Bytes and Bygones – Abstracts

T. Adekoya, A. Anderson

Advancing Named Entity Disambiguation in Ancient Text Networks through Homonym Cohort Analysis

The accurate identification and disambiguation of named entities pose critical challenges in historical research, particularly for ancient texts with numerous homonyms—individuals sharing identical names. This paper presents an innovative computational methodology developed for disambiguating homonyms within extensive corpora of ancient Akkadian texts, using what we term Homonym Cohort Analysis. Building on prior work in Named Entity Recognition and initial network formation in the FactGrid Cuneiform Project, our method systematically constructs subnetworks around homonyms, exploiting contextual metadata such as patronyms, occupations, and relational roles (e.g., author, recipient, witness). By creating weighted links between homonyms and their associated cohorts (i.e. groups of entities frequently co-mentioned across texts), we establish structured, quantifiable criteria for distinguishing between individuals sharing the same name. Recursive network analyses iteratively refine these associations, significantly enhancing disambiguation accuracy compared to traditional manual methods or basic clustering techniques. Preliminary evaluations using Neo-Assyrian, Middle Assyrian, and Old Assyrian social network datasets have demonstrated substantial improvements in both accuracy and scalability. Our Homonym Cohort Analysis thus provides a powerful new tool for historical network analyses, enabling more precise reconstructions of historical interactions and societal structures. Future research will refine cohort detection algorithms and further integrate Natural Language Processing to automate and enhance the identification of relational roles, advancing our capabilities for event-scale historical analysis.

E. Alexandrova

Quantitative and Qualitative Approaches in Studying Egyptian Funerary Literature

This project applies a combination of computational and philological methods to investigate the evolution and semantic shifts of lexical fields related to sacrifice, offering, and hunting in Egyptian funerary literature. Using the Thesaurus Linguae Aegyptiae (TLA) as the primary data source, I analyze three major corpora — the Pyramid Texts, the Book of the Dead, and the Netherworld Books — which represent different stages and branches in the development of funerary traditions and textual practices.

My core aim is to trace how practical, ritual vocabulary — originally tied to concrete actions and real-world contexts such as animal slaughter and hunting — becomes increasingly abstracted and embedded within the mythological and cosmological discourse of the afterlife. To achieve this, I first manually extracted a list of semantically relevant lemmas, verified and refined it using LDA, and then applied frequency-based metrics, lexical diversity indices, and distribution analysis across the sub-corpora. I identified candidate lemmas that appear to 'migrate' from documentary

and economic spheres into funerary texts, reflecting shifts in religious semantics.

The preliminary results confirm significant differences in the density and variability of this lexical field between the corpora, reflecting genre-specific narrative strategies. The Pyramid Texts preserve remnants of concrete sacrificial imagery, while the Netherworld Books demonstrate fully mythologized representations.

However, an essential intermediate stage — the Coffin Texts — is not yet included in the TLA database. Given their importance for understanding transitional phases in funerary thought, I will complement the quantitative findings with a qualitative analysis of selected Coffin Text passages. Focusing on the usage contexts of specific lemmas, this analysis will help to verify the tendencies detected computationally and refine the proposed interpretation of lexical shifts.

Overall, this project demonstrates the potential of combining large-scale digital analysis with traditional close reading to uncover semantic transformations in ancient Egyptian religious texts.

H. Ashkenazi, O. Ze'evi-Berger, B. Gross, G. D. Stiebel

GIS, Photogrammetry, and Field Survey: Understanding Roman Siege System of Masada

At the end of the First Jewish Revolt against the Roman empire (66-73/74 CE), the last Jewish rebels sought refuge in the remote desert fortress of Masada. Determined to eliminate all resistance, the Roman army pursued them and, in 73 or 74 CE, laid siege to the fortress. To encircle Masada, the Romans constructed an extensive siege system, including a 4.3 km circumvallation wall, 15 towers, eight army camps, built access trails, and a massive siege ramp leading to the fortress walls. According to historical accounts, after the Romans breached the defences, nearly all of the thousand rebels chose mass suicide over enslavement.

