yannisassael@google.com, thea.sommerschild@unive.it, annis@aueb.gr

Action Editor: Nianwen Xue. Submission received: 17 January 2023; revised version received: 20 March 2023; accepted for publication: 16 April 2023.

<https://doi.org/10.1162/coli.a.00481>

Ion Androutsopoulos

Athens University of  
Economics and Business  
Department of Informatics  
ion@aueb.gr

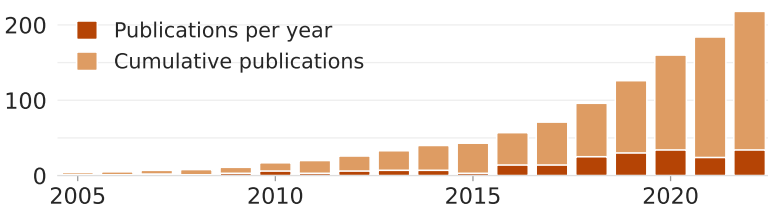
Nando de Freitas

Google DeepMind  
nandodefrees@google.com

*Ancient languages preserve the cultures and histories of the past. However, their study is fraught with difficulties, and experts must tackle a range of challenging text-based tasks, from deciphering lost languages to restoring damaged inscriptions, to determining the authorship of works of literature. Technological aids have long supported the study of ancient texts, but in recent years advances in artificial intelligence and machine learning have enabled analyses on a scale and in a detail that are reshaping the field of humanities, similarly to how microscopes and telescopes have contributed to the realm of science. This article aims to provide a comprehensive survey of published research using machine learning for the study of ancient texts written in any language, script, and medium, spanning over three and a half millennia of civilizations around the ancient world. To analyze the relevant literature, we introduce a taxonomy of tasks inspired by the steps involved in the study of ancient documents: digitization, restoration, attribution, linguistic analysis, textual criticism, translation, and decipherment. This work offers three major contributions: first, mapping the interdisciplinary field carved out by the synergy between the humanities and machine learning; second, highlighting how active collaboration between specialists from both fields is key to producing impactful and compelling scholarship; third, highlighting promising directions for future work in this field. Thus, this work promotes and supports the continued collaborative impetus between the humanities and machine learning.*

## 1. Introduction

Ancient languages are key conveyors and repositories of ancient civilizations, as they encode the thought, cultures, and histories of the past. The texts that preserve these languages were written over the centuries on a variety of media (bone, metal, palm leaf, paper, papyri, parchment, potsherds, stone) and in different scripts (Brahmi, Old Chinese, Egyptian hieroglyphs, ancient Greek, Indus, Latin, Mayan, and others). Over the last twenty years, the introduction of advanced technologies has spurred transformational advances in the study of ancient languages and texts, with the rise of machine learning leaving a particular mark. Machine learning models can discover and harness intricate statistical patterns in vast quantities of data. Recent increases in computational power and advances in deep neural network models, a sub-area of machine learning known as deep learning, have enabled these models to tackle challenges of growing sophistication in several fields (LeCun, Bengio, and Hinton 2015) (see Section 2), including the study of ancient languages (Parker et al. 2019; Kang et al. 2021; Assael et al. 2022; Yoo et al. 2022). The patterns discovered by these models can be leveraged to advance the state-of-the-art in tasks ranging from character recognition to stylometrics, from author attribution



**Figure 1**  
Bar chart depicting the number of articles published on the topic of machine learning for ancient languages per year. As can be seen, the last 5 years have seen a substantial increase in the number of publications.

to textual restoration. Similarly to how microscopes and telescopes have contributed to the realm of science, the humanities can now be assisted by machine learning methods and techniques.

The steep increase in scholarly efforts in this field can be connected to the wider availability of digitized datasets, comprising publicly accessible high-quality photographs or transcriptions of ancient texts. In parallel, the field of machine learning is constantly being advanced by novel learning methods and architectures, deep learning being one of the most recent examples. This fecund situation has engendered a “virtuous circle” of sorts, whereby greater availability of data and fast-paced scientific progress have inspired more people to both digitize ancient texts and explore novel machine learning methods to study them, fuelling a dynamic feedback loop (competitions are a clear example of this trend). As a direct consequence, new interdisciplinary research questions are being posed—the growing number of publications per year as surveyed below and seen in Figure 1 demonstrates this momentum.

In a machine learning setting, ancient languages exhibit several differences from their modern counterparts. For ancient languages, often very limited data is available today, owing to low survival record, damaged state of material preservation, or complex transmission traditions. Moreover, only a small fraction of the extant ancient data has been digitized in a standardized, open-access and metadata-rich format, which is crucial to machine learning tasks. Ancient languages were written in a variety of writings systems, some of which are now extinct (e.g., Mayan hieroglyphs) or still undeciphered (e.g., the Indus script). What is more, a single script might encode several languages, but the relationship between these languages might be unclear. Even within the same script or language, variations between genres, written supports, and text types (e.g., epigraphic and literary texts) and geographically or chronologically specific variants (e.g., local dialects) make the generalization of machine learning methods complex. Language-specific idiosyncrasies (e.g., Latin abbreviations), complex textual transmission histories (e.g., disputed authorship), and semantic shifts between ancient and modern languages also add a layer of sophistication to an already complex endeavour. Finally, the lack of “ground truths” concerning, for instance, the restoration of textual lacunae or the dating of a text also makes evaluating a model’s performance extremely difficult. But it is this very complexity that makes the study of machine learning for ancient languages a worthwhile research topic and an interesting benchmark.

In this survey, we offer a detailed review of scholarly contributions to this field. We have chosen to discuss scholarship focusing on *ancient texts written in ancient languages*. More specifically, with “ancient languages” we consider all languages in use across the world, written on any medium and in any script, between the birth of writing systems

in ancient Mesopotamia (3400 BCE) up until the conventional end of “ancient history” in the late first millennium CE. Owing to space limitations, this article exclusively considers those works dealing with interdisciplinary machine learning research for ancient languages: works that did not use machine learning models, as well as those published before the 2000s, were excluded. Compared to previous literature reviews (Stamatatos 2009; Dexter et al. 2017; Papantoniou and Tzitzikas 2020; Mantovan and Nanni 2020; Fiorucci et al. 2020; Bhurke et al. 2020; Narang, Jindal, and Kumar 2020; Sahala 2021; Bogacz and Mara 2022; Faigenbaum-Golovin, Shaus, and Sober 2022), which are either task- or language-specific (e.g., artificial intelligence [AI] for the Greek language, digital Assyriology, handwritten text recognition, AI for archaeology), this work strives to encompass all tasks and all ancient languages benefiting from the synergistic collaboration between machine learning and the humanities.

Our goal is twofold: First, we seek to map this interdisciplinary field to aid future research; second, we aim to highlight how the joint, collaborative effort of specialists in both the sciences and the humanities is key to producing relevant, robust, and cogent scholarship. Our target audience therefore comprises humanities researchers (historians, classicists, linguists, philologists), reviewing the plethora of machine learning methods available to tackle ancient textual challenges; and computer scientists, examining the many idiosyncrasies of ancient writing systems. Our final aim is to promote and support an increased collaborative impetus between the humanities and AI (Palaniappan and Adhikari 2017; Popović, Dhali, and Schomaker 2021; Assael et al. 2022).

To best undertake our review work, we designed a taxonomy (Figure 2) following the different, but not necessarily sequential, steps involved in the study of ancient documents:

- **Digitization:** bringing textual sources to a high-quality machine-readable format, for example, through optical character recognition and handwritten text recognition.
- **Restoration:** the process of recovery of missing text and reassembly of fragmented written artifacts.
- **Attribution:** contextualizing a document within its original geographical, chronological, and authorial setting (i.e., who wrote the text, when, and where).
- **Linguistic analysis:** involving linguistic tasks such as semantic analysis, part of speech (POS) tagging, text parsing, and segmentation.



**Figure 2**  
Proposed taxonomy to study machine learning for ancient languages, inspired by the different steps involved in the study of ancient documents.

- **Textual criticism:** the process of reconstructing a text's philological tradition of textual transmission.
- **Translation and decipherment:** which aim to make a text's language comprehensible and interpretable to modern-day researchers.

The majority of the surveyed works operate on text, apart from those focusing on quality enhancement (*digitization*), fragment reassembly (*restoration*), palaeographic analysis, and writer identification (*attribution*), which instead operate mainly on visual inputs; whereas works focusing on recognition (*digitization*) harness both modalities.

The arrangement of our taxonomy into separate sections is intended to enhance readability. Each section's content dictates the underlying paper review order, be it chronological, by language, or by model type. By assigning a logical structure to each section, the reader will be able to follow related works in a more straightforward manner.

## 2. Background

Before commencing the discussion of each step in our pipeline, we will spare a few words to the history of machine learning over the last two decades, for the benefit of historians and non-experts. Before the advent of deep learning, scientists developed hand-crafted features to describe the input data and better address different tasks. These features ranged from image descriptors (e.g., Histogram of Oriented Gradients [HoG] and Scale Invariant Feature Transform [SIFT] for object recognition and image classification [Forsyth and Ponce 2011, p. 155]), to frequency-based or categorical text features (e.g., term frequency-inverse document frequency [TF-IDF] [Manning and Schütze 1999, p. 543], grammatical features). These features were then used as inputs to clustering algorithms (e.g., *k*-means, Gaussian mixture models, which automatically discover groups of similar texts), statistical models (e.g., hidden Markov models [HMMs] and conditional random fields [CRFs], which learn to tag sequences by assigning POS tags to word sequences), and classification algorithms (e.g., support vector machines [SVMs], random forests, and boosting algorithms, which learn to classify texts by author, historical period).

However, the capabilities of these hand-crafted feature representations were limited and tailored to tackle only specific tasks, and as such were gradually superseded starting in 2012 with the advent of deep learning and neural network (NN) models. NN models have the ability to process raw data and automatically learn the representations needed for the task. This representation learning in a supervised setting relies on labeled data, where the model is trained to learn the optimal features of the inputs to predict specific labels. In contrast, unsupervised representation learning focuses on identifying patterns and structures within the data without explicit guidance, allowing the model to capture intrinsic characteristics and form meaningful features. In NNs this process is implemented hierarchically by using non-linear layer modules that transform data from one layer of representation to the next, gradually increasing the abstraction level. Complex functions can be learned by the composition of enough layers (LeCun, Bengio, and Hinton 2015).

Massive increases in computing power engendered by the parallel processing capabilities of Graphics Processing Units (GPUs), the growing availability of large datasets, and continued methodological advances could now enable the NN models to learn better and more generalizable feature representations from the data itself as part of

their training process. The initial focus was on convolutional neural networks (CNNs) working on images. The mechanisms behind CNNs were inspired by visual neuroscience: simple cells that are sensitive to specific orientations of edges, and complex cells that respond to patterns of edges with a particular orientation and spatial frequency (Hubel and Wiesel 1962). CNNs were soon extended from images to learning word representations, for example, word2vec (Mikolov et al. 2013): In this setting, words are represented by means of embeddings, a learned representation for text where words with the same meaning will also have a similar representation. At the same time, recurrent neural networks (RNNs) (Goodfellow, Bengio, and Courville 2016) started to show enormous potential for modeling text sequences, for example, long short-term memory (Hochreiter and Schmidhuber 1997) and gated recurrent units (Chung et al. 2014). NN models continued to evolve, classifying inputs, mapping textual sequences to sequences (seq2seq) (Sutskever, Vinyals, and Le 2014), featuring attention mechanisms for RNNs, introducing novel generative models (e.g., variational autoencoders and generative adversarial networks [GANs] for text and image generation [Goodfellow, Bengio, and Courville 2016; Goodfellow et al. 2020]). One of the most important breakthroughs was the Transformer (Vaswani et al. 2017) in 2017—a deep learning model that relied extensively on attention mechanisms to better capture contextual information, and which could process sequences in parallel, unlike RNNs. Soon, Transformer-based models became the standard for extracting features from texts (e.g., BERT; Devlin et al. 2019), while other larger Transformer models—for example, GPT-3 (Brown et al. 2020), PaLM (Chowdhery et al. 2022), and ChatGPT—trained on even larger corpora demonstrated emergent capabilities on a wide range of tasks. Today, such large-scale models are routinely applied to protein folding, video generation and translation, image captioning, as well as the study of ancient languages. The works surveyed in this research retrace the above-mentioned chronological progression.

To conclude this preamble, we must stress that progress in machine learning relies not only on powerful models, but also on the quality and quantity of datasets, evaluation metrics, and experimental protocols. For this reason, we emphasize: (a) the direct correlation between the choice of dataset and a model's performance; (b) the importance of robust hypothesis testing, with data partitioning (into train, validation, and test sets) or resampling to train different models and measure their effectiveness and stability in generalizing results (e.g., with cross-validation); and (c) the value of statistical significance tests, which ensure that observed differences in performance across models are not merely random artefacts.

### 3. Digitization

#### 3.1 Recognition

The first steps in the study of ancient texts using machine learning are digitization and recognition. The task of automated transcription of a text from the image of a written support (e.g., a photograph, drawing, or scan of an inscription, manuscript, or papyrus fragment) to its digitized form is a central task in the conservation of ancient documents, making them digitally accessible for downstream tasks. Optical Character Recognition (OCR) and its sub-field Handwritten Text Recognition (HTR) are indeed well-studied areas of research.

The tasks of OCR/HTR have been attracting interest of early traditional machine learning approaches since the 1980s. These early works were followed by efforts

to use self-adapting software to digitize the Latin writing tablets from Vindolanda (Terras and Robertson 2005) and open and closed cavity character detection on early Christian manuscripts in Greek (Gatos et al. 2006; Ntzios et al. 2007). In more recent years, research focused on training new models or on adapting existing OCR engines to Latin and ancient Greek text recognition. Indeed, the impact of open-source OCR engines such as Abbyy FineReader, Gamera, Tesseract, OCRopus<sup>1</sup> has been essential to this field, granting free access to off-the-shelf OCR solutions to humanities researchers. For the purpose of this review, we will focus exclusively on the *ex novo* development of models and tools for OCR/HTR. Firmani et al. (2018) used a word to character segmentation algorithm followed by a CNN to classify characters in Latin historical documents, and then used language modeling to yield their word transcriptions. More recently, Swindall et al. (2021) introduced two “Ancient Lives” datasets consisting of more than 490k labeled character images of ancient Greek manuscripts manually annotated by volunteers. The authors evaluated multiple CNN models, and the highest accuracy was obtained using a residual neural network (ResNet) model (He et al. 2016a). The importance of such manually annotated datasets cannot be underestimated, an observation which applies throughout this review.

Multiple efforts have focused on cuneiform sign recognition. Early works (Edan 2013; Mostofi and Khashman 2014) used *k*-nearest neighbors (*k*-NN, classifying each test instance to the majority class of its most similar training instances) or NNs with very few layers, to classify small subsets of cuneiform signs. Further work has attempted to build an automated pipeline for transliterating entire lines of text (Bogacz, Klingmann, and Mara 2017). Working on a small number of tablets, the pipeline segments the lines, extracts image features, and aligns them to their transliteration. The best performing model was an HMM, but the accuracy was low, leading the authors to conclude that transliteration is tractable, but requires significantly cleaner data. A fully automated approach for automatic transliteration was proposed by Dencker et al. (2020). The authors began by weakly aligning sign transliterations to their corresponding tablet images, using a CNN and a CRF, and then trained a CNN sign detector. Combining these steps in an iterative process enabled the training of a better aligner and, as a result, a better sign detector. The model was evaluated on tablets from the Oracc and CDLI datasets. Several other ancient writing systems have been the focus of HTR/OCR efforts, including: Devanagari (Narang, Jindal, and Kumar 2019; Narang et al. 2020; Narang, Kumar, and Jindal 2021; Jindal and Ghosh 2022), Egyptian hieroglyphs (Franken and van Gemert 2013; Haliassos et al. 2020; Barucci et al. 2021; Moustafa et al. 2022), Old Tamil (Suganya and Murugavalli 2017; Subramani and Murugavalli 2019; Devi et al. 2022), ancient Ge’ez (Demilew and Sekeroglu 2019), Brahmi (Wijerathna et al. 2019), Grantha (Raj, Jyothi, and Anilkumar 2017), Indus script (Palaniappan and Adhikari 2017), Maya glyphs (Can, Odoñez, and Gatica-Perez 2016), Oracle bone (jiāguwén) script (Zhang et al. 2019), Phoenician (Rizk et al. 2021), and Tibetan script (Liu et al. 2022).

HTR remains among the most challenging tasks in machine learning for ancient writing systems. The implementation of recognition pipelines relies upon the existence of digital images, and a successful pipeline requires high quality and quantity of digitizations; for example, compare the rich datasets of cuneiform tablet images to their paucity for Greek inscriptions. A standard recognition pipeline comprises: image

1 Abbyy FineReader: <https://www.abbyy.com>, Gamera: <https://gamera.informatik.hsnr.de>, Tesseract: <https://github.com/tesseract-ocr/tesseract>, OCRopus: <https://github.com/ocropus>.

pre-processing, text segmentation, feature extraction and classification, and post-processing. Segmentation can work at a line-, word- or character-level, and is a crucial phase of an OCR/HTR system, as it can directly affect the overall accuracy of transliterations (Narang, Jindal, and Kumar 2020). Several studies propose HTR as a classification problem of pre-segmented character images (Terras and Robertson 2005; Edan 2013; Franken and van Gemert 2013; Mostofi and Khashman 2014; Can, Odobez, and Gatica-Perez 2016; Raj, Jyothi, and Anilkumar 2017; Firmani et al. 2018; Zhang et al. 2019; Subramani and Murugavalli 2019; Narang, Jindal, and Kumar 2019; Narang et al. 2020; Haliassos et al. 2020; Swindall et al. 2021; Barucci et al. 2021; Rizk et al. 2021).