The Neustadter Masada Expedition set out to document and study the Roman siege works. This was done by ground survey, 3D modelling with drones and digital cameras, and GIS analysis. The exceptional preservation of the siege system in the arid desert environment, combined with 3D modelling, allowed for the calculation of the volume of stone in the circumvallation wall, and therefore its original dimensions. These findings provided insights into the wall's function, construction duration, and the overall length of the siege. The integration of ground survey and GIS analysis further refined our understanding of the placement of towers and camps, as well as the post-depositional processes affecting the site.

Our study concludes that the circumvallation wall stood no higher than 2 to 2.5 meters and served multiple purposes: acting as a physical barrier (especially for preventing infiltrating into the fortress), exerting psychological pressure on the besieged, and serving as a platform for launching counterattacks. Based on our measurements and workload estimations, we argue that the construction of the siege wall and accompanying camps was completed in no more than two weeks.

L. Bampfield

Exploring Digital Pathways: Image Annotation and Multi-Method Digitisation of Seals

The integration of digital methodologies has transformed the documentation, analysis, and dissemination of ancient artefacts, providing scholars with new tools to engage with cultural heritage. This paper examines the role of image annotation in the digital recording of Mesopotamian cylinder seals and ongoing trials of different digitisation techniques for stamp seals. By employing high-resolution imaging, photogrammetry, and unwrapping techniques, this project generates structured digital representations that facilitate comparative analyses across collections and improve the accessibility of seal imagery. Central to this approach is the use of image annotation tools, which enable the systematic identification of motifs, enhance documentation consistency, and support interoperability through linked open data (LOD).

Beyond cylinder seals, this study investigates the challenges associated with the digitisation of stamp seals, a class of artefacts that presents distinct complexities due to variations in material, scale, and surface detail. Ongoing trials assess the efficacy of different imaging techniques to refine best practices for capturing and analysing these objects in alignment with broader digital cultural heritage initiatives. While manual annotation remains essential for ensuring accuracy in motif classification, machine-assisted approaches, including an object detector trained to identify recurring motifs, offer promising avenues for automating aspects of image analysis and facilitating large-scale comparative studies.

By situating this research within the wider fields of image annotation and digital cultural heritage, this study highlights the transformative potential of structured digital methodologies in the documentation and interpretation of ancient artefacts. Through a critical evaluation of both established and emerging techniques, this work contributes to ongoing efforts to standardise digitisation protocols, enhance data interoperability, and expand scholarly engagement with glyptic materials in digital humanities.

B. Brown-deVost, B. Kurar-Barakat, N. Dershowitz

Segmenting the Dead Sea Scroll Fragment Images

The Dead Sea Scrolls comprise close to a thousand ancient manuscripts discovered near Qumran between 1947 and 1956, primarily written on parchment and papyrus and dating from the third century BCE to the first century CE. Their texts hold significant historical, religious, and linguistic value, including some of the earliest Hebrew Bible manuscripts.

Since 2011, the Israel Antiquities Authority has been digitizing the scroll fragments, both recto and verso sides, creating high-resolution images using twelve wavelengths across visible and infrared spectra. Despite computational advantages of a blue background, black was chosen for minimal reflectivity. Each image includes calibration objects whose placement varies between fragments.

Accurate segmentation of the fragments in these images is essential for comparing historical and modern images, facilitating computational analyses, letter recognition, paleographic studies, and testing digital hypotheses about fragment joins. Segmentation challenges include variable calibration object placement, the similarity

between background and ink colors, and rice paper used for conservation.

We have created the publicly available "Qumran Segmentation Dataset," consisting of 138 fragment images (99 training, 20 validation, and 19 testing), as a benchmark for evaluating segmentation methods. Fragments were randomly selected without manuscript overlap to ensure generalizability. Each image includes bounding box annotations for the calibration objects. The annotations for recto images in the test set are provided as well-known text (WKT) polygons, offering precise contour information.

Our own segmentation method uses faster R-CNN to detect calibration objects, followed by the computation of an alignment matrix of recto and verso images using SIFT feature detection and RANSAC-based feature matching.

We threshold grayscale infrared images, which provide higher contrast between the fragment and the background, using a dynamic lower bound calculated from the average intensity of dark pixels plus a buffer. The resulting masks for the recto and verso sides are aligned and combined using the alignment matrix. This initial mask includes both the fragment (ink and parchment) and the surrounding rice paper regions, separating them from the background and parchment holes.