But character segmentation is not a solved task, especially in scripts where character boundaries overlap (e.g., cursive handwriting). Such challenges are made even more taxing by the state of preservation of ancient written media (damaged supports, low quality images, etc.). As a direct consequence, other studies have chosen to approach the task more holistically by introducing pipelines that include handcrafted or trained alignment and segmentation models (Gatos et al. 2006; Ntzios et al. 2007; Palaniappan and Adhikari 2017; Bogacz, Klingmann, and Mara 2017; Suganya and Murugavalli 2017; Demilew and Sekeroglu 2019; Wijerathna et al. 2019; Dencker et al. 2020; Gordin et al. 2020; Narang, Kumar, and Jindal 2021; Devi et al. 2022; Moustafa et al. 2022; Liu et al. 2022; Jindal and Ghosh 2022). Both approaches saw substantial improvements when using CNNs. For further details on recognition and digitization, we refer the reader to the recent subject-specific surveys by Bhurke et al. (2020) and Narang, Jindal, and Kumar (2020).

### 3.2 Quality Enhancement

When faced with a lack of high-quality image datasets or of significant variability in the data (owing to the paucity of digitized documents), *enhancing or restoring the quality* of existing datasets can yield better results in downstream tasks.

In 2003, Molton et al. (2003) focused on the visual enhancement of Roman stylus tablets using edge detection methods. In 2016, Faigenbaum-Golovin et al. (2016) analyzed ancient Hebrew inscriptions as parallel evidence for dating early biblical texts. However, because of the damaged state of the inscriptions, visual restoration was required to compare different handwriting styles and determine the inscriptions' author. The authors approached the problem of restoring characters on the basis of their composing strokes and representing them as spline-based structures, estimated using optimization.

More recent studies have resorted to NN models. Parker et al. (2019) presented a non-invasive digital recovery method for the carbonized texts of Herculaneum, showing that X-ray-based micro-computed tomography data can capture the presence of carbon ink. The authors used a 3D CNN to detect the volumetric presence of ink using reference papyrus rolls that had been already opened and inspected for writing. Then, using a virtual "unwrapping" pipeline, they were able to align these labels with the tomography volume and reveal the presence of letters in unopened scrolls. In 2020, Zhao et al. (2020) introduced a Laplacian pyramid GAN to enhance low-resolution inputs for ancient Shui handwriting recognition. Subsequently, the authors used an unsupervised clustering algorithm based on information entropy for automatically annotating the manuscript's character images. Similarly, Brandenbusch, Rusakov, and Fink (2021) used a conditional GAN for cuneiform sign inpainting in existing images of tablets. The model was trained on hundreds of photographs of tablets with 45k signs



annotated by bounding boxes, where random patches would be cropped around the signs to be infilled. Further encoder-decoder architectures were evaluated by Yu et al. (2022) for the visual restoration of the Mogao caves findings in Dunhuang.

But small datasets may be imbalanced, thereby introducing bias in the results: One must then seek to improve not just the quality but also augment the quantity of existing datasets. Swindall et al. (2022) used a GAN to generate synthetic characters of ancient Greek letters to balance a dataset of 400k papyri images (Swindall et al. 2021). The synthetic dataset resulted in a 12% recognition accuracy increase. Huang et al. (2022) also introduced a GAN architecture for Oracle bone and cuneiform glyph generation to address the data scarcity problem. The proposed model architecture cascaded a glyph transformation and a texture-transfer GAN. Finally, Nguyen et al. (2021) proposed an encoder-decoder NN architecture for de-noising the images of inscriptions in ancient Cham script. The architecture used attention over multiple scales to enhance the de-noised images. In an artificial noise-generation setting, the proposed model outperformed component analysis methods and other NN models.

## 4. Restoration

### 4.1 Textual Restoration

Over the centuries and millennia, ancient texts can be fragmented or become illegible owing to the deterioration or destruction of writing supports. Historians must then reconstruct the lost or illegible parts of the text, a process known as **textual restoration** (Matsumoto 2022). This is a complex and time-consuming task (Woodhead 1959): Specialists typically rely on textual and contextual “parallels” (recurring expressions or linguistic peculiarities) to reconstruct missing parts in similar texts.

Early modeling attempts to automate textual restoration used  $n$ -gram Markov chains for texts in the Indus script (Rao et al. 2009b; Yadav et al. 2010). Assael, Sommerschild, and Prag (2019) were the first to address the problem of text restoration using deep learning. Their work focused on Greek inscriptions and introduced an autoregressive sequence-to-sequence RNN model called Pythia. Pythia operated at both word- and character-level, the intuition being that words convey context, but parts of words may be damaged. On a purpose-made dataset based on the Packard Humanities’ Institute (PHI) dataset of ancient Greek inscriptions, Pythia achieved a 30% character error rate, compared with the 57% of two evaluated human specialists. Moreover, in three out of four cases, the ground-truth sequence was among the model’s Top-20 restorations. Fetaya et al. (2020) presented an RNN language model for token prediction and missing word completion in fragmentary Akkadian tablets. The RNN model was far more accurate than traditional  $n$ -gram baselines. Similar trends were observed in the work of Papavassiliou, Kosmopoulos, and Owens (2022), who used a bidirectional RNN trained on original and augmented data of Linear B inscriptions (Papavassiliou, Owens, and Kosmopoulos 2020), which exhibited a higher accuracy compared with  $n$ -gram baselines.

The introduction of Transformer-based models has led to significant advances in this field. Shen et al. (2020) introduced a Transformer-based architecture capable of generating sequences by dynamically creating and filling in blanks. The architecture was evaluated on the dataset introduced by Assael, Sommerschild, and Prag (2019) and performed on-par with Pythia, but it could also generate arbitrary sequences without the need of experts specifying the target length. Bamman and Burns (2020)

pre-trained a BERT (Devlin et al. 2019) model on Latin texts from Perseus, PROIEL, and Index Thomisticus Treebank, targeting restoration and several other downstream tasks. The effectiveness of Latin BERT's restoration accuracy was evaluated against the emendations made by experts with an accuracy of 33%, and many of its restorations were within its Top-10 predictions. Another masked language modeling Transformer architecture was proposed by Lazar et al. (2021) for the restoration of cuneiform tablets (in Akkadian). The model achieved an 83% accuracy on the Oracc dataset. The authors also evaluated two human annotators, who reviewed the model's Top-5 predictions. In the majority of cases, they accepted the model's restorations when up to 2 characters were missing, whereas when 3 characters were missing they accepted only half of the restorations generated by the model. Kang et al. (2021) used a Transformer-based model to restore and translate Korean historical records dating to the Joseon Dynasty. The model's Top-10 restoration accuracy was 89%. Finally, Assael et al. (2022) introduced Ithaca, a sparse-attention Transformer-based architecture for restoring, dating, and attributing ancient Greek inscriptions. Like its predecessor Pythia, Ithaca operates at both a character- and a word-level. To train Ithaca, the authors created a processed version of the PHI dataset of Greek inscriptions. While Ithaca alone achieved 62% accuracy when restoring damaged texts, as soon as evaluated historians used Ithaca, their accuracy leaped from 25% to 72%, thus effectively demonstrating the impact of this synergistic research aid. Finally, Ithaca uses saliency maps as a visual aid to highlight and inform the historians about which inputs were most important for the model's predictions.

Several studies relied on human baselines to measure effectiveness: Assael, Sommerschild, and Prag (2019), Lazar et al. (2021), Assael et al. (2022), but only Assael et al. (2022) sought to augment the interpretability of predictions using tools such as saliency maps and distributional outputs for human experts to evaluate in a real-world setting. The inclusion of humans in the training loop could result in more effective research.

## 4.2 Fragment Reassembly

Written supports may be broken into several pieces, which must be *reassembled* or visually restored to make the text legible again.

In 2012, Tyndall (2012) proposed naive Bayes and maximum entropy classifiers for rejoining the texts of fragmentary Hittite cuneiform tablets. Collins et al. (2014) introduced a matching algorithm for 3D scans of cuneiform tablets. In 2019, Pirrone, Aimar, and Journet (2019) used a Siamese-network architecture to reassemble papyrus fragments. By extracting "patches" of papyrus fragment images, they deployed a NN to score each matching pair. The model achieved a high accuracy on a synthetic dataset comprising gapless fragments in Coptic, Greek, Arabic, Hebrew, Hieratic, Latin, and Demotic from the APIS UM Papyrology Database. In 2021, Abitbol, Shimshoni, and Ben-Dov (2021) introduced a more complex system for the reassembly of the Dead Sea Scrolls. They first identified continuous natural fibre thread patterns in the papyrus fragments by processing square patches of different fragments through a CNN. The resulting local matching scores were then fed into a voting mechanism enhanced by geometric alignment techniques and a random forest classifier. The system produced a list of candidates for expert evaluation. In 2022, Zhang et al. (2022a) proposed a self-supervised network to rejoin bone fragments (inscribed in Oracle Bone script) based on shape similarity between joining fragments. The model consisted of a GAN for augmenting positive pairs of re-joinable fragments and a Siamese network trained on

the augmented data to retrieve the matching Oracle Bone fragments from a fragment gallery. The network could reassemble half of the previously disjointed fragments.

Fragment reassembly is a challenging problem, as the lack of real-world datasets poses significant obstacles. To overcome this hurdle, the development of real-world datasets and establishment of benchmark challenges could facilitate future research in this field.

## 5. Attribution

### 5.1 Language Identification

Ancient languages evolve over time and vary in space: Words fall in and out of use, grammar changes, regional dialects develop. Thus, *attributing* ancient texts to their place and time of writing is key to grounding them within their original historical and cultural context.

Identifying what language a text might be written in is a task that has received particular attention in machine learning competitions. An influential effort in language identification was initiated by Jauhiainen et al. (2019), who introduced a corpus of texts from Oracc for the Cuneiform Language Identification shared task, part of the 2019 VarDial Evaluation Campaign (Zampieri et al. 2019). The authors evaluated multiple statistical frequency methods to classify different languages in cuneiform script, and the product of relative uni- to four-gram frequencies exhibited the best performance. On the same challenge, Bernier-Colborne, Goutte, and Léger (2019) proposed a modified version of the BERT model taking characters as input, which led to substantial performance improvements. A similar performance was exhibited by an SVM classifier used by Wu et al. (2019), which used character and word weighted  $n$ -grams. The authors utilized test-time adaptation to label the validation set, and then retrained the model on the whole dataset. Several other teams also proposed SVM-based models (Benites de Azevedo e Souza, von Däniken, and Cieliebak 2019; Paetzold and Zampieri 2019; Doostmohammadi and Nassajian 2019) on the same task. Meloni, Ravfogel, and Goldberg (2021) introduced a seq2seq RNN model to identify phonetic alterations between words in Latin and in “daughter” Romance languages. The authors constructed a dataset of 8k comparative entries, and showed that NN models outperform non-NN models in detecting historic language change.

### 5.2 Chronological Attribution

When *dating* ancient documents, specialists often rely on internal (paleographical, prosopographical) and external (archaeological) contextual clues. Modern techniques (e.g., C-14 radiocarbon dating) are unviable when the writing supports are made of inorganic materials (stone), and are often prohibitively expensive.

In 2003, Kashyap and Koushik (2003) were the first to build a probabilistic NN for dating texts in Kannada script. In 2014, Soumya and Kumar (2014) used binarized image features and random forests to date the images of 110 Kannada inscriptions to 6 historical periods. To increase the amount of data available for the dating task, Adam et al. (2018) created KERTAS, a dataset of 2k high-resolution images of Arabic manuscripts dating between the 8th and 14th century CE. The authors used k-NN to predict each image’s century with 86% accuracy. In 2019, Yu and Huangfu (2019) proposed an RNN for dating ancient Chinese documents. The authors extracted 800K characters from several ancient documents and then dated to three different

historical periods with very high accuracy. The same year, Goler et al. (2019) used Raman spectrography to date the carbon ink in Egyptian papyri dating between the 4th century BCE and the 10th century CE. They trained a Gaussian mixture model on a dataset of 17 papyri to model the distribution of a particular set of Raman spectral parameters, and their discoveries had important implications for the authenticity of two controversial papyri, the “Gospel of Jesus’ Wife” and a fragment of the Coptic Gospel of John. In 2020, Bogacz and Mara (2020) used a CNN to classify cuneiform tablets among four historical periods. The model exhibited high accuracy on the Heidelberg Cuneiform Benchmark dataset, but half of the samples were attributed to a single historical period. Further research operating on per century classification of ancient Greek papyri images was conducted by Paparigopoulou, Pavlopoulos, and Konstantinidou (2022) and the best results were obtained using a multilayer perceptron (MLP) trained on CNN-derived features. Harnessing recent advances in large-scale language models, Assael et al. (2022) introduced Ithaca, a sparse-attention Transformer-based architecture for the chronological attribution of ancient Greek inscriptions. On the I.PHI dataset, Ithaca could date texts to less than 30 years of their ground-truth ranges, outperforming the evaluated human baseline four times over. The authors also used Ithaca to re-date some of the most important decrees of classical Athens whose dating is controversial. Using a similar large-scale architecture inspired by BERT, Yoo et al. (2022) presented a dataset, HUE, and a model for dating, topic classification, named entity recognition, and summary retrieval tasks of ancient Korean Hanja documents. The model was pre-trained on two large textual datasets and fine-tuned on two smaller datasets containing historical annotations. The model could attribute texts to the different kings of the Joseon dynasty with very high accuracy. Finally, Chang et al. (2022) modeled the historical evolution of characters in the Oracle Bone script using a GAN architecture.

An issue shared by all these efforts is data circularity. The dates recorded in the models’ training datasets are the product of accumulated scholarly knowledge, which may imply circularity in results. Emphasis on dataset analysis could avoid pinning misleading objectivity to dating predictions.

### 5.3 Geographical Attribution

Written monuments or supports may have been moved in ancient or modern times for a multitude of reasons, and experts must then establish their *geographical attribution* (Tsirogiannis 2020).

Materials analysis offers one possibility (Harper et al. 2020), but Assael et al. (2022) was the first to use a deep NN architecture to attribute Greek inscriptions among 84 ancient regions (among other tasks) with an accuracy three times higher than the evaluated human baseline. Yamshchikov et al. (2022) fine-tuned existing BERT language models trained on ancient Greek to attribute texts among different authors and four regions. Focusing on pseudo-Plutarchian works, they demonstrated that the texts could have originated from an Alexandrian context.

Further work on language identification and geographical-chronological attribution should attempt to shed light upon the possible reasons underlying a model’s hypotheses: Although historians know that linguistic variation and regional-thematic practices contribute to the distinctiveness of writing habits, expanding the interpretability of models’ results (e.g., saliency maps, retrieval) could illuminate previously unknown patterns, habits, and regionalities.

## 5.4 Topic Modeling, Genre Detection

Texts can be grouped within the system of literature on the basis of their shared features of form, style, and contents. Automatic genre detection has been approached through **topic modeling**, a machine learning technique for clustering and classifying document topics.

Early works on topic modeling focused on text clustering. In 2013, Bracco et al. (2013) used the *k*-means clustering algorithm to group transliterated cuneiform texts sharing stylistic features. The authors computed the frequency features for ancient Babylonian letters and experimented with different clusters. In 2017, Wishart and Prokopidis (2017) adapted a POS tagger and a lemmatizer to Hellenistic Greek. The processed texts from a dataset comprising the Greek New Testament, the ancient Greek Dependency Treebank, and the O'Donnell corpus were used as inputs to a Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) statistical model for topic modeling, determining the most significant words per topic. Similarly, in 2020 Köntges (2020) used an LDA model on ancient Greek philosophical texts from the First1kGreek and Perseus corpora. Using the topics discovered, the author distilled three numeric scores for philosophical text in Ancient Greek: One score measured “good and virtue,” the second score measured “scientific inquiry,” and a third combined the two to measure the “philosophicalness” of a given text. Kaše, Heřmánková, and Sobotková (2021) extracted frequency features from Latin inscriptions and compared extremely randomized trees, random forests, and SVM classifiers in predicting different inscription categories (honorific, epitaph, curse, etc.). The categories originated from the Epigraphic Database Heidelberg (EDH) and were used to label inscriptions from the Epigraphic Database Clauss-Slaby (EDCS) with high accuracy, significantly enriching the original dataset's inconsistent metadata. Finally, in 2022 Yoo et al. (2022) introduced a dataset and a Transformer-based model for modeling historical documents written in Hanja, ancient Korean. Among other tasks, the model was able to classify the documents among hundreds of major and minor topics.

## 5.5 Authorship Attribution

*Authorship attribution* is the task of determining the author(s) of a text, often based on salient stylistic markers and characteristics (Koppel, Schler, and Argamon 2009). This task has been supported by statistical or computational methods since the 19th century (Stamatatos 2009). Today, “quantitative authorship attribution” (Grieve 2007) can be essentially envisaged as a classification task, building upon a background training dataset of multiple authors, from which textual features can be extracted and compared in order to classify text(s) to author(s).

Following this approach, Koentges (2020) paired the analysis of word and character *n*-grams with philological arguments concerning the authorship of the “Menexenus,” a contested Platonic dialogue in ancient Greek. He extracted a feature set from a background corpus of literary works in ancient Greek, and computed the similarities to conclude that the “Menexenus” is in all likelihood not Platonic. Tang, Liang, and Liu (2019) selected a small set of linguistic features to classify the novel “The Golden Lotus” (in vernacular Chinese) among four authors, using a background dataset of poems. Their experiments concluded that Wei Xu's writing style was closest to that of the “The Golden Lotus.” The *k*-NN approach of Martins et al. (2021) was unable to offer definitive results for the multi-author classification task of the contested “Historia

Augusta” (in Latin). Corbara, Moreo, and Sebastiani (2022) used syllabic quantity to derive an additional set of stylistic features for attributing texts by Latin authors using an SVM, based on the fact that certain authors show a preference for specific rhythmic patterns obtained by specific sequences of long and short syllables. Yamshchikov et al. (2022) fine-tuned existing BERT language models trained on ancient Greek. The model focused on pseudo-Plutarchian works, and could attribute texts to different authors active in different regions of the ancient world.

A background dataset is key to extracting meaningful features, but datasets are often imbalanced. Unless sampling is done robustly, there is the risk of introducing bias in the results. For example, Reisi and Mahboob Farimani (2020) used a CNN with a self-attention mechanism to determine that the “Khān al-Ikhwān” (in ancient Persian) was written with high probability by Nāsir-i Khusraw, but their background dataset of Persian literature comprised only 4k randomly selected sentences from the train and test sets. Others have preferred instead to select a fixed number of texts per author to test their SVM classifier on ten ancient Arabic travel writers (Ouamour and Sayoud 2012, 2013a, b, 2018), while Kestemont et al. (2016) applied approximate randomization to validate the statistical significance of their results on the authorship verification of the Latin “Corpus Caesarianum.” Following prior work, Koppel and Winter (2014) and Stover et al. (2016) used repeated feature sub-sampling and a pool of impostor authors to attribute the newly discovered Latin “Compendiosa Expositio” to the author Apuleius, based on the text’s similarities to his “De Platone.” These results confirmed philological analyses on the new text.

Further works investigated the possibility of authorial variability taking place within the same literary work. Manousakis and Stamatatos (2018) applied an SVM classifier with character  $n$ -grams to both full-texts and text segments from their background dataset of ancient Greek plays to capture the authorial variability in the play “Rhesus,” whose Euripidean authorship is contested. Finally, variability within the same work might be due to interpolation: Tuccinardi (2017) analyzed Pliny the Younger’s Latin correspondence to the Roman emperor Trajan, using profile-based methods on  $n$ -grams features to detect a large amount of interpolation in the text of the letters. Pavlopoulos and Konstantinidou (2022) used a statistical language model’s perplexity on samples of the ancient Greek epic poems “The Iliad” and “The Odyssey” to identify outlier-passages from other texts of the Homeric canon.

## 5.6 Palaeographic Analysis and Writer Identification

Machine learning has been successfully used to identify the hand(s) of the people writing ancient texts, based primarily on digitized handwriting images (Davis 2007; Stokes 2015). *Palaeographic analysis* enables, for example, identifying writing hands, joining or grouping segments, dating documents and their writers, and making historical observations concerning ancient scribal cultures.

Palaeography is based on the analysis of letter shapes, and the main difficulty lies in distinguishing between variations in one writer’s style, and similarities in different writers’ styles (Popović, Dhali, and Schomaker 2021). Significant research has been carried out on distinguishing medieval scribes working on, for example, ancient Greek, Latin, or Hebrew texts, but this time-frame is beyond the remit of the present work, as the transmission of ancient authors’ works cannot in most cases be traced back to antiquity. The contribution of conferences such as ICFHR and ICDAR (Fiel et al. 2017; Christlein et al. 2019; Seuret et al. 2020; Lai, Zhu, and Jin 2020; Chammas et al. 2022) to the field is substantial, as they sponsor competitions addressing topical challenges.

Relevant reports, datasets, methodologies, and evaluation metrics are made public after the conclusion of the competitions, increasing accessibility to key resources.

Current approaches to writer identification focus on either global (text-based) or local (grapheme-based) features (De Stefano et al. 2018). Recent studies have shown not only that the former are more effective (He et al. 2016b), but also that joint feature representations can improve the classifier's performance (Dhali et al. 2017). For ancient Greek, Panagopoulos et al. (2008) and Tracy and Papaodysseus (2009) used image segmentation and statistical analysis to isolate the writing styles of different cutters on a small dataset of Athenian inscriptions; follow-up work by Papaodysseus et al. (2010) also operated on a character-level, but also created an ideal prototype of each alphabet character, from which they extracted geometric features using heuristics and maximum likelihood estimation procedures. The same research group (Papaodysseus et al. 2014; Arabadjis et al. 2013) dated Greek inscriptions and Byzantine codices based on analytical comparisons between letter curvatures and "ideal prototype" curvatures. Arabadjis et al. (2019) also identified specific writers across codices recording editions of "The Iliad" based on letter curvature and on the ideal prototype of letter rendition. Considerable work has also been done on ancient Hebrew texts: Faigenbaum-Golovin et al. (2016) and Shaus (2017) worked on a set of Hebrew ostraca (ink on clay fragments) from the Arad fortress in the First Temple period (8th–6th century BCE). They focused (among other tasks) on writer attribution using a combination of features (including binary pixel patterns and Fisher's method) to estimate a minimum of 6 different writers. Dhali et al. (2017) worked on the Dead Sea Scrolls (3rd century BCE to 1st century CE in Hebrew, Greek and Aramaic): They used a feature representation method that relied on the curvature information of the neighboring fragments to detect regions of interest and attribute them to different scribes. Popović, Dhali, and Schomaker (2021) revealed a break in a section of the Great Isaiah Scroll (part of the Dead Sea Scrolls): They observed that the handwritten columns of the first and second halves of the manuscript showed substantially different feature-representations. A further visual inspection concluded that the break was due to the activity of two different scribes working on the scroll. We refer the reader to Faigenbaum-Golovin, Shaus, and Sober (2022) for further details on handwriting analysis and writer identification for Hebrew texts.

As in the case of authorship attribution, imbalanced datasets can penalize a model's performance, especially when training datasets are limited in size. For example, Mohammed, Marthot-Santaniello, and Märgner (2019) created a new dataset of 50 handwriting samples in Greek papyri. The authors used a Naive Bayes Nearest Neighbor classifier with FAST (Features from Accelerated Segment Test) keypoints to identify 10 different hands. Compare this situation with the automated palaeography of post-antique Arabic-language documents, for which there exist far richer image datasets (Asi et al. 2017; Abdelhaleem et al. 2017; Adam et al. 2018). For example, Fecker et al. (2014a, 2014b) attributed a dataset of thousands of images of Arabic manuscripts to different hands using traditional image features (SIFT, HoG) as inputs to a voting procedure and an SVM. Some works have attempted to work around the problems of data availability: for example, Nasir and Siddiqi (2020) evaluated different pre-trained CNNs for the palaeographic analysis of Greek papyri, and compared them with FAST features and a Nearest-Neighbor classifier. The models were first tuned on contemporary handwriting images (relatively larger dataset) and later tuned to the smaller dataset of papyri. Researchers must be aware of this data imbalance: Certain languages or writing methods (ink texts written with a pen or brush vs inscribed texts on stone or metal) are underrepresented in this task owing to the lack of digitized data. Moreover, just as for the chronological attribution task, the risk of data circularity within

palaeographical arguments and the ground truth labels they produce should be taken into serious account.

## 6. Linguistic Analysis

### 6.1 Representation Learning

The success of machine learning algorithms depends on *data representation*, as different representations can either reveal or obscure explanatory factors in the data (Bengio, Courville, and Vincent 2013). Although domain-specific knowledge may help manual feature engineering, learning automated generic representations has been shown to be especially helpful in downstream tasks.

A large body of literature tackling learning representations for ancient languages has focused on Latin. Bjerva and Praet (2015, 2016) used word2vec (Mikolov et al. 2013) to investigate the relationships between persons and themes of interest emerging in the works of the 6th century CE scholar Cassiodorus. Based on the extracted representations, they analyzed the text embedding associations between Cassiodorus, Liberius, Symmachus, and Boethius, and a selection of abstract concepts such as “liberty,” “antiquity,” “modern,” “Greekness,” “Romanness,” and “Gothness.” In 2019, Sprugnoli, Passarotti, and Moretti (2019) compared word2vec embeddings with fastText word embeddings (Grave et al. 2018) by using cosine similarity to fetch similar lemmas. FastText’s skip-gram embeddings exhibited a higher success rate in the task.

Bamman and Burns (2020) trained a BERT model on Latin corpora (Perseus, PROIEL, Index Thomisticus Treebank) and used the learned embeddings for word sense disambiguation and to identify words occurring in similar context, for POS tagging and language modeling. Finally, Burns et al. (2021) assembled a new benchmark and dataset for Latin synonym detection based on Valerius Flaccus’ “Argonautica.” The authors compared different implementations of BERT, word2vec, and fastText embeddings: Using a newly lemmatized Latin corpus, they showed that their embeddings could enhance intertextual search.

The work of Svärd et al. (2018) used word2vec embeddings and Pointwise Mutual Information to find collocations and associations between words in Akkadian. The model was trained on transliterated and lemmatized Akkadian cuneiform texts from the Oracc dataset. More recently, Karajgikar, Al-Khulaidy, and Berea (2021) used word2vec embeddings for Linear A glyphs, as part of an extended analysis of the Minoan writing system, which is yet to be deciphered.

### 6.2 Word Segmentation and Boundary Detection

*Tokenization*, the process of locating word or character boundaries in a text, and *sentence segmentation*, the task of identifying sentence boundaries, can aid the automation of linguistic analysis of an ancient text (Palmer 2000). However, the ambiguities and varieties in human languages and writing systems (e.g., logographic such as cuneiform and classical Chinese vs. logosyllabic such as ancient Japanese) must be taken into account.

In 2010, Huang, Sun, and Chen (2010) presented a CRF model for segmenting classical Chinese texts into sentences and clauses. The model used *n*-gram features, word class, and phonetic information. Similarly, Wang et al. (2016) introduced an RNN for the same task in ancient Chinese. The performance was comparable to that of traditional CRF-based models, and the authors improved the model’s effectiveness by introducing a length-based penalty term. A bidirectional RNN model was presented by Hellwig



(2016) for Sanskrit sentence boundary detection, using a combination of morphological and lexical features as inputs. The model clearly outperformed the CRF baseline, but its accuracy was insufficient for real-world segmentation without human supervision. Li et al. (2018a) were the first to introduce a capsule-based model for word segmentation, which had high accuracy on a dataset of ancient Chinese medicine books developed by the authors. More recently, Zhang et al. (2021) presented a bidirectional RNN with attention and a CRF for identifying the boundaries of historical figures' first names. To generate training labels, the authors matched personal names from the ancient Chinese corpus of "Song History" with a dictionary of historical names.

In 2012, Yoshimura, Kimura, and Maeda (2012) presented a method for segmenting sentences in the ancient Japanese manuscript "Genji Monogatari" into words, by calculating the likelihood of character 2- to 10-grams being words. In 2016, Homburg and Chiarcos (2016) benchmarked rule-based, dictionary-based, and machine learning methods for the word segmentation of cuneiform tablets in Akkadian. The authors evaluated CRF, HMM, SVM, and NN models, but the dictionary-based approaches produced the best-performing classification. More recently, Yu et al. (2020) proposed a word segmentation model for ancient Chinese based on a non-parametric Bayesian model and BERT. Both models were repeatedly trained on large-scale unlabeled data. An and Long (2021) used a bidirectional RNN with a CRF for ancient Tibetan word segmentation, which achieved a high accuracy and outperformed the HMM and other baselines. Tupman, Kangin, and Christmas (2021) and Paolanti et al. (2022) also worked on word segmentation in Roman inscriptions and Medieval notary documents, respectively. Finally, recent work on ancient Chinese has combined the problem of word segmentation with POS tagging, and are presented jointly in the following section.

### 6.3 POS Tagging and Parsing

*Part-of-speech* (POS) tagging involves the grammatical mark up of a word in a text as corresponding to a particular part of speech, while syntax *parsing* generates parse trees, showing how words and phrases combine to form larger syntactic constituents. In our context, interest in both tasks has been spurred by conference challenges, especially with regard to the *writer identification* task.

In 2017, the Computational Natural Language Learning (CoNLL) conference featured a challenge (Zeman et al. 2017) that involved training dependency parsers on several languages, including ancient ones, and a real-world setting with noisy annotated labels. The goal was to detect syntactic dependencies and classify the dependency relation type. In 2018, the challenge expanded to morphological feature extraction, POS tagging, and lemmatization. The inputs consisted of simply raw text, without any segmentation or morphological annotations. A large number of submissions competed for the parsing task (Bhat, Bhat, and Bangalore 2018; Boroş, Dumitrescu, and Burtica 2018; Chen et al. 2018; Duthoo and Mesnard 2018; Jawahar et al. 2018; Ji et al. 2018; Kanerva et al. 2018; Kirnap, Dayanık, and Yuret 2018; Li et al. 2018b; Nguyen and Verspoor 2018; Qi et al. 2019; Rybak and Wróblewska 2020; Smith et al. 2018; Straka 2018; Wan et al. 2018): Each language was well represented in the datasets, which comprised treebanks from the Universal Dependencies 2.2 collection. Ancient Greek, for example, included 160k labeled words from Perseus and another 187k from PROIEL, while Latin data gathered 460k words. It is worth noting that the best performing method in ancient Greek and Latin was also the best method overall, combining contextual embeddings with ensembling (Che et al. 2018). These methods, such as Straka (2018), were built on prior work (Straka, Hajic, and Straková 2016) and were followed by Transformer-based

architectures (Straka, Straková, and Hajič 2019; Straka and Straková 2020), achieving an even higher performance. One of the most significant contributions of the CoNLL challenge was the resulting dataset for ancient languages, which has allowed subsequent research to investigate the data further (de Lhoneux, Stymne, and Nivre 2017) and expand existing resources to other ancient languages. For example, Keersmaekers et al. (2019) focused on ancient Greek and Bamman and Burns (2020) on Latin, both achieving the state-of-the-art in POS tagging with a Transformer-based architecture.

Ancient language-specific campaigns have also been organized by the workshop on Language Technologies for Historical and Ancient Languages (LT4HALA). In 2020, the EvaLatin (Sprugnoli et al. 2020) challenge focused on Latin POS tagging and lemmatization using texts from the Perseus dataset. The generalizability of competition submissions was evaluated using additional cross-genre and cross-time test sets. Most participants proposed RNN-based architectures (Straka and Straková 2020; Wu and Nicolai 2020; Bacon 2020; Stoeckel et al. 2020), while Celano (2020) used gradient boosting with pre-trained word embeddings, and Stoeckel et al. (2020) used an ensemble of classifiers for POS tagging. LT4HALA's EvaLatin 2022 campaign (Sprugnoli et al. 2022) focused on Latin POS tagging, lemmatization, and morphological feature identification using texts from the LASLA corpus, containing nearly 2 million words and corresponding to 133k unique tokens annotated by trained classicists, and 24k lemmas. The best performing participants Wróbel and Nowak (2022) trained Transformer-based models: an XLM-RoBERTa pre-trained on Latin for POS tagging and feature identification, and a ByT5 for lemmatization. Similarly, Mercelis and Keersmaekers (2022) started from a pre-trained small ELECTRA Transformer-based model for the POS tagging task, and handcrafted rules were added to handle lemmatization. The same year, LT4HALA introduced the first ancient Chinese word segmentation and POS tagging challenge, EvaHan 2022 (Li et al. 2022). The challenge used texts from ancient Chinese chronicles and featured a "closed" part involving limited data and a pre-trained RoBERTa model, and an "open" part without resource limitations. Some participants used traditional RNN architectures (Tang, Lin, and Li 2022) and CRFs (Yang 2022), while others focused on Transformer-based alternatives, adversarial training (Zhang et al. 2022b; Yang 2022), data augmentations, and ensemble learning (Zhang et al. 2022b; Yang 2022; Wei et al. 2022) to compensate for the limited and imbalanced training data.

Outside competitions, several recent studies have focused on Transformer-based architectures. For example, Singh, Rutten, and Lefever (2021) used a corpus of modern, ancient, and Byzantine Greek texts to further pre-train a BERT model and then fine-tune it for ancient Greek POS tagging. A similar study was performed by Tian et al. (2021), who used Chinese articles, poems, and couplets dating between 1000 BCE and 200 BCE for pre-training a BERT model and then fine-tuning it to the classification and text generation tasks. Others have used more traditional approaches: Hellwig (2015) used maximum entropy classifiers and CRFs for the tokenization and morphosyntactic analysis of writings in ancient Sanskrit. A joint RNN-CRF architecture for ancient Chinese word segmentation and POS tagging was also presented by Cheng et al. (2020). Sahala et al. (2020b) used a finite-state transducer (FST) to address lemmatization and POS tagging for cuneiform tablets in Babylonian from the Oracc corpus. Phonological transcription is essential for the automatic morphological analysis of cuneiform. The same group of Sahala et al. (2020a) presented a character-level sequence-to-sequence model with attention for the automated phonological transcription of transliterated text. This was the first attempt to automatically transcribe Akkadian, and the predictions were evaluated using the FST of Sahala et al. (2020b). Finally, other efforts (Celano, Crane, and Majidi 2016; Vatri and McGillivray 2018, 2020) used off-the-shelf software.

## 6.4 Semantics

*Computational semantics* seeks techniques for automatically constructing semantic representations of expressions in natural language (Blackburn and Bos 2005).