We then segment the rice paper using HSV color space thresholds, refined by morphological operations with an elliptical kernel, leveraging the consistent pixel density of the images. A rice paper mask is subtracted from the initial mask to produce the final fragment segmentation. Evaluation metrics show a mean IoU of 97%, precision of 98%, and recall of 99% at the pixel level.

J. Ph. Bullenkamp, F. Linsel, H. Mara

Neural Network-Based Wedge Segmentation for 3D-Scanned Cuneiform Tablets Using Synthetic Data

Motivated by the question, how to analyse the rich record of cuneiform tablets for making its ancient texts readable for us, we developed a workflow, which allows wedge segmentation on 3D scanned cuneiform tablets using neural networks.

Cuneiform tablets were inscribed by pressing wedge-shaped styluses into clay, making cuneiform script inherently three-dimensional. With the increasing availability of 3D scanned cuneiform tablets, analyzing their geometric and structural properties has become crucial. Wedge segmentation and classification are fundamental steps toward sign detection in an OCR pipeline. However, manually annotating wedges on 3D meshes is highly time-consuming, making large-scale ground truth datasets unavailable for machine learning tasks.

We address the challenge of training neural networks for wedge segmentation in the absence of large-scale annotated datasets. To overcome this limitation, we developed two complementary strategies for generating 3D ground truth data. First, we create synthetic cuneiform tablets by modeling wedge imprints on simulated surfaces, allowing us to automatically derive vertex-level labels from the mesh construction process. Second, we introduce an experimental cuneiform writing series, in which tablets were inscribed and 3D scanned after each wedge insertion, providing real-world training data with precise ground truth. Based on these datasets, we initiated the training of neural network architectures for mesh and point cloud segmentation, such as PointNet, as a first step toward wedge detection and recognition. Our approach lays

the groundwork for future large-scale applications of machine learning on 3D scanned cuneiform tablets.

Ch. D. Casey, M. X. Kudela, A. Alieva, J. Sido Bozan

Preserving Digital Humanities Data on the Ancient World

Digital-humanities projects focused on the ancient world frequently generate critical datasets, yet their long-term preservation remains uncertain. This presentation examines the sustainability of digital resources related to ancient texts, particularly those in non-Latin scripts. Drawing on research from Closing the Gap in Non-Latin Script Data, which catalogs digital projects working with historical and premodern textual data, we highlight significant risks to the longevity of these resources. Our mixed-methods study, combining quantitative analysis with qualitative case studies, reveals a troubling pattern: many grant-funded digital projects become inaccessible after their funding period ends, leading to a substantial loss of scholarly work and cultural knowledge.

This crisis is especially pronounced for projects involving non-Latin scripts, where technical limitations, institutional barriers, and funding models disproportionately hinder preservation efforts. Many of these resources are crucial for understanding ancient civilizations, yet they lack the long-term infrastructure afforded to more traditional forms of scholarship. Additionally, current academic evaluation systems often fail to recognize the importance of maintaining digital archives, further marginalizing research in this area.

To address these challenges, we propose a framework for sustainable digital preservation that integrates both technical and institutional strategies. Our recommendations outline the roles of universities, funding agencies, and academic publishers in ensuring that digital resources on the ancient world remain accessible beyond the lifespan of individual projects. The presentation concludes with concrete actions that Closing the Gap and similar initiatives can take to support a more equitable and durable future for digital humanities scholarship on ancient texts.

K. De Graef

Stylometric Explorations of Scribal Traditions and Ideological Discourse in Cuneiform Corpora.

The Neo-Assyrian empire (ca. 900–600 BCE) is known for its highly militarized and war-driven nature, with a reputation for ruthless and cruel military tactics. In contrast, the Old Babylonian kingdom (ca. 1900–1600 BCE) is remembered for its kings who styled themselves as shepherds of the people, bringing abundance, justice, and order. Although rulers in both periods waged war and violently expanded their territories, the perception of them could hardly be more different: fear and terror on the one hand, unity and peace on the other.

This paper presents a case study that analyzes whether these stereotypical narratives are reflected in, reinforced by, or challenged through the language of royal inscriptions. Over more than two millennia, royal inscriptions developed into a highly formalized genre that commemorated military victories, construction activities, and pious dedications to the gods using codified literary strategies.