In 2011, Bamman and Crane (2011) used k-NN, naive Bayes, and statistical language modeling for measuring Latin word sense variation using a processed collection of 7k books. Aligning a small collection of parallel texts, the authors introduced a bilingual sense inventory that was then used to tag a 389 million word corpus and track the rise and fall of word senses over 2,000 years. More recently, Yoo et al. (2022) introduced a dataset and a Transformer-based model for analyzing historical documents written in the Hanja writing system. Among other tasks, the model performed named entity recognition.

The rest of the literature focuses on ancient Greek literary texts. Perrone et al. (2019) designed a Bayesian mixture model for measuring the evolution of word sense over time, based on distributional information of lexical nature and genre. The model was evaluated on the Diorisis Ancient Greek Corpus (Vatri and McGillivray 2018), which contains a large collection of automatically and carefully lemmatized and POS tagged texts released by the same research group. The authors used expert-assigned sense labels for a small subset of words, presenting improvements over the previous state-of-the-art. In 2020, the follow-up work by Vatri and McGillivray (2020) benchmarked major lemmatizers (CLTK, GLEM) and datasets (Diorisis Corpus and the Lemmatized Ancient Greek Texts repository) against three highly proficient readers of ancient Greek. The most accurate labels came from the Diorisis corpus and the CLTK backoff lemmatizer. In 2020, Keersmaekers (2020) used a random forest to perform the semantic parsing of the Ancient Greek Dependency Treebanks, Harrington Trees, and Pedalion Treebanks with high accuracy. In the same year, Palladino, Karimi, and Mathiak (2020) introduced a CRF model for named entity recognition, based on  $n$ -gram features close to the target word and POS information. The model was evaluated on Herodotus' "Histories" (in ancient Greek) discovering ethnonyms and place names.

## 7. Textual Criticism

### 7.1 Stylometrics

*Stylometric analysis* attempts to statistically quantify the linguistic features of authorial style (Holmes 1998). In 2019, Gianitsos et al. (2019) introduced a stylometric feature-set for ancient Greek enabling the identification of texts as either prose or verse using a Random Forest classifier. The feature-set included several primarily syntactic features. Then, the authors classified a selection of the verses as belonging to either the epic or the drama genre. In an effort to better understand stylometric patterns, Ochab and Essler (2019) used different unsupervised clustering methods to group the authors of ancient Greek papyri on the basis of their stylistic features. Two years later, Alqasemi et al. (2021) compared the performance of a neural network, an SVM, and a decision tree for classifying different poetic metres occurring in Arabic poetry.

### 7.2 Stemmatology

The goal of *computational stemmatology* is to reconstruct the genealogy of different versions of a text, in order to obtain a text as close as possible to the authorial original (Roos and Heikkilä 2009). In 2010, Roelli and Bachmann (2010) computed the Character

Edit Distance between text strings from different versions of the Latin “*Dialogus contra Iudaeos*” by Petrus Alfonsi. The distances were used to produce a distance matrix and tree graphs visualizing the evolution of different parts of the text. In 2016, Koppel, Michaely, and Tal (2016) introduced a method based on expectation-maximization; given multiple corrupted versions of the same text, they aimed to reconstruct the authorial original. The method was applied to artificially generated manuscripts and the Talmud, showing how automated methods for reconstruction can be more effective than a naive majority rule. More recently, Jones, Romano, and Mohd (2022) cast the problem of stemmatology as a classification task. More specifically, using verses from Greek New Testament manuscripts with slight variations, they proposed a feature-set to identify whether a given verse belonged to the “gold standard” (the authorial original) or to a variant.

### 7.3 Intertextuality

Authors often convey meaning by referring to or imitating another text (e.g., prior works of literature), a process that creates complex networks of literary relationships, known as *intertextuality*. In recent years, computational approaches have introduced quantitative measures to aid large-scale analyses (Dexter et al. 2017).

Most research on this topic has focused on Latin. Early efforts worked on string-matching approaches and on the identification of lexical correspondences (Coffee et al. 2012a, b; Scheirer, Forstall, and Coffee 2016). In 2011, Forstall, Jacobson, and Scheirer (2011) used an SVM with character bi- and tri-grams and word bi-grams to determine to what extent, if any, the classical Roman poet Catullus had influenced the 8th century CE Latin poem “*Angustae Vitae*” by Paul the Deacon. The results showed notable stylistic similarities between two poems of Catullus and the “*Angustae Vitae*.” Bernstein, Gervais, and Lin (2015) computed the word frequencies of the Tesseræ corpus comprising over 300 works of Latin literature to identify instances where short passages, written between 1st century BCE and 6th century CE, shared two or more repeated words. Bjerva and Praet (2015, 2016) used word embeddings to analyse Cassiodorus’ “*Variae*,” a corpus of hundreds of state letters. The authors used word2vec and network analysis to find associations between Latin and Greek authors and a selection of ideological concepts (“liberty,” “antiquity,” “modern,” “Greekness” or “Romanness”). Focusing on the Roman authors Seneca and Livy, Dexter et al. (2017) proposed different stylometric features to distinguish citational material, including non-content words (e.g., articles, prepositions), syntactic constructions, and the length of sentences and clauses. They then used an SVM to identify the citational and non-citation material Livy might have loosely appropriated from earlier sources. Burns et al. (2021) compared different implementations of both word2vec and fastText on the CLTK-lemmatized (Johnson et al. 2021) “*Argonautica*” by Valerius Flaccus. By comparing the cosine similarities of bi-gram pairs, they showed that embeddings could enhance Latin intertextual detection, and produce state-of-the-art results.

Research on intertextuality has also been carried out on Biblical texts in Greek: Lee (2007) studied text reuse (“source alternation patterns”) in the New Testament. Considering the Gospel of Luke as the target text and the Gospel of Mark as the source text, the authors introduced a model for sentence-level quantitative text-reuse discovery. The model’s predictions were fine-tuned and evaluated against scholarly hypotheses, demonstrating the model’s ability to capture the researchers’ expert understanding of text reuse. Moritz et al. (2016) presented a linguistic analysis of text reuse in non-literal translations of Bible verses in ancient Greek and Latin. The authors used hundreds of

reused verse pairs, and used lexical databases of semantic relations of words and lemmas, together with POS information, to identify reuse. Their results showed that simple pre-processing, such as stemming and lemmatizing, may not be sufficient to capture the richness of the qualitative manual analysis. Shifting to ancient Greek literature, Büchler et al. (2012) studied text-reuse in Athenaeus' "Deipnosophistai": Editors have explicitly marked hundreds of instances of text being quoted or paraphrased from the Homeric epic poems. Using uni- and bi-gram frequency features and a wide window to preserve locality, the authors identified nearly all references annotated by editors. Finally, Monroe (2018) used frequencies of cuneiform signs to study the scholarly practices behind the composition of damaged and fragmentary examples of late Babylonian astrology.

## 7.4 Sentiment Analysis

*Sentiment analysis* is an NLP task where the goal is to extract subjective information and affective states from a text, for example, whether the text expresses positive, negative, or neutral emotions (Medhat, Hassan, and Korashy 2014). In ancient languages, the lack of labeled data can pose an obstacle to this task. To overcome this issue, Kumar, Pathania, and Raman (2022) introduced a zero-shot method for sentiment analysis using cross-lingual data. The authors collected a dataset of 12k samples of online English–ancient Sanskrit translations, to train a Transformer model to translate from Sanskrit to English. An additional GAN loss was used to improve the quality of the translations. Finally, the sentiment of the resulting English translations was classified with high accuracy using an RNN model. Pavlopoulos, Xenos, and Picca (2022) showed that the linguistic expression of sentiments may diverge between ancient and modern Greek. The authors annotated the sentiment of verses from the first Book of Iliad (translated into modern Greek) and fine-tuned Greek BERT on the task.

## 8. Translation and Decoding

### 8.1 Decipherment

Ancient texts are usable only in proportion to their intelligibility, but many ancient languages and scripts remain undeciphered (Robinson 2009). **Deciphering** an ancient written language involves understanding the original meaning of words in their context, often using descended and cognate languages or multilingual keys as aids.

Early statistical techniques focused on reconstructing linguistic structures. More specifically, Rao et al. (2009a, 2010) compared the statistical structure of sign sequences in the Indus script to those of a representative group of languages: Sumerian, Old Tamil, Vedic Sanskrit, English words and characters, and non-linguistic systems such as DNA and protein sequences. Using conditional entropy, they showed that Indus script inscriptions have an increased probability of representing language. Using the same corpus, Rao et al. (2009b) computed pairwise statistics using a Markov model. Their work suggested that specific signs often occur at the beginning of Indus script inscriptions and that, for any sign, there are other signs that have a high probability of occurring after. Such syntax patterns could pave the way to decipherment. Their Markov model was also applied to textual restoration, and Yadav et al. (2010) used it for further  $n$ -gram analysis of the Indus script. Their model could restore signs with a high accuracy. However, the statistical approach of these works was challenged by Sproat (2010), and Sproat (2014) later introduced a novel measure based on repetition turn

out, classifying the data for the Indus Valley script as a non-linguistic symbol system, thereby contradicting those earlier works.

The work of Snyder, Barzilay, and Knight (2010) focused on the alphabetic mappings and translations of Ugaritic words to their corresponding cognates in Hebrew. Using a non-parametric Bayesian model, they estimated distributions over bilingual morpheme pairs and assigned probability based on recurrent patterns: Each character in one language would map to a small number of characters in the other. The accuracy of cognate translations was measured with respect to complete word forms and morphemes. Berg-Kirkpatrick and Klein (2011) modeled the same problem as a combinatorial optimization, minimizing the edit-distance between a source word and target word, given alphabetical sign matching. Their results were better than Snyder, Barzilay, and Knight (2010) in cognate word accuracy, but lower in alphabet accuracy. The same model was also used to identify phonetic cognates between Spanish, Portuguese, and Italian. In 2013, Bouchard-Côté et al. (2013) presented a probabilistic model of sound change for reconstructing words occurring in the proto-languages from which modern Austronesian languages evolved. Over 85% of the system's reconstructions were within one character of the manual reconstruction provided by a linguist. In 2019, Luo, Cao, and Barzilay (2019) introduced a more general approach for automated decipherment based on a sequence-to-sequence neural network model, NeuroCipher, which captured character-level correspondences between cognates using optimization. NeuroCipher was used to map Ugaritic to Hebrew and Linear B to ancient Greek. In 2021, Luo et al. (2021) presented a model for deciphering unsegmented languages using phonetic conversion. The model was able to identify related known languages, and was used to extract cognates from undersegmented texts in Gothic, Ugaritic, and the undeciphered Iberian scripts.

Using images as inputs, Daggumati and Revesz (2018) used a CNN with an SVM to generate similarity matrices and map linguistic family trees, showing that Indus script is visually close to Sumerian pictographs, while the Linear B script is close to the Cretan Hieroglyphic script. In a similar setting, de Lima-Hernandez and Vergauwen (2021) used a CNN to show that the Phoenician alphabet is much closer to the Indus script than to the Brahmi script. Recent studies such as Karajgikar, Al-Khulaidy, and Barea (2021) have carried out computational analyses using  $n$ -grams and word2vec embeddings on undeciphered scripts such as Linear A, and also tried to group symbols. Papavassiliou, Owens, and Kosmopoulos (2020) increased the amount of data by including related writing systems, such as Linear B, which could be the key to solving decipherment challenges. Recently, Corazza et al. (2022) introduced Sign2Vec, an unsupervised clustering method for analysing signs from 200 inscriptions in the undeciphered Cypro-Minoan syllabary. Sign2Vec used  $k$ -means on the outputs of a ResNet50, and incorporated additional contextual information from the surrounding signs, classifying two out of three signs correctly.

## 8.2 Machine Translation

The *translation* of ancient texts takes us an interpretative step closer to the mentality and *milieux* of ancient authors. Recent efforts of neural machine translation have allowed historians to harness all available data to create automated pipelines for ancient languages.

For Sumerian, in 2018 Chiarcos et al. (2018) presented a dictionary- and rule-based method for the morphological and syntactic annotation of administrative texts pertaining to the third Ur dynasty. The dataset was then used by Punia et al. (2020) to create

the first machine translation system for Sumerian transliterations. The authors used a stacked RNN sequence-to-sequence with GloVe embeddings and a Transformer model. Two human experts were asked to score 50 translations generated by each model. The problem of automatic transliteration of glyphs into Latin script was approached by Gordin et al. (2020), evaluating multiple models on 23k Neo-Assyrian cuneiform tablets from the Oracc dataset. The highest transliteration and segmentation accuracy was achieved using a bidirectional RNN model.

Zhang, Li, and Su (2019) proposed a bidirectional RNN sequence-to-sequence model for translating old Chinese documents into contemporary Chinese and vice versa. The model had a copying mechanism and local attention. Using only a small sentence-aligned corpus of 4k pairs, the authors addressed the matter of limited aligned corpora by introducing an unsupervised sentence alignment model using dynamic programming. However, the semantics of ancient Chinese are complex—for example, word polysemy introduces a one-to-many alignment with modern Chinese. Yang et al. (2021) showed that a BLEU (Bilingual Evaluation Understudy) score could not identify potentially correct translation results. Inspired by unsupervised dual learning, the authors introduced a Dual-based Translation Evaluation, able to evaluate the one-to-many alignment of ancient Chinese, and outperform BLEU in a human expert evaluation.

Park et al. (2020) presented an attention RNN and a Transformer-based model for ancient Korean translation. The authors used a shared vocabulary, byte pair encoding, and  $n$ -gram decoding. Using a processed dataset crawled from the Institute for the Translation of Korean Classics, the Transformer model performed better when combined with the RNN. On the same dataset, Park et al. (2022) presented a model using bilingual sub-word embedding initialization and priming, inspired by the cognitive science theory that two different stimuli influence each other. Their RNN model surpassed the previous transformer results. Furthermore, Kang et al. (2021) worked on translating and restoring the Hanja historical records of the Annals of Joseon Dynasty into old Korean using a Transformer-based model, which achieved fluent translations. Follow up research from Son et al. (2022), supports both translation into contemporary Korean and into English and uses a newer version of the Annals of Joseon Dynasty corpus.

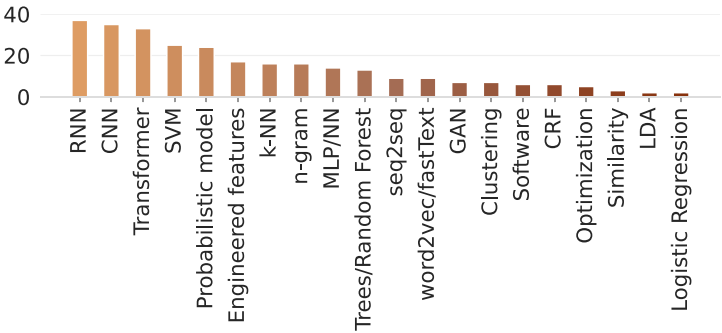
Finally, Yousef et al. (2022) fine-tuned a pre-trained multilingual BERT-based language model to automatically translate ancient Greek to Latin texts following a novel alignment workflow.

## 9. Conclusions

In this survey, we set out to examine all interdisciplinary machine learning contributions to the study of ancient languages to date. While reviewing the literature, we identified a recurring set of factors that are either driving research or posing challenges to be overcome.

### 9.1 Impact and Data Availability

The increased availability of digitized, linked, open, and rich data for ancient languages has been recognized as the *sine qua non* condition for advancing machine learning research for ancient languages. Many such datasets are created and exploited in the context of conferences and competitions (such as ICFHR, ICDAR, CoNLL, LT4HALA). At the same time, large datasets paired with large-scale models such as Transformer-based architectures have resulted in significant improvements over traditional approaches, allowing a scale and precision unattainable by human researchers alone. To support



**Figure 3**  
Distribution of machine learning model architectures (with  $\geq 2$  articles per architecture). Under Engineered Features we’ve grouped works using PCA, HOG, HOOSC, word frequencies, and other similar methods for further analysis. Under Software, we include all methods using third party or standalone software.

this momentum, standardized data encoding (Bodard 2010) in accordance with Findable, Accessible, Interoperable, Reusable principles (Wilkinson et al. 2016) is crucial to advancing future research. Indeed, our evaluation has also shown that the adoption of shared data standards (as seen in certain works) successfully fosters a more scientific approach to evaluation and metrics, which are vital to tracking progress and impact in machine learning research.

9.2 Machine Learning Observations

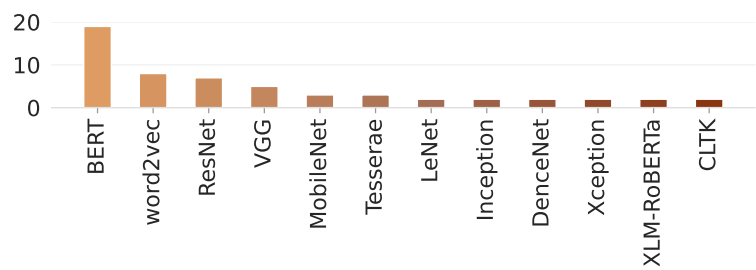
In this survey we analyzed over 230 interdisciplinary works. The majority, 149 in total, utilized textual inputs, while 59 operated on visual inputs and 18 on both modalities.<sup>2</sup> Out of the works reviewed, 137 used supervised learning, 33 were self-supervised, and 26 used unsupervised or weakly supervised methods. In Figure 3 we present the distribution of machine learning model architectures utilized: 117 studies used deep learning architectures, 66 used machine learning, and 42 used statistical models. It is particularly noteworthy that several works used existing architectures, such as computer vision or language models, that were retrained to solve new tasks. This is illustrated in Figure 4: One may also note that, among others, BERT, word2vec ResNet, and VGG exhibit a substantial uptake. A subset of 36 works used existing pre-trained models, which once again goes to demonstrate the impact of open-source pre-trained models on such research. We refer the reader to Appendix A for further details.

9.3 Future Research

Future research should address the extant challenges. Firstly, machine learning methods are quintessentially data-dependent, and all major breakthroughs surveyed in this article build upon digitization and labeling efforts—which should therefore be prioritized and rewarded. At the same time, given the current extent of unlabeled data, it

<sup>2</sup> The numbers may not add up to the total number of studies due to some being reviews or summaries of competitions.





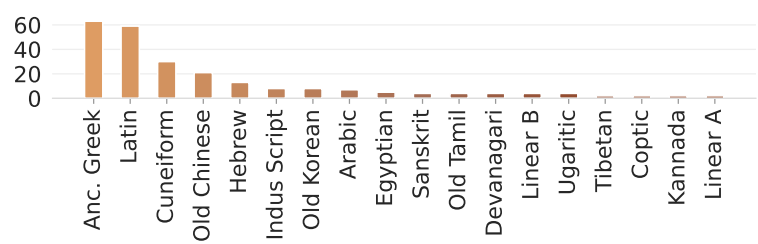
**Figure 4**  
Distribution of existing machine learning models utilized (with  $\geq 2$  articles per model).

would be auspicious to explore the potential of pre-trained large-scale foundation models, further fine-tuned to the tasks addressed in our taxonomy.

Secondly, some of the most impactful works reviewed were those developed by interdisciplinary teams bringing together computer scientists and historians, linguists, or subject-specific specialists. This can easily be appreciated by the more thoughtful experiment designs, the use of accurate terminology, and the overall better results achieved in the reports (e.g., Tracy and Papaodyseus 2009; Popović, Dhali, and Schomaker 2021; Assael et al. 2022). Multidisciplinary teams may more effectively address the challenges posed to machine learning methods by ancient writing systems, as they will be better informed of the idiosyncrasies of the textual material, more aware of the machine learning techniques best suited to address them, and will devote themselves to veritably worthwhile research questions. Moreover, it is only through such interdisciplinary collaborations that greater trust in digital methods may be built within scholarly communities in the humanities on one hand, and on the other hand the truly pressing questions and challenges posed by ancient texts might be more meticulously addressed by computer scientists.

Thirdly, it is a commonly acknowledged fact that ground truths are unattainable when dealing with ancient texts, as the original written form (physical or textual), exact date and place of writing, and so forth, could have been lost over the centuries. One can only test a model’s predictions exclusively against the assumptions of experts, a situation where “data circularity” might arise, whereby existing scholarly conjectures are included within the training set. We found the studies that did acknowledge this situation were particularly insightful, and hopefully will motivate further research harnessing machine learning methods for denoising and debiasing data. On that note, imbalanced datasets are known to introduce bias, prejudice, and unfairness, which may perpetuate systemic bias, obfuscate evidence, or point to misleading patterns in the data. This review has highlighted how not all languages, histories, or geographies are equally represented (Figure 5) in the field under review, and this lack of representation may result in “digital colonialism” (McGillivray et al. 2020). This remains an active area of research in AI ethics and in studies of the ancient world.

Fourthly, historians are constantly seeking novel methodological aids to advance their research, and should therefore be open to the opportunities offered by technology. In tandem, scientists working with ancient texts should direct their efforts towards augmenting the interpretability of their model’s results, rather than on merely maximizing metrics. We furthermore found that comparing prior literature of our taxonomy was especially challenging due to the lack of standardized benchmarks and the constant use



**Figure 5** Distribution of publications per language (with  $\geq 2$  articles per language). Under “cuneiform” script, we include the Akkadian, Sumerian, and Babylonian languages.

of different datasets. Such inconsistencies hamper the ability to draw clear conclusions and evaluate the true progress of different approaches. Moreover, the absence of universally accepted evaluation metrics and benchmarks further complicates the process of comparing the performance of models, as the results are often not directly comparable. This could ultimately slow down the advancement of research, as it becomes more challenging for researchers to identify promising directions, replicate results, and build upon previous work in a reliable and transparent manner. Thus, we’ve included an Appendix section tracking the uptake of different models per year, which confirms our assumptions expounded in Section 2. We hope that this survey will be adopted as a reference point for prior work and help bridge this gap.

Finally, and building upon that point, it is essential to emphasize that progress in machine learning relies not only on models, but also on the quality and quantity of data, metrics, and evaluation. We wish to emphasize: (a) the direct correlation between the characteristics of a dataset and a model’s performance, and (b) the importance of robust hypothesis testing, with data partitioning (train, test and validation sets), or data resampling to train different models and statistically analyze generalizability (e.g., cross-validation).

9.4 The Value of Interdisciplinarity

To conclude, the synergy between the study of ancient languages and machine learning achieves its full potential when historians and scientists work together to identify the problems and find the solutions best tailored to the ancient data’s idiosyncrasies. In this review, we set out to map a nascent field and highlight the scholarly benefits of collaboration between two seemingly unrelated disciplines. Our review has determined that machine learning for ancient languages is not only a well-established field with its own research questions, but holds significant potential for the large scale and scientific exploration of a wide-range of historical questions, and in doing so can open up new areas of research.

Appendix A. Taxonomy Analysis: Modalities used per Section

**Table A.1**  
Summary of each taxonomy section, showing the number of works using the text and visual modalities.

Section	Text	Visual	Both
Authorship attribution	12	0	0
Chronological attribution	5	4	1
Decipherment	10	1	2
Fragment reassembly	1	4	0
Geographical attribution	2	0	0
Intertextuality	12	0	0
Language identification	8	0	0
Machine translation	9	1	0
POS tagging and Parsing	45	0	1
Palaeographic analysis and writer identification	0	22	1
Quality enhancement	0	9	0
Recognition	0	17	12
Representation learning	6	0	0
Semantics	7	0	0
Sentiment analysis	2	0	0
Stemmatology	3	0	0
Stylometrics	3	0	0
Textual restoration	10	0	0
Topic modeling, genre detection	5	0	0
Word segmentation and boundary detection	9	1	1

Appendix B. Taxonomy Analysis: Model Types per Year

**Table B.1**  
Summary of the different machine learning model types employed per year. Models that were used 2 or less times were omitted. Under Engineered Features we've grouped works using PCA, HOG, HOOSC, word frequencies, and other similar methods for further analysis. Under Software, we include all methods using third party or standalone software.

Model	2000–2010	2010–2015	2015–2020	2020–2023	Total
CNN	0	0	10	25	35
CRF	0	2	2	2	6
Clustering	0	1	2	4	7
Engineered features	0	1	13	3	17
GAN	0	0	0	7	7
k-NN	0	2	9	5	16
MLP/NN	1	3	6	4	14
Optimization	0	2	3	0	5
Probabilistic model	2	6	9	7	24
RNN	0	1	20	16	37
SVM	0	4	14	7	25
Similarity	0	1	1	1	3
Software	1	2	2	1	6
Transformer	0	0	3	30	33
Trees/Random Forest	0	1	5	7	13
<i>n</i> -gram	2	6	4	4	16
seq2seq	0	0	3	6	9
word2vec/fastText	0	0	5	4	9

Downloaded from [http://direct.mit.edu/col/article-pdf/49/3/703/217413/col\\_a\\_00481.pdf](http://direct.mit.edu/col/article-pdf/49/3/703/217413/col_a_00481.pdf) by guest on 12 October 2024

Appendix C. Taxonomy Analysis: Ancient Languages Researched per Year

**Table C.1**  
Summary of the different ancient languages researched per year. Languages that were used 2 or less times were omitted. Finally the results may not sum to the total number of papers as we include review works in our literature.

Language	2000–2010	2010–2015	2015–2020	2020–2022	Total
Anc. Greek	6	4	32	21	63
Arabic	0	3	3	1	7
Cuneiform	0	5	14	11	30
Devanagari	0	0	1	3	4
Egyptian	0	1	1	3	5
Hebrew	0	2	8	3	13
Indus Script	3	2	2	1	8
Latin	3	6	29	21	59
Linear B	0	0	2	2	4
Old Chinese	0	1	6	14	21
Old Korean	0	0	0	8	8
Old Tamil	0	0	3	1	4
Sanskrit	0	0	3	1	4
Ugaritic	0	2	1	1	4

## Acknowledgments

The authors would like to thank Çağlar Gülçehre and Francesco Nori for their helpful comments and advice on this article. TS acknowledges that this project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement no. 101026185.

## References

- Abdelhaleem, Alaa, Ahmed Droby, Abdelkader Asi, Majeed Kassis, Reem Al Asam, and Jihad El-sanaa. 2017. WAHD: A database for writer identification of Arabic historical documents. In *International Workshop on Arabic Script Analysis and Recognition (ASAR)*, pages 64–68. <https://doi.org/10.1109/ASAR.2017.8067761>
- Abitbol, Roy, Ilan Shimshoni, and Jonathan Ben-Dov. 2021. Machine learning based assembly of fragments of ancient papyrus. *Journal on Computing and Cultural Heritage (JOCCH)*, 14(3):1–21. <https://doi.org/10.1145/3460961>
- Adam, Kalthoum, Asim Baig, Somaya Al-Maadeed, Ahmed Bouridane, and Sherine El-Menshawy. 2018. KERTAS: Dataset for automatic dating of ancient Arabic manuscripts. *International Journal on Document Analysis and Recognition (IJ DAR)*, 21(4):283–290. <https://doi.org/10.1007/s10032-018-0312-3>
- Alqasemi, Fahd, Salah AL-Hagree, Nail Adeeb Ali Abdu, Baligh Al-Helali, and Ghaleb Al-Gaphari. 2021. Arabic poetry meter categorization using machine learning based on customized feature extraction. In *International Conference on Intelligent Technology, System and Service for Internet of Everything (ITSS-IOE)*, pages 1–4. <https://doi.org/10.1109/ITSS-IOE53029.2021.9615302>
- An, Bo and Congjun Long. 2021. Ancient Tibetan word segmentation based on deep learning. In *International Conference on Asian Language Processing (IALP)*, pages 292–297. IEEE. <https://doi.org/10.1109/IALP54817.2021.9675149>
- Arabadjis, DDimitrios, Fotios Giannopoulos, Michail Panagopoulos, Michail Exarchos, Christopher Blackwell, and Constantin Papaodysseus. 2019. A general methodology for identifying the writer of codices. Application to the celebrated “twins.” *Journal of Cultural Heritage*, 39:186–201. <https://doi.org/10.1016/j.culher.2019.04.002>
- Arabadjis, Dimitris, Fotios Giannopoulos, Constantin Papaodysseus, Solomon Zannos, Panayiotis Rousopoulos, Mihalios Panagopoulos, and Christopher Blackwell. 2013. New mathematical and algorithmic schemes for pattern classification with application to the identification of writers of important ancient documents. *Pattern Recognition*, 46(8):2278–2296. <https://doi.org/10.1016/j.patcog.2013.01.019>
- Asi, Abdelkadir, Alaa Abdalhaleem, Daniel Fecker, Volker Märgner, and Jihad El-Sana. 2017. On writer identification for Arabic historical manuscripts. *International Journal on Document Analysis and Recognition (IJ DAR)*, 20(3):173–187. <https://doi.org/10.1007/s10032-017-0289-3>
- Assael, Yannis, Thea Sommerschild, and Jonathan Prag. 2019. Restoring ancient text using deep learning: A case study on Greek epigraphy. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 6368–6375. <https://doi.org/10.18653/v1/D19-1668>
- Assael, Yannis, Thea Sommerschild, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283. <https://doi.org/10.1038/s41586-022-04448-z>, PubMed: 35264762
- Bacon, Geoff. 2020. Data-driven choices in neural part-of-speech tagging for Latin. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 111–113.
- Bamman, David and Patrick J. Burns. 2020. Latin BERT: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*.
- Bamman, David and Gregory Crane. 2011. Measuring historical word sense variation. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 1–10. <https://doi.org/10.1145/1998076.1998078>
- Barucci, Andrea, Costanza Cucci, Massimiliano Franci, Marco Loschiavo, and Fabrizio Argenti. 2021. A deep learning approach to ancient Egyptian hieroglyphs classification. *IEEE Access*, 9:123438–123447. <https://doi.org/10.1109/ACCESS.2021.3110082>
- Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A

- review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828. <https://doi.org/10.1109/TPAMI.2013.50>, PubMed: 23787338
- Benites de Azevedo e Souza, Fernando, Pius von Däniken, and Mark Cieliebak. 2019. TwistBytes - Identification of Cuneiform languages and German dialects at VarDial 2019. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 194–201.
- Berg-Kirkpatrick, Taylor and Dan Klein. 2011. Simple effective decipherment via combinatorial optimization. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 313–321.
- Bernier-Colborne, Gabriel, Cyril Goutte, and Serge Léger. 2019. Improving Cuneiform language identification with BERT. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 17–25. <https://doi.org/10.18653/v1/W19-1402>
- Bernstein, Neil, Kyle Gervais, and Wei Lin. 2015. Comparative rates of text reuse in classical Latin hexameter poetry. *DHQ: Digital Humanities Quarterly*, 9(3).
- Bhat, Riyaz Ahmad, Irshad Bhat, and Srinivas Bangalore. 2018. The SLT-interactions parsing system at the CoNLL 2018 shared task. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 153–159.
- Bhurke, Shubham S., Vina M. Lomte, Pranay M. Kolhe, and Akshay U. Pednekar. 2020. Survey on Sanskrit script recognition. In *International Conference on Mobile Computing and Sustainable Informatics*, pages 771–782. [https://doi.org/10.1007/978-3-030-49795-8\\_73](https://doi.org/10.1007/978-3-030-49795-8_73)
- Bjerva, Johannes and Raf Praet. 2015. Word embeddings pointing the way for Late Antiquity. In *SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 53–57. <https://doi.org/10.18653/v1/W15-3708>
- Bjerva, Johannes and Raf Praet. 2016. Rethinking intertextuality through a word-space and social network approach – the case of Cassiodorus. *Journal of Data Mining and Digital Humanities*, pages 1–25.
- Blackburn, Patrick and Johannes Bos. 2005. *Representation and Inference for Natural Language: A First Course in Computational Semantics*, Center for the Study of Language and Information Stanford.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.
- Bodard, Gabriel. 2010. EpiDoc: Epigraphic documents in XML for publication and interchange. *Latin on Stone: Epigraphic Research and Electronic Archives*, pages 101–18.
- Bogacz, Bartosz, Maximilian Klingmann, and Hubert Mara. 2017. Automating transliteration of cuneiform from parallel lines with sparse data. In *IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 615–620. <https://doi.org/10.1109/ICDAR.2017.106>
- Bogacz, Bartosz and Hubert Mara. 2020. Period classification of 3D cuneiform tablets with geometric neural networks. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 246–251. <https://doi.org/10.1109/ICFHR2020.2020.00053>
- Bogacz, Bartosz and Hubert Mara. 2022. Digital Assyriology—Advances in visual cuneiform analysis. *Journal on Computing and Cultural Heritage (JOCCH)*, 15(2):1–22. <https://doi.org/10.1145/3491239>
- Boroş, Tiberiu, Ștefan Daniel Dumitrescu, and Ruxandra Burtica. 2018. NLP-Cube: End-to-end raw text processing with neural networks. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179.
- Bouchard-Côté, Alexandre, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences (PNAS)*, 110(11):4224–4229. <https://doi.org/10.1073/pnas.1204678110>, PubMed: 23401532
- Bracco, Giovanni, Silvio Migliori, Giorgio Mencuccini, Daniela Alderuccio, and Giovanni Ponti. 2013. Data mining tools and GRID infrastructure for Assyriology text analysis (an Old-Babylonian situation studied through text analysis and data mining tools). In *RAI - Rencontre Assyriologique Internationale - Private and State in the Ancient Near East*, pages 82–88.
- Brandenbusch, Kai, Eugen Rusakov, and Gernot A. Fink. 2021. Context aware generation of cuneiform signs. In *International Conference on Document Analysis and Recognition*, pages 65–79.