Based on two lemmatized corpora of royal inscriptions (Oracc RINAP and RIME 4), this study investigates (1) genre continuity, (2) intra-corpus variation, and (3) semantic ideology. Using relative lemma frequency (TF), Principal Component Analysis (PCA), and cosine distance, it explores both the overall stylistic and semantic evolution of the genre and differences between individual kings. Keyness measures (log-likelihood ratio) and defined semantic categories (war, peace, neutral) further identify the overrepresentation of ideologically charged vocabulary.

K. De Graef (poster)

Who is Who in Sukkalmah Susa: Disambiguating Homonyms in Flawed Corpora via Ego-Network Similarity.

The identification of individuals in ancient administrative corpora is often complicated by homonymy, especially in cases where traditional markers of identity—such as patronymics, titles, seal impressions, or date formulas—are absent. This challenge is particularly acute in the cuneiform corpus from Sukkalmah-period Susa (2000-1500 BCE), with approximately 900 published texts lacking clear archaeological or archival context. This poster presents a network-based methodology for disambiguating homonymous personal names in such flawed corpora. The approach is grounded in the assumption that if two (near-)identical names occur in different texts but share a highly similar set of co-occurring individuals who play active roles in the respective transactions, they are more likely to refer to the same person. By constructing ego-networks for each name attestation and calculating their similarity using the Jaccard index and overlap coefficient, the method probabilistically assesses identity, merging homonyms that exceed certain thresholds.

A case study focused on the frequently attested name Nur-Inšušinak illustrates the method's potential. Out of 323 names in 32 undated Susa texts, 100 attestations were merged into 27 distinct individuals, including 18 references to Nur-Inšušinak likely representing a single agricultural entrepreneur. His reconstructed social network reveals a well-structured, socially diverse community including women, priests, and scribes, with clear sub-grouping and strong internal cohesion.

L. Foket, G. R. Smidt, K. De Graef

Bringing Mesopotamia Alive in the Belgian History Class through Interoperable Digital Exhibitions

Mesopotamia, often referred to as the cradle of civilization, has shaped human history for over 5000 years. Yet, in Flemish secondary history education, it is limited to two classes in the first grade. Despite the availability of digitized cuneiform tablets online, history teachers in Flanders face significant challenges in accessing, interpreting, and effectively integrating these materials into their lessons.

The CUNE-IIIF-ORM project addresses these challenges by creating a digital exhibition tailored to the needs of history teachers using the digital heritage collections from the Royal Museum of Art and History (RMAH) in Belgium. The digital exhibition offers an interactive, inquiry-based learning environment that allows history teachers and

students to explore Mesopotamian culture through curated, high-quality digital heritage resources.

This study draws on in-depth interviews with Flemish history teachers, an analysis of educational policy documents, and a review of teaching materials. Findings reveal that teachers acknowledge the importance of Mesopotamian history but struggle to engage students due to a lack of accessible, high-quality resources. Additionally, teachers report difficulties in visualizing Mesopotamian daily life and contextualizing historical sources. The disconnect between heritage collections, academic expertise, and classroom needs limits the effective use of Mesopotamian digital heritage.

In this project presentation we demonstrate how our project bridges these gaps by providing curriculum-aligned digital tools that promote historical thinking and inquiry-based learning. The curated digitised materials are made available on the digital exhibition via IIIF (International Image Interoperable Framework). Unlocking the materials through IIIF ensures not only access to these high-quality resources, catering to the needs of history teachers, but also interoperability with the collection management system of the RMAH. We argue that tailored, curriculum-aligned digital materials can transform the teaching of ancient cultures and serve as a model for integrating digital heritage into education more broadly.

Sh. Gordin

Measuring Scribal Literacy Through Quantitative Analysis: A Landmark-Based Approach to Cuneiform Signs

This study addresses a fundamental question in Assyriology and ancient literacy studies: can we develop quantifiable metrics to distinguish between different levels of scribal literacy based on the quality of handwriting? While paleographic studies have traditionally relied on descriptive terminology for sign variations, such assessments remain largely qualitative (Devecchi, Müller and Mynářová, 2015; Devecchi, Mynářová and Müller, 2019). The present paper proposes a methodological framework integrating geometric morphometrics (GM) and computational analysis to establish objective measurements of scribal competence across diverse textual corpora.