- <https://doi.org/10.1007/978-3-030-86549-8.5>
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.
- Büchler, Marco, Gregory Crane, Maria Moritz, and Alison Babeu. 2012. Increasing recall for text re-use in historical documents to support research in the Humanities. In *International Conference on Theory and Practice of Digital Libraries*, pages 95–100. [https://doi.org/10.1007/978-3-642-33290-6\\_11](https://doi.org/10.1007/978-3-642-33290-6_11)
- Burns, Patrick J., James Brofos, Kyle Li, Pramit Chaudhuri, and Joseph P. Dexter. 2021. Profiling of intertextuality in Latin literature using word embeddings. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4900–4907. <https://doi.org/10.18653/v1/2021.naacl-main.389>
- Can, Gülcan, Jean-Marc Odobez, and Daniel Gatica-Perez. 2016. Evaluating shape representations for Maya glyph classification. *Journal on Computing and Cultural Heritage (JOCCH)*, 9(3):1–26. <https://doi.org/10.1145/2905369>
- Celano, Giuseppe G. A. 2020. A gradient boosting-seq2seq system for Latin POS tagging and lemmatization. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 119–123.
- Celano, Giuseppe G. A., Gregory Crane, and Saeed Majidi. 2016. Part of speech tagging for ancient Greek. *Open Linguistics*, 2(1):393–399. <https://doi.org/10.1515/opli-2016-0020>
- Chammas, Michel, Abdallah Makhoul, Jacques Demerjian, and Elie Dannaoui. 2022. A deep learning based system for writer identification in handwritten Arabic historical manuscripts. *Multimedia Tools and Applications*, pages 1–16. <https://doi.org/10.1007/s11042-022-12673-x>
- Chang, Xiang, Fei Chao, Changjing Shang, and Qiang Shen. 2022. Sundial-GAN: A cascade generative adversarial networks framework for deciphering Oracle Bone inscriptions. In *ACM International Conference on Multimedia*, pages 1195–1203. <https://doi.org/10.1145/3503161.3547925>
- Che, Wanxiang, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64.
- Chen, Danlu, Mengxiao Lin, Zhifeng Hu, and Xipeng Qiu. 2018. A simple yet effective joint training method for cross-lingual universal dependency parsing. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 256–263. <https://doi.org/10.18653/v1/K18-2026>
- Cheng, Ning, Bin Li, Liming Xiao, Changwei Xu, Sijia Ge, Xingyue Hao, and Minxuan Feng. 2020. Integration of automatic sentence segmentation and lexical analysis of ancient Chinese based on BiLSTM-CRF model. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 52–58.
- Chiarcos, Christian, Ilya Khait, Émilie Pagé-Perron, Niko Schenk, Christian Fäth, Julius Steuer, William Mcgrath, and Jinyan Wang. 2018. Annotating a low-resource language with LLOD technology: Sumerian morphology and syntax. *Information*, 9(11):290. <https://doi.org/10.3390/info9110290>
- Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Christlein, Vincent, Angelos Nicolaou, Mathias Seuret, Dominik Stutzmann, and Andreas Maier. 2019. ICDAR 2019 competition on image retrieval for historical handwritten documents. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1505–1509. <https://doi.org/10.1109/ICDAR.2019.00242>
- Chung, Junyoung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Advances in Neural Information Processing Systems Workshop on Deep Learning*.
- Coffee, Neil, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde, and Sarah L. Jacobson. 2012a. The Tesseræ Project: Intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, 28(2):221–228. <https://doi.org/10.1093/l1c/fqs033>



- Coffee, Neil, Jean-Pierre Koenig, Shakthi Poornima, Roelant Ossewaarde, Christopher Forstall, and Sarah Jacobson. 2012b. Intertextuality in the digital age. *Transactions of the American Philological Association*, pages 383–422. <https://doi.org/10.1353/apa.2012.0010>
- Collins, Tim, Sandra I. Woolley, Luis Hernandez Munoz, Andrew Lewis, Eugene Ch'ng, and Erlend Gehlken. 2014. Computer-assisted reconstruction of virtual fragmented cuneiform tablets. In *International Conference on Virtual Systems & Multimedia (VSMM)*, pages 70–77. <https://doi.org/10.1109/VSMM.2014.7136691>
- Corazza, Michele, Fabio Tamburini, Miguel Valério, and Silvia Ferrara. 2022. Unsupervised deep learning supports reclassification of Bronze age cypriot writing system. *PLOS ONE*, 17(7):1–22. <https://doi.org/10.1371/journal.pone.0269544>, PubMed: 35834491
- Corbara, Silvia, Alejandro Moreo, and Fabrizio Sebastiani. 2022. Syllabic quantity patterns as rhythmic features for Latin authorship attribution. *Journal of the Association for Information Science and Technology*, 74(1):128–141. <https://doi.org/10.2139/ssrn.4214742>
- Daggumati, Shruti and Peter Z. Revesz. 2018. Data mining ancient script image data using convolutional neural networks. In *International Database Engineering & Applications Symposium*, pages 267–272. <https://doi.org/10.1145/3216122.3216163>
- Davis, Tom. 2007. The practice of handwriting identification. *Library*, 8(3):251–276. <https://doi.org/10.1093/library/8.3.251>
- de Lhoneux, Miryam, Sara Stymne, and Joakim Nivre. 2017. Arc-hybrid non-projective dependency parsing with a static-dynamic oracle. In *International Conference on Parsing Technologies (IWPT)*, pages 99–104.
- de Lima-Hernandez, Roberto and Maarten Vergauwen. 2021. A generative and entropy-based registration approach for the reassembly of ancient inscriptions. *Remote Sensing*, 14(1):6. <https://doi.org/10.3390/rs14010006>
- De Stefano, Claudio, Marilena Maniaci, Francesco Fontanella, and A. Scotto di Freca. 2018. Reliable writer identification in medieval manuscripts through page layout features: The “Avila” Bible case. *Engineering Applications of Artificial Intelligence*, 72:99–110. <https://doi.org/10.1016/j.engappai.2018.03.023>
- Demilew, Fitehalew Ashagrie and Boran Sekeroglu. 2019. Ancient Geez script recognition using deep learning. *SN Applied Sciences*, 1(11):1–7. <https://doi.org/10.1007/s42452-019-1340-4>
- Dencker, Tobias, Pablo Klinkisch, Stefan M. Maul, and Björn Ommer. 2020. Deep learning of cuneiform sign detection with weak supervision using transliteration alignment. *PLOS ONE*, 15(12):e0243039. <https://doi.org/10.1371/journal.pone.0243039>, PubMed: 33326435
- Devi, S. Gayathri, Subramaniaswamy Vairavasundaram, Yuvaraja Teekaraman, Ramya Kuppusamy, and Arun Radhakrishnan. 2022. A deep learning approach for recognizing the cursive Tamil characters in palm leaf manuscripts. *Computational Intelligence And Neuroscience*, 2022. <https://doi.org/10.1155/2022/3432330>
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186.
- Dexter, Joseph P., Theodore Katz, Nilesh Tripuraneni, Tathagata Dasgupta, Ajay Kannan, James A. Brofos, Jorge A. Bonilla Lopez, Lea A. Schroeder, Adriana Casarez, Maxim Rabinovich, Ayelet Haimson Lushkov, and Pramit Chaudhuri. 2017. Quantitative criticism of literary relationships. *Proceedings of the National Academy of Sciences (PNAS)*, 114(16):E3195–E3204. <https://doi.org/10.1073/pnas.1611910114>, PubMed: 28373557
- Dhali, Maruf A., Sheng He, Mladen Popović, Eibert Tigchelaar, and Lambert Schomaker. 2017. A digital palaeographic approach towards writer identification in the Dead Sea Scrolls. In *International Conference on Pattern Recognition Applications and Methods*, volume 2017, pages 693–702. <https://doi.org/10.5220/0006249706930702>
- Doostmohammadi, Ehsan and Minoo Nassajian. 2019. Investigating machine learning methods for language and dialect identification of cuneiform texts. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 188–193. <https://doi.org/10.18653/v1/W19-1420>

- Duthoo, Elie and Olivier Mesnard. 2018. CEA LIST: Processing low-resource languages for CoNLL 2018. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 34–44.
- Edan, Naktal M. 2013. Cuneiform symbols recognition based on *k*-means and neural network. *AL-Rafidain Journal of Computer Sciences and Mathematics*, 10(1):195–202. <https://doi.org/10.33899/csmj.2013.163436>
- Faigenbaum-Golovin, S., A. Shaus, B. Sober, D. Levin, N. Na'aman, B. Sass, E. Turkel, E. Piasetzky, and I. Finkelstein. 2016. Algorithmic handwriting analysis of Judah's military correspondence sheds light on composition of biblical texts. *Proceedings of the National Academy of Sciences (PNAS)*, 113(17):4664–4669. <https://doi.org/10.1073/pnas.1522200113>, PubMed: 27071103
- Faigenbaum-Golovin, Shira, Arie Shaus, and Barak Sober. 2022. Computational handwriting analysis of ancient Hebrew inscriptions—A survey. *IEEE BITS the Information Theory Magazine*, 2(1):90–101. <https://doi.org/10.1109/MBITS.2022.3197559>
- Fecker, Daniel, Abedelkadir Asi, Werner Pantke, Volker Märgner, Jihad El-Sana, and Tim Fingscheidt. 2014a. Document writer analysis with rejection for historical Arabic manuscripts. In *International Conference on Frontiers in Handwriting Recognition*, pages 743–748. <https://doi.org/10.1109/ICFHR.2014.130>
- Fecker, Daniel, Abedelkadir Asit, Volker Märgner, Jihad El-Sana, and Tim Fingscheidt. 2014b. Writer identification for historical Arabic documents. In *International Conference on Pattern Recognition*, pages 3050–3055. <https://doi.org/10.1109/ICPR.2014.526>
- Fetaya, Ethan, Yonatan Lifshitz, Elad Aaron, and Shai Gordin. 2020. Restoration of fragmentary Babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 117(37):22743–22751. <https://doi.org/10.1073/pnas.2003794117>, PubMed: 32873650
- Fiel, Stefan, Florian Kleber, Markus Diem, Vincent Christlein, Georgios Louloudis, Stamatopoulos Nikos, and Basilis Gatos. 2017. ICDAR2017 competition on historical document writer identification. In *IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1377–1382. <https://doi.org/10.1109/ICDAR.2017.225>
- Fiorucci, Marco, Marina Khoroshiltseva, Massimiliano Pontil, Arianna Travaglia, Alessio Del Bue, and Stuart James. 2020. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133:102–108. <https://doi.org/10.1016/j.patrec.2020.02.017>
- Firmani, Donatella, Marco Maiorino, Paolo Merialdo, and Elena Nieddu. 2018. Towards knowledge discovery from the Vatican secret archives. In *Codice Ratio - episode 1: Machine transcription of the manuscripts*. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 263–272. <https://doi.org/10.1145/3219819.3219879>
- Forstall, Christopher W., Sarah L. Jacobson, and Walter J. Scheirer. 2011. Evidence of intertextuality: Investigating Paul the Deacon's *Angustae Vitae*. *Literary and Linguistic Computing*, 26(3):285–296. <https://doi.org/10.1093/lc/fqr029>
- Forsyth, David and Jean Ponce. 2011. *Computer Vision: A Modern Approach*. Prentice Hall.
- Franken, Morris and Jan C. van Gemert. 2013. Automatic Egyptian hieroglyph recognition by retrieving images as texts. In *ACM International Conference on Multimedia*, pages 765–768, Association for Computing Machinery. <https://doi.org/10.1145/2502081.2502199>
- Gatos, Basilios, Kostas Ntzios, Ioannis Pratikakis, Sergios Petridis, Thomas Konidakis, and Stavros J. Perantonis. 2006. An efficient segmentation-free approach to assist old Greek handwritten manuscript OCR. *Pattern Analysis and Applications*, 8(4):305–320. <https://doi.org/10.1007/s10044-005-0013-7>
- Gianitsos, Efthimios, Thomas Bolt, Pramit Chaudhuri, and Joseph Dexter. 2019. Stylometric classification of ancient Greek literary texts by genre. In *SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 52–60. <https://doi.org/10.18653/v1/W19-2507>
- Goler, Sarah, James T. Yardley, David M. Ratzan, Roger Bagnall, Alexis Hagadorn, and James McInerney. 2019. Dating ancient Egyptian papyri through Raman spectroscopy: Concept and application to the fragments of the Gospel of Jesus' wife and the Gospel of John. *Journal for the*

- Study of the New Testament*, 42(1):98–133. <https://doi.org/10.1177/0142064X19855583>
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
- Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144. <https://doi.org/10.1145/3422622>
- Gordin, Shai, Gai Gutherz, Ariel Elazary, Avital Romach, Enrique Jiménez, Jonathan Berant, and Yoram Cohen. 2020. Reading Akkadian cuneiform using Natural Language Processing. *PLOS ONE*, 15(10):1–16. <https://doi.org/10.1371/journal.pone.0240511>, PubMed: 33112872
- Grave, Edouard, Piotr Bojanowski, Prakhara Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Language Resources and Evaluation Conference (LREC)*.
- Grieve, Jack. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270. <https://doi.org/10.1093/l1c/fqm020>
- Haliassos, Alexandros, Panagiotis Barmapoutis, Tania Stathaki, Stephen Quirke, and Anthony Constantinides. 2020. Classification and detection of symbols in ancient papyri. *Visual Computing for Cultural Heritage*, pages 121–140, Springer. [https://doi.org/10.1007/978-3-030-37191-3\\_7](https://doi.org/10.1007/978-3-030-37191-3_7)
- Harper, Kyle, Michael McCormick, Matthew Hamilton, Chantal Peiffert, Raymond Michels, and Michael Engel. 2020. Establishing the provenance of the Nazareth Inscription: Using stable isotopes to resolve a historic controversy and trace ancient marble production. *Journal of Archaeological Science: Reports*, 30:102228. <https://doi.org/10.1016/j.jasrep.2020.102228>
- He, Kaiping, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. <https://doi.org/10.1109/CVPR.2016.90>, PubMed: 26180094
- He, Sheng, Petros Samara, Jan Burgers, and Lambert Schomaker. 2016b. Image-based historical manuscript dating using contour and stroke fragments. *Pattern Recognition*, 58:159–171. <https://doi.org/10.1016/j.patcog.2016.03.032>
- Hellwig, Oliver. 2015. Morphological disambiguation of classical Sanskrit. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 41–59. [https://doi.org/10.1007/978-3-319-23980-4\\_3](https://doi.org/10.1007/978-3-319-23980-4_3)
- Hellwig, Oliver. 2016. Detecting sentence boundaries in Sanskrit texts. In *International Conference on Computational Linguistics: Technical Papers (COLING)*, pages 288–297.
- Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>, PubMed: 9377276
- Holmes, David I. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117. <https://doi.org/10.1093/l1c/13.3.111>
- Homburg, Timo and Christian Chiacros. 2016. Word segmentation for Akkadian cuneiform. In *Language Resources and Evaluation Conference (LREC)*, pages 4067–4074.
- Huang, Hongxiang, Daihui Yang, Gang Dai, Zhen Han, Yuyi Wang, Kin-Man Lam, Fan Yang, Shuangping Huang, Yongge Liu, and Mengchao He. 2022. AGTGAN: Unpaired image translation for photographic ancient character generation. In *ACM International Conference on Multimedia*, pages 5456–5467. <https://doi.org/10.1145/3503161.3548338>
- Huang, Hen Hsen, Chuen-Tsai Sun, and Hsin-Hsi Chen. 2010. Classical Chinese sentence segmentation. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 15–22.
- Hubel, David H. and Torsten N. Wiesel. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154. <https://doi.org/10.1113/jphysiol.1962.sp006837>, PubMed: 14449617
- Jauhiainen, Tommi, Heidi Jauhiainen, Tero Alstola, and Krister Lindén. 2019. Language and dialect identification of cuneiform texts. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 89–98. <https://doi.org/10.18653/v1/W19-1409>
- Jawahar, Ganesh, Benjamin Muller, Amal Fethi, Louis Martin, Éric Villemonte De La Clergerie, Benoît Sagot, and Djamé

- Seddah. 2018. ELMoLex: Connecting ELMo and lexicon features for dependency parsing. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–16.
- Ji, Tao, Yufang Liu, Yijun Wang, Yuanbin Wu, and Man Lan. 2018. AntNLP at CoNLL 2018 shared task: A graph-based parser for universal dependency parsing. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 248–255.
- Jindal, Amar and Rajib Ghosh. 2022. Text line segmentation in Indian ancient handwritten documents using faster R-CNN. *Multimedia Tools and Applications*, pages 1–20. <https://doi.org/10.1007/s11042-022-13709-y>
- Johnson, Kyle P., Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Association for Computational Linguistics*, pages 20–29. <https://doi.org/10.18653/v1/2021.acl-demo.3>
- Jones, Mason, Francesco Romano, and Abidalrahman Mohd. 2022. Machine learning in textual criticism: An examination of the performance of supervised machine learning algorithms in reconstructing the text of the Greek New Testament. In *2022 7th International Conference on Machine Learning Technologies (ICMLT)*, pages 1–5. <https://doi.org/10.1145/3529399.3529400>
- Kanerva, Jenna, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133–142.
- Kang, Kyeongpil, Kyohoon Jin, Soyoung Yang, Soojin Jang, Jaegul Choo, and Youngbin Kim. 2021. Restoring and mining the records of the Joseon dynasty via neural language modeling and machine translation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4031–4042. <https://doi.org/10.18653/v1/2021.naacl-main.317>
- Karajgikar, Jajwalya, Amira Al-Khulaidy, and Anamaria Berea. 2021. Computational pattern recognition in Linear A. <https://hal.science/hal-03207615/document>.
- Kaše, V., P. Heřmánková, and A. Sobotková. 2021. Classifying Latin inscriptions of the Roman empire: A machine-learning approach. In *Workshop on Computational Humanities Research*, volume 2989, pages 123–35.
- Kashyap, K. Harish and P. A. Koushik. 2003. Hybrid neural network architecture for age identification of ancient Kannada scripts. In *International Symposium on Circuits and Systems*, volume 5, pages V–V, IEEE. <https://doi.org/10.1109/ISCAS.2003.1206399>
- Keersmaekers, Alek. 2020. Automatic semantic role labeling in ancient Greek using distributional semantic modeling. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 59–67.
- Keersmaekers, Alek, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. Creating, enriching and valorizing treebanks of ancient Greek. In *International Workshop on Treebanks and Linguistic Theories (TLT)*, pages 109–117. <https://doi.org/10.18653/v1/W19-7812>
- Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016. Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, 63:86–96. <https://doi.org/10.1016/j.eswa.2016.06.029>
- Kirnap, Ömer, Erenay Dayanık, and Deniz Yuret. 2018. Tree-stack LSTM in transition based dependency parsing. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 124–132.
- Koentges, Thomas. 2020. The un-Platonic Menexenus: A stylometric analysis with more data. *Greek, Roman, and Byzantine Studies*, 60(2):211–241.
- Köntges, Thomas. 2020. Measuring philosophy in the first thousand years of Greek literature. *Digital Classics Online*, pages 1–23.
- Koppel, Moshe, Moty Michaely, and Alex Tal. 2016. Reconstructing ancient literary texts from noisy manuscripts. In *Workshop on Computational Linguistics for Literature*, pages 40–46. <https://doi.org/10.18653/v1/W16-0205>
- Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26. <https://doi.org/10.1002/asi.20961>

- Koppel, Moshe and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187. <https://doi.org/10.1002/asi.22954>
- Kumar, Puneet, Kshitij Pathania, and Balasubramanian Raman. 2022. Zero-shot learning based cross-lingual sentiment analysis for Sanskrit text with insufficient labeled data. *Applied Intelligence*, pages 1–18. <https://doi.org/10.1007/s10489-022-04046-6>
- Lai, Songxuan, Yecheng Zhu, and Lianwen Jin. 2020. Encoding pathlet and SIFT features with bagged VLAD for historical writer identification. *IEEE Transactions on Information Forensics and Security*, 15:3553–3566. <https://doi.org/10.1109/TIFS.2020.2991880>
- Lazar, Koren, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. Filling the gaps in ancient Akkadian texts: A masked language modeling approach. *arXiv preprint arXiv:2109.04513*. <https://doi.org/10.18653/v1/2021.emnlp-main.384>
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444. <https://doi.org/10.1038/nature14539>, PubMed: 26017442
- Lee, John S. Y. 2007. A computational model of text reuse in ancient literary texts. In *Annual Meeting of the Association of Computational Linguistics*, pages 472–479.
- Li, Bin, Yiguo Yuan, Jingya Lu, Minxuan Feng, Chao Xu, Weiguang Qu, and Dongbo Wang. 2022. The first international ancient Chinese word segmentation and POS tagging bakeoff: Overview of the EvaHan 2022 evaluation campaign. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 135–140.
- Li, Si, Mingzheng Li, Yajing Xu, Zuyi Bao, Lu Fu, and Yan Zhu. 2018a. Capsules based Chinese word segmentation for ancient Chinese medical books. *IEEE Access*, 6:70874–70883. <https://doi.org/10.1109/ACCESS.2018.2881280>
- Li, Zuchao, Shexia He, Zhuosheng Zhang, and Hai Zhao. 2018b. Joint learning of POS and dependencies for multilingual universal dependency parsing. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 65–73.
- Liu, Yashan, Jie Zhu, Zezhou Xu, Songsi Yan, Chao Wang, and Zehui Xu. 2022. Research on multi-line recognition algorithm for Tibetan document. In *2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML)*, pages 72–76, IEEE. <https://doi.org/10.1109/PRML56267.2022.9882261>
- Luo, Jiaming, Yuan Cao, and Regina Barzilay. 2019. Neural decipherment via minimum-cost flow: From Ugaritic to Linear B. In *Annual Meeting of the Association for Computational Linguistics*, pages 3146–3155. <https://doi.org/10.18653/v1/P19-1303>
- Luo, Jiaming, Frederik Hartmann, Enrico Santus, Regina Barzilay, and Yuan Cao. 2021. Deciphering undersegmented ancient scripts using phonetic prior. *Transactions of the Association for Computational Linguistics*, 9:69–81. [https://doi.org/10.1162/tac1\\_a\\_00354](https://doi.org/10.1162/tac1_a_00354)
- Manning, Christopher and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.
- Manousakis, Nikos and Efsthios Stamatatos. 2018. Devising Rhesus: A strange collaboration between Aeschylus and Euripides. *Digital Scholarship in the Humanities*, 33(2):347–361. <https://doi.org/10.1093/dlsc/fqx021>
- Mantovan, Lorenzo and Loris Nanni. 2020. The computerization of archaeology: Survey on artificial intelligence techniques. *SN Computer Science*, 1(5):1–32. <https://doi.org/10.1007/s42979-020-00286-w>
- Martins, Armando, Clara Grácio, Cláudia Teixeira, Irene Pimenta Rodrigues, Juan Luís García Zapata, and Lígia Ferreira. 2021. Historia Augusta authorship: An approach based on measurements of complex networks. *Applied Network Science*, 6(1):1–23. <https://doi.org/10.1007/s41109-021-00390-7>
- Matsumoto, Mallory E. 2022. Archaeology and epigraphy in the digital era. *Journal of Archaeological Research*, 30(2):285–320. <https://doi.org/10.1007/s10814-021-09162-4>
- McGillivray, Barbara, Beatrice Alex, Sarah Ames, Guyda Armstrong, David Beavan, Arianna Ciula, Giovanni Colavizza, James Cummings, David De Roure, Adam Farquhar, Simon Hengchen, Anouk Lang, James Loxley, Eirini Goudarouli, Federico Nanni, Andrea Nini, Julianne Nyhan, Nicola Osborne, Thierry Poibeau, Mia Ridge, Sonia Ranade, James Smithies, Melissa Terras, Andreas Vlachidis, and Pip Willcox. 2020. The challenges and prospects of the intersection of humanities

- and data science: A white paper from the Alan Turing Institute. *Alan Turing Institute*. Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113. <https://doi.org/10.1016/j.asej.2014.04.011>
- Meloni, Carlo, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4460–4473. <https://doi.org/10.18653/v1/2021.naacl-main.353>
- Mercelis, Wouter and Alek Keersmaekers. 2022. An electra model for Latin token tagging tasks. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 189–192.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.
- Mohammed, Hussein, Isabelle Marthot-Santaniello, and Volker Märgner. 2019. GRK-papyri: A dataset of Greek handwriting on papyri for the task of writer identification. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 726–731. <https://doi.org/10.1109/ICDAR.2019.00121>
- Molton, Nicholas, Xiaobo Pan, Michael Brady, Alan K. Bowman, Charles Crowther, and Roger Tomlin. 2003. Visual enhancement of incised text. *Pattern Recognition*, 36(4):1031–1043. [https://doi.org/10.1016/S0031-3203\(02\)00112-7](https://doi.org/10.1016/S0031-3203(02)00112-7)
- Monroe, M. Willis. 2018. Using quantitative methods for measuring inter-textual relations in cuneiform. *Digital Biblical Studies*, pages 257–280. <https://doi.org/10.1163/9789004375086.010>
- Moritz, Maria, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler. 2016. Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to bible reuse. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1849–1859. <https://doi.org/10.18653/v1/D16-1190>
- Mostofi, Fahimeh and Adnan Khashman. 2014. Intelligent recognition of ancient Persian cuneiform characters. In *International Conference on Neural Computation Theory and Applications*, pages 119–123. <https://doi.org/10.5220/0005035401190123>
- Moustafa, Ragaa, Farida Hesham, Samiha Hussein, Badr Amr, Samira Refaat, Nada Shorim, and Taraggy M. Ghanim. 2022. Hieroglyphs language translator using deep learning techniques (Scriba). In *International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 125–132. <https://doi.org/10.1109/MIUCC55081.2022.9781784>
- Narang, Sonika, M. K. Jindal, and Munish Kumar. 2019. Devanagari ancient documents recognition using statistical feature extraction techniques. *Sādhanā*, 44(6):1–8. <https://doi.org/10.1007/s12046-019-1126-9>
- Narang, Sonika Rani, Manish Kumar Jindal, Shruti Ahuja, and Munish Kumar. 2020. On the recognition of Devanagari ancient handwritten characters using SIFT and Gabor features. *Soft Computing*, 24(22):17279–17289. <https://doi.org/10.1007/s00500-020-05018-z>
- Narang, Sonika Rani, Manish Kumar Jindal, and Munish Kumar. 2020. Ancient text recognition: A review. *Artificial Intelligence Review*, 53(8):5517–5558. <https://doi.org/10.1007/s10462-020-09827-4>
- Narang, Sonika Rani, Munish Kumar, and Manish Kumar Jindal. 2021. DeepNetDevanagari: A deep learning model for Devanagari ancient character recognition. *Multimedia Tools and Applications*, 80(13):20671–20686. <https://doi.org/10.1007/s11042-021-10775-6>
- Nasir, Sidra and Imran Siddiqi. 2020. Learning features for writer identification from handwriting on papyri. In *Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, pages 229–241. [https://doi.org/10.1007/978-3-030-71804-6\\_17](https://doi.org/10.1007/978-3-030-71804-6_17)
- Nguyen, Dat Quoc and Karin Verspoor. 2018. An improved neural network model for joint POS tagging and dependency parsing. *arXiv preprint arXiv:1807.03955*. <https://doi.org/10.18653/v1/K17-3014>
- Nguyen, Tien Nam, Jean-Christophe Burie, Thi-Lan Le, and Anne-Valerie Schweyer. 2021. On the use of attention in deep learning based denoising method for ancient Cham inscription images. In *International Conference on Document Analysis and Recognition*, pages 400–415. [https://doi.org/10.1007/978-3-030-86549-8\\_26](https://doi.org/10.1007/978-3-030-86549-8_26)

- Ntzios, Kostas, Basilios Gatos, Ioannis Pratikakis, Thomas Konidakis, and Stavros J. Perantonis. 2007. An old Greek handwritten OCR system based on an efficient segmentation-free approach. *International Journal on Document Analysis and Recognition (IJ DAR)*, 9(2):179–192. <https://doi.org/10.1007/s10032-006-0031-z>
- Ochab, Jeremi K. and Holger Essler. 2019. Stylometry of literary papyri. In *International Conference on Digital Access to Textual Cultural Heritage*, pages 139–142. <https://doi.org/10.1145/3322905.3322930>
- Ouamour, Siham and Halim Sayoud. 2012. Authorship attribution of ancient texts written by ten Arabic travelers using a SMO-SVM classifier. In *International Conference on Communications and Information Technology (ICCIT)*, pages 44–47. <https://doi.org/10.1109/ICCITechnol.2012.6285841>
- Ouamour, Siham and Halim Sayoud. 2013a. Authorship attribution of ancient texts written by ten Arabic travelers using character  $n$ -grams. In *International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–5. <https://doi.org/10.1109/CITS.2013.6705713>
- Ouamour, Siham and Halim Sayoud. 2013b. Authorship attribution of short historical Arabic texts based on lexical features. In *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 144–147. <https://doi.org/10.1109/CyberC.2013.31>
- Ouamour, Siham and Halim Sayoud. 2018. A comparative survey of authorship attribution on short Arabic texts. In *International Conference on Speech and Computer*, pages 479–489. [https://doi.org/10.1007/978-3-319-99579-3\\_50](https://doi.org/10.1007/978-3-319-99579-3_50)
- Paetzold, Gustavo and Marcos Zampieri. 2019. Experiments in cuneiform language identification. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 209–213. <https://doi.org/10.18653/v1/W19-1423>
- Palaniappan, Satish and Ronjojoy Adhikari. 2017. Deep learning the Indus script. *arXiv preprint arXiv:1702.00523*.
- Palladino, Chiara, Farimah Karimi, and Brigitte Mathiak. 2020. NER on ancient Greek with minimal annotation. In *Digital Humanities 2020*, pages 1–3.
- Palmer, David D. 2000. Tokenization and sentence segmentation. *Handbook of Natural Language Processing*, pages 11–35.
- Panagopoulos, Michail, Constantin Papaodysseus, Panayiotis Rousopoulos, Dimitra Dafi, and Stephen Tracy. 2008. Automatic writer identification of ancient Greek inscriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1404–1414. <https://doi.org/10.1109/TPAMI.2008.201>, PubMed: 19542575
- Paolanti, Marina, Rocco Pietrini, Laura Della Sciuca, Emanuele Balloni, Benedetto Luigi Compagnoni, Antonella Cesarini, Luca Fois, Pierluigi Feliciati, and Emanuele Frontoni. 2022. PernaNet: A deep learning framework for automatic appearance-based analysis of ancient parchment collections. In *International Conference on Image Analysis and Processing*, pages 290–301. [https://doi.org/10.1007/978-3-031-13324-4\\_25](https://doi.org/10.1007/978-3-031-13324-4_25)
- Papantoniou, Katerina and Yannis Tzitzikas. 2020. NLP for the Greek language: A brief survey. In *Hellenic Conference on Artificial Intelligence*, pages 101–109. <https://doi.org/10.1145/3411408.3411410>
- Papaodysseus, Constantin, Panayiotis Rousopoulos, Dimitris Arabadjis, Fivi Panopoulou, and Michalis Panagopoulos. 2010. Handwriting automatic classification: Application to ancient Greek inscriptions. In *International Conference on Autonomous and Intelligent Systems*, pages 1–6. <https://doi.org/10.1109/AIS.2010.5547045>
- Papaodysseus, Constantin, Panayiotis Rousopoulos, Fotios Giannopoulos, Solomon Zannos, Dimitris Arabadjis, Michalis Panagopoulos, E. Kalfa, Christopher Blackwell, and Stephen Tracy. 2014. Identifying the writer of ancient inscriptions and Byzantine codices. A novel approach. *Computer Vision and Image Understanding*, 121:57–73. <https://doi.org/10.1016/j.cviu.2014.01.003>
- Papargioglou, Asimina, John Pavlopoulos, and Maria Konstantinidou. 2022. Dating Greek papyri images with machine learning. In *ICDAR Workshop on Computational Paleography*, <https://doi.org/10.21203/rs.3.rs-2272076/v1>
- Papavassileiou, Katerina, Dimitrios I. Kosmopoulos, and Gareth Owens. 2022. A generative model for the Mycenaean Linear B script and its application in

- infilling text from ancient tablets. *ACM Journal on Computing and Cultural Heritage*. <https://doi.org/10.1145/3593431>
- Papavassiliou, Katerina, Gareth Owens, and Dimitrios Kosmopoulos. 2020. A dataset of Mycenaean Linear B sequences. In *Language Resources and Evaluation Conference*, pages 2552–2561.
- Park, Chanjun, Chanhee Lee, Yeongwook Yang, and Heuseok Lim. 2020. Ancient Korean neural machine translation. *IEEE Access*, 8:116617–116625. <https://doi.org/10.1109/ACCESS.2020.3004879>
- Park, Chanjun, Seolhwa Lee, Jaehyung Seo, Hyeonseok Moon, Sugyeong Eo, and Heuseok Lim. 2022. Priming ancient Korean neural machine translation. In *Language Resources and Evaluation Conference (LREC)*.
- Parker, Clifford Seth, Stephen Parsons, Jack Bandy, Christy Chapman, Frederik Coppens, and William Brent Seales. 2019. From invisibility to readability: Recovering the ink of Herculaneum. *PLOS ONE*, 14(5):1–17. <https://doi.org/10.1371/journal.pone.0215775>, PubMed: 31067260
- Pavlopoulos, John and Maria Konstantinidou. 2022. Computational authorship analysis of the Homeric poems. *International Journal of Digital Humanities*, 4:45–64. <https://doi.org/10.1007/s42803-022-00046-7>
- Pavlopoulos, John, Alexandros Xenos, and Davide Picca. 2022. Sentiment analysis of Homeric text: The 1st Book of Iliad. In *Language Resources and Evaluation Conference (LREC)*, pages 7071–7077.
- Perrone, Valerio, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2019. GASC: Genre-aware semantic change for ancient Greek. In *International Workshop on Computational Approaches to Historical Language Change*, pages 56–66. <https://doi.org/10.18653/v1/W19-4707>
- Pirrone, Antoine, Marie Beurton Aimar, and Nicholas Journet. 2019. Papy-S-Net: A Siamese network to match papyrus fragments. In *International Workshop on Historical Document Imaging and Processing*, pages 78–83. <https://doi.org/10.1145/3352631.3352646>
- Popović, Mladen, Maruf A. Dhali, and Lambert Schomaker. 2021. Artificial intelligence based writer identification generates new evidence for the unknown scribes of the Dead Sea Scrolls exemplified by the Great Isaiah Scroll (1qisaa). *PLOS ONE*, 16(4):1–28. <https://doi.org/10.1371/journal.pone.0249769>, PubMed: 33882053
- Punia, Ravneet, Niko Schenk, Christian Chiacaros, and Émilie Pagé-Perron. 2020. Towards the first machine translation system for Sumerian transliterations. In *International Conference on Computational Linguistics*, pages 3454–3460. <https://doi.org/10.18653/v1/2020.coling-main.308>
- Qi, Peng, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2019. Universal dependency parsing from scratch. *arXiv preprint arXiv:1901.10457*. <https://doi.org/10.18653/v1/K18-2016>
- Raj, V. Amrutha, R. L. Jyothi, and A. Anilkumar. 2017. Grantha script recognition from ancient palm leaves using histogram of orientation shape context. In *International Conference on Computing Methodologies and Communication (ICCMC)*, pages 790–794. <https://doi.org/10.1109/ICCMC.2017.8282574>
- Rao, Rajesh P. N., Nisha Yadav, Mayank N. Vahia, Hrishikesh Joglekar, R. Adhikari, and Iravatham Mahadevan. 2009a. Entropic evidence for linguistic structure in the Indus script. *Science*, 324(5931):1165. <https://doi.org/10.1126/science.1170391>, PubMed: 19389998
- Rao, Rajesh P. N., Nisha Yadav, Mayank N. Vahia, Hrishikesh Joglekar, Ronjojoy Adhikari, and Iravatham Mahadevan. 2009b. A Markov model of the Indus script. *Proceedings of the National Academy of Sciences (PNAS)*, 106(33):13685–13690. <https://doi.org/10.1073/pnas.0906237106>, PubMed: 19666571
- Rao, Rajesh P. N., Nisha Yadav, Mayank N. Vahia, Hrishikesh Joglekar, Ronjojoy Adhikari, and Iravatham Mahadevan. 2010. Entropy, the Indus script, and language: A reply to R. Sproat. *Computational Linguistics*, 36(4):795–805. [https://doi.org/10.1162/coli.c\\_00030](https://doi.org/10.1162/coli.c_00030)
- Reisi, Ehsan and Hassan Mahboob Farimani. 2020. Authorship attribution in historical and literary texts by a deep learning classifier. *Journal of Applied Intelligent Systems and Information Sciences*, 1(2):118–127.
- Rizk, Rodrigue, Dominick Rizk, Frederic Rizk, and Ashok Kumar. 2021. A hybrid capsule network-based deep learning framework for deciphering ancient scripts with scarce annotations: A case study on Phoenician epigraphy. In *IEEE*