Our approach leverages a new annotation system to identify standardized landmarks on cuneiform signs—a method adapted from evolutionary biology that captures distinct shape variables in geometric forms (Okumura and Araujo, 2019). While previous computational approaches to ancient inscriptions faced challenges with cuneiform's three-dimensional nature, our methodology directly measures wedge stroke execution patterns, including angles, proportions, and spatial relationships between components from 2D images. Case studies presented will include both peripheral Akkadian (Hattusa, Ugarit, Alalakh, Canaan), and Mesopotamian examples (Old and Middle Assyrian).

By analyzing these quantifiable features across texts from different archives, genres, and historical periods, we can establish baseline metrics for expert scribal production and identify statistically significant deviations that may indicate varying literacy levels. Preliminary results from the annotation of signs for Mikulinsky & Alper et al. (2025) demonstrate that consistent measurements of wedge stroke variation correlate with known paleographic typologies, while also revealing subtle distinctions

previously undetected through traditional observational methods.

This research contributes to broader scholarly discourse on literacy acquisition in ancient societies by providing a data-driven framework for distinguishing between professional scribes, apprentices, and occasional writers. Moreover, it establishes a methodological foundation for investigating knowledge transmission patterns in scribal education and practice. By transforming traditional qualitative assessments into measurable parameters, this study enhances our understanding of how scribal competence manifested physically in cuneiform sign execution, offering new perspectives on literacy as both technical skill and cultural practice in the ancient Near East.

References:

Devecchi, E., Müller, G.G.W. and Mynářová, J. (eds) (2015) Current research in cuneiform palaeography: proceedings of the workshop organised at the 60th Rencontre Assyriologique Internationale, Warsaw 2014. Rencontre assyriologique internationale, Gladbeck: PeWe-Verlag.

Devecchi, E., Mynářová, J. and Müller, G.G.W. (eds) (2019) Current Research in Cuneiform Palaeography 2: proceedings of the workshop organised at the 64th Rencontre Assyriologique Internationale, Innsbruck 2018. Rencontre Assyriologique Internationale, Gladbeck: PeWe-Verlag.

Mikulinsky, R., Alper, M., Gordin, S., Jiménez, E., Cohen, Y., & Averbuch-Elor, H. (2025). ProtoSnap: Prototype Alignment For Cuneiform Signs. The Thirteenth International Conference on Learning Representations. https://arxiv.org/abs/2502.00129.

Okumura, M. and Araujo, A.G.M. (2019) 'Archaeology, biology, and borrowing: A critical examination of Geometric Morphometrics in Archaeology', Journal of Archaeological Science, 101, pp. 149–158. Available at: https://doi.org/10.1016/j.jas.2017.09.015.

P. Hacigüzeller

Archaeological Typologies and Language Acts in the Age of Metadata and Large Language Models

KEYNOTE

T. Homburg, C. Ziegeler, L. Wilhelmi

Extending Unicode Cuneiform with Cuneiform Sign Modifiers

Cuneiform signs have been added to the Unicode standard for quite some time but do not fully cover the number of attested sign variants in scholarly literature, as evidenced by the ORACC sign list or recent work of collecting cuneiform signs in Wikidata.

The latter counts 808 cuneiform signs that are not attested in Unicode cuneiform but in different sources, such as sign lists.

While many cuneiform signs are unique in shape and hence require their own single Unicode codepoint proposed in a Unicode proposal, many missing attested cuneiform signs in Unicode are simply allographs (e.g. TENU, GUNU versions of signs such as A TENU) or compounds (e.g. A TIMES AN) of cuneiform signs that have a Unicode code point. Having no unified way of describing these allographs and compound signs aside from a few compound signs having their own Unicode code points, font creators usually resolve to use the private Unicode code space, making so-defined fonts incompatible with other fonts attempting the same.

This leads to incompatibility issues with fonts, even when these fonts attempt to describe cuneiform signs of the same time period.

We propose a solution to include these missing allographs and compound signs into Unicode by defining them as combinations of already existing Unicode code points and sign modifiers with their own Unicode code point akin to modifiers common in Emoji fonts. This allows for the instant definition of more than 600 missing Unicode cuneiform signs and the opportunity to encode future undiscovered currently unattested cuneiform sign allographs and compound signs based on already defined Unicode cuneiform signs.

In our talk, we