- International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 617–620. <https://doi.org/10.1109/MWSCAS47672.2021.9531798>
- Robinson, Andrew. 2009. *Writing and Script: A Very Short Introduction*, volume 208. Oxford University Press. <https://doi.org/10.1093/actrade/9780199567782.001.0001>
- Roelli, Philipp and Dieter Bachmann. 2010. Towards generating a stemma of complicated manuscript traditions: Petrus Alfonsi's Dialogus. *Revue d'Histoire des Textes*, 5:307–331. <https://doi.org/10.1484/J.RHT.5.101260>
- Roos, Teemu and Tuomas Heikkilä. 2009. Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets. *Literary and Linguistic Computing*, 24(4):417–433. <https://doi.org/10.1093/llc/fqp002>
- Rybak, Piotr and Alina Wróblewska. 2020. Semi-supervised neural system for tagging, parsing and lematization. *arXiv preprint arXiv:2004.12450*.
- Sahala, Aleksi. 2021. *Contributions to Computational Assyriology*. Ph.D. thesis, Helsingin yliopisto.
- Sahala, Aleksi, Miikka Silfverberg, Antti Arppe, and Krister Lindén. 2020a. Automated phonological transcription of Akkadian cuneiform text. In *Language Resources and Evaluation Conference (LREC)*.
- Sahala, Aleksi, Miikka Silfverberg, Antti Arppe, and Krister Lindén. 2020b. BabyFST: Towards a finite-state based computational model of ancient Babylonian. In *Language Resources and Evaluation Conference (LREC)*.
- Scheirer, Walter, Christopher Forstall, and Neil Coffee. 2016. The sense of a connection: Automatic tracing of intertextuality by meaning. *Digital Scholarship in the Humanities*, 31(1):204–217. <https://doi.org/10.1093/llc/fqu058>
- Seuret, Mathias, Anguelos Nicolaou, Dominique Stutzmann, Andreas Maier, and Vincent Christlein. 2020. ICFHR 2020 competition on image retrieval for historical handwritten fragments. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 216–221. <https://doi.org/10.1109/ICFHR2020.2020.00048>
- Shaus, Arie. 2017. *Computer Vision and Machine Learning Methods for Analyzing First Temple Period Inscriptions*. Ph.D. thesis, Tel Aviv University.
- Shen, Tianxiao, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020. Blank language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 5186–5198. <https://doi.org/10.18653/v1/2020.emnlp-main.420>
- Singh, Pranaydeep, Gorik Rutten, and Els Lefever. 2021. A pilot study for BERT language modeling and morphological analysis for ancient and medieval Greek. In *SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–137. <https://doi.org/10.18653/v1/2021.latechclfl-1.15>
- Smith, Aaron, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal dependency parsing with multi-treebank models. *arXiv preprint arXiv:1809.02237*. <https://doi.org/10.18653/v1/K18-2011>
- Snyder, Benjamin, Regina Barzilay, and Kevin Knight. 2010. A statistical model for lost language decipherment. In *Association for Computational Linguistics*, pages 1048–1057.
- Son, Juhee, Jiho Jin, Haneul Yoo, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. Translating Hanja historical documents to contemporary Korean and English. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 1260–1272.
- Soumya, A. and G. Hemantha Kumar. 2014. Classification of ancient epigraphs into different periods using random forests. In *International Conference on Signal and Image Processing*, pages 171–178. <https://doi.org/10.1109/ICSHIP.2014.33>
- Sproat, Richard. 2010. Last words: Ancient symbols, computational linguistics, and the reviewing practices of the general science journals. *Computational Linguistics*, 36(3):585–594. <https://doi.org/10.1162/coli.a.00011>
- Sproat, Richard. 2014. A statistical comparison of written language and nonlinguistic symbol systems. *Language*, 90(2):457–481. <https://doi.org/10.1353/lan.2014.0031>
- Sprugnoli, Rachele, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvaLatin 2020 evaluation campaign. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 105–110.
- Sprugnoli, Rachele, Marco Passarotti, Cecchini Flavio Massimiliano, Margherita

- Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 183–188.
- Sprugnoli, Rachele, Marco Passarotti, and Giovanni Moretti. 2019. Vir is to moderatus as mulier is to intemperans-lemma embeddings for Latin.. In *CLiC-it*. <https://ceur-ws.org/Vol-2481/paper69.pdf>
- Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556. <https://doi.org/10.1002/asi.21001>
- Stoeckel, Manuel, Alexander Henlein, Wahed Hemati, and Alexander Mehler. 2020. Voting for POS tagging of Latin texts: Using the flair of flair to better ensemble classifiers by example of Latin. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 130–135.
- Stokes, Peter A. 2015. Digital approaches to paleography and book history: Some challenges, present and future. *Frontiers in Digital Humanities*. <https://doi.org/10.3389/fdigh.2015.00005>
- Stover, Justin Anthony, Yaron Winter, Moshe Koppel, and Mike Kestemont. 2016. Computational authorship verification method attributes a new work to a major 2nd century African author. *Journal of the Association for Information Science and Technology*, 67(1):239–242. <https://doi.org/10.1002/asi.23460>
- Straka, Milan. 2018. UDpipe 2.0 prototype at CoNLL 2018 UD shared task. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.
- Straka, Milan, Jan Hajič, and Jana Straková. 2016. UDpipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Language Resources and Evaluation Conference (LREC)*, pages 4290–4297.
- Straka, Milan and Jana Straková. 2020. UDpipe at EvaLatin 2020: Contextualized embeddings and treebank embeddings. *arXiv preprint arXiv:2006.03687*.
- Straka, Milan, Jana Straková, and Jan Hajič. 2019. Evaluating contextualized embeddings on 54 languages in POS tagging, lemmatization and dependency parsing. *arXiv preprint arXiv:1908.07448*.
- Subramani, Kavitha and S. Murugavalli. 2019. Recognizing ancient characters from Tamil palm leaf manuscripts using convolution based deep learning. *International Journal of Recent Technology and Engineering*, 8(3):6873–6880. <https://doi.org/10.35940/ijrte.C5842.098319>
- Suganya, T. S. and Subramaniam Murugavalli. 2017. Feature selection for an automated ancient Tamil script classification system using machine learning techniques. In *International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*, pages 1–6. <https://doi.org/10.1109/ICAMMAET.2017.8186731>
- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27.
- Svärd, Saana, Heidi Jauhiainen, Aleks Sahala, and Krister Lindén. 2018. Semantic domains in Akkadian texts. *CyberResearch on the Ancient Near East and Neighboring Regions. Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving*, 2:224–256.
- Swindall, Matthew I., Gregory Croisdale, Chase C. Hunter, Ben Keener, Alex C. Williams, James H. Brusuelas, Nita Krevans, Melissa Sellew, Lucy Fortson, and John F. Wallin. 2021. Exploring learning approaches for ancient Greek character recognition with citizen science data. In *International Conference on eScience*, pages 128–137. <https://doi.org/10.1109/eScience51609.2021.00023>
- Swindall, Matthew I., Timothy Player, Ben Keener, Alex C. Williams, James H. Brusuelas, Federica Nicolardi, Marzia D'Angelo, Claudio Vergara, Michael McOsker, and John F. Wallin. 2022. Dataset augmentation in papyrology with generative models: A study of synthetic ancient Greek character images. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4973–4979. <https://doi.org/10.24963/ijcai.2022/689>
- Tang, Binghao, Boda Lin, and Si Li. 2022. Simple tagging system with RoBERTa for ancient Chinese. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 159–163.
- Tang, Xuemei, Shichen Liang, and Zhiying Liu. 2019. Authorship attribution of the Golden Lotus based on text classification methods. In *International Conference on*

- Innovation in Artificial Intelligence*, pages 69–72. <https://doi.org/10.1145/3319921.3319958>
- Terras, Melissa and Paul Robertson. 2005. Image and interpretation using artificial intelligence to read ancient Roman texts. *Human IT*, 7(3):1–56.
- Tian, Huishuang, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. 2021. AnchiBERT: A pre-trained model for ancient Chinese language understanding and generation. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. <https://doi.org/10.1109/IJCNN52387.2021.9534342>
- Tracy, Stephen V. and Constantin Papaodysseus. 2009. The study of hands on Greek inscriptions: The need for a digital approach. *American Journal of Archaeology*, pages 99–102. <https://doi.org/10.3764/aja.113.1.99>
- Tsirogiannis, Christos. 2020. The itinerary of a stolen stele. *UNESCO Courier*, 2020(4):18–20. <https://doi.org/10.18356/c002fa64-en>
- Tuccinardi, Enrico. 2017. An application of a profile-based method for authorship verification: Investigating the authenticity of Pliny the Younger’s letter to Trajan concerning the Christians. *Digital Scholarship in the Humanities*, 32(2):435–447. <https://doi.org/10.1093/llc/fqw001>
- Tupman, Charlotte, Dmitry Kangin, and Jacqueline Christmas. 2021. Reconsidering the Roman workshop: Using computer vision to analyse the making of ancient inscriptions. *Umanistica Digitale*, 10:461–473.
- Tyndall, Stephen. 2012. Toward automatically assembling Hittite-language Cuneiform tablet fragments into larger texts. In *Annual Meeting of the Association for Computational Linguistics*, pages 243–247.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.
- Vatri, Alessandro and Barbara McGillivray. 2018. The Diorisis ancient Greek corpus: Linguistics and literature. *Research Data Journal for the Humanities and Social Sciences*, 3(1):55–65. <https://doi.org/10.1163/24523666-01000013>
- Vatri, Alessandro and Barbara McGillivray. 2020. Lemmatization for ancient Greek: An experimental assessment of the state of the art. *Journal of Greek Linguistics*, 20(2):179–196. <https://doi.org/10.1163/15699846-02002001>
- Wan, Hui, Tahira Naseem, Young-Suk Lee, Vittorio Castelli, and Miguel Ballesteros. 2018. IBM research at the CoNLL 2018 shared task on multilingual parsing. In *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 92–102.
- Wang, Boli, Xiaodong Shi, Zhixing Tan, Yidong Chen, and Weili Wang. 2016. A sentence segmentation method for ancient Chinese texts based on NNLM. In *Workshop on Chinese Lexical Semantics*, pages 387–396. [https://doi.org/10.1007/978-3-319-49508-8\\_36](https://doi.org/10.1007/978-3-319-49508-8_36)
- Wei, Xinyuan, Weihao Liu, Qing Zong, Shaoqing Zhang, and Baotian Hu. 2022. Glyph features matter: A multimodal solution for EvaHan in LT4HALA2022. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 178–182.
- Wijerathna, K. A. S. A. Nilupuli, Rashmi Sepalitha, Thuiyadura Indika, Harshana Athauda, P. D. Suranjini, J. A. D. C. Silva, and Anuradha Jayakodi. 2019. Recognition and translation of ancient Brahmi letters using deep learning and NLP. In *International Conference on Advancements in Computing (ICAC)*, pages 226–231.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9.
- Wishart, Ryder and Prokopis Prokopidis. 2017. Topic modeling experiments on Hellenistic corpora. In *CDH@ TLT*, pages 39–47.
- Woodhead, Arthur Geoffrey. 1959. *The Study of Greek Inscriptions*, volume 424, Cambridge University Press. <https://doi.org/10.2307/147248>
- Wróbel, Krzysztof and Krzysztof Nowak. 2022. Transformer-based part-of-speech tagging and lemmatization for Latin. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 193–197.
- Wu, Nianheng, Eric DeMattos, Kwok Him So, Pin-zhen Chen, and Çağrı Çöltekin. 2019. Language discrimination and transfer learning for similar languages:

- Experiments with feature combinations and adaptation. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 54–63. <https://doi.org/10.18653/v1/W19-1406>
- Wu, Winston and Garrett Nicolai. 2020. JHUBC's submission to LT4HALA EvaLatin 2020. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 114–118.
- Yadav, Nisha, Hrishikesh Joglekar, Rajesh P. N. Rao, Mayank N. Vahia, Ronojoy Adhikari, and Iravatham Mahadevan. 2010. Statistical analysis of the Indus script using  $n$ -grams. *PLOS ONE*, 5(3):e9506. <https://doi.org/10.1371/journal.pone.0009506>, PubMed: 20333254
- Yamshchikov, Ivan P., Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. BERT in Plutarch's shadows. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 6071–6080.
- Yang, Kexin, Dayiheng Liu, Qian Qu, Yongsheng Sang, and Jiancheng Lv. 2021. An automatic evaluation metric for ancient-modern Chinese translation. *Neural Computing and Applications*, 33(8):3855–3867. <https://doi.org/10.1007/s00521-020-05216-8>
- Yang, Shuxun. 2022. A joint framework for ancient Chinese WS and POS tagging based on adversarial ensemble learning. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 174–177.
- Yoo, Haneul, Jiho Jin, Juhee Son, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. HUE: Pretrained model and dataset for understanding Hanja documents of ancient Korea. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1832–1844. <https://doi.org/10.18653/v1/2022.findings-naacl.140>
- Yoshimura, Mamoru, Fuminori Kimura, and Akira Maeda. 2012. Word segmentation for text in Japanese ancient writings based on probability of character  $n$ -grams. In *International Conference on Asian Digital Libraries*, pages 313–316. [https://doi.org/10.1007/978-3-642-34752-8\\_38](https://doi.org/10.1007/978-3-642-34752-8_38)
- Yousef, Tariq, Chiara Palladino, David J. Wright, and Monica Berti. 2022. Automatic translation alignment for ancient Greek and Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107. <https://doi.org/10.31219/osf.io/8epsy>
- Yu, J. S., Y. Wei, Y. W. Zhang, and H. Yang. 2020. Word segmentation for ancient Chinese texts based on nonparametric Bayesian models and deep learning. *Journal of Chinese Information Processing*, 34(6):1–8.
- Yu, Tianxiu, Cong Lin, Shijie Zhang, Chunxue Wang, Xiaohong Ding, Huili An, Xiaoxiang Liu, Ting Qu, Liang Wan, Shaodi You, et al. 2022. Artificial intelligence for Dunhuang cultural heritage protection: The project and the dataset. *International Journal of Computer Vision*. 130:1–28. <https://doi.org/10.1007/s11263-022-01665-x>
- Yu, Xuejin and Wei Huangfu. 2019. A machine learning model for the dating of ancient Chinese texts. In *International Conference on Asian Language Processing (IALP)*, pages 115–120. <https://doi.org/10.1109/IALP48816.2019.9037653>
- Zampieri, Marcos, Shervin Malmasi, Yves Scherrer, Tanja Samardzic, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, et al. 2019. A report on the third VarDial evaluation campaign. In *Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–16.
- Zeman, Daniel, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umüt Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyong Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaraj, and Josie Li. 2017.

- CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19. <https://doi.org/10.18653/v1/K17-3001>
- Zhang, Chongsheng, Bin Wang, Ke Chen, Ruixing Zong, Bo-feng Mo, Yi Men, George Almpandis, Shanxiong Chen, and Xiangliang Zhang. 2022a. Data-driven Oracle Bone rejoining: A dataset and practical self-supervised learning scheme. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4482–4492. <https://doi.org/10.1145/3534678.3539050>
- Zhang, Hailin, Ziyu Yang, Yingwen Fu, and Ruoyao Ding. 2022b. BERT 4ever@EvaHan 2022: Ancient Chinese word segmentation and part-of-speech tagging based on adversarial learning and continual pre-training. In *Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 150–154.
- Zhang, Hailin, Hai Zhu, Junsong Ruan, and Ruoyao Ding. 2021. People name recognition from ancient Chinese literature using distant supervision and deep learning. In *International Conference on Artificial Intelligence and Information Systems*, pages 1–6. <https://doi.org/10.1145/3469213.3470270>
- Zhang, Yi Kang, Heng Zhang, Yong-Ge Liu, Qing Yang, and Cheng-Lin Liu. 2019. Oracle character recognition by nearest neighbor classification with deep metric learning. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 309–314. <https://doi.org/10.1109/ICDAR.2019.00057>
- Zhang, Zhiyuan, Wei Li, and Qi Su. 2019. Automatic translating between ancient Chinese and contemporary Chinese with limited aligned corpora. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 157–167. [https://doi.org/10.1007/978-3-030-32236-6\\_13](https://doi.org/10.1007/978-3-030-32236-6_13)
- Zhao, Hongshuai, Haozhen Chu, Yuanyuan Zhang, and Yu Jia. 2020. Improvement of ancient Shui character recognition model based on convolutional neural network. *IEEE Access*, 8:33080–33087. <https://doi.org/10.1109/ACCESS.2020.2972807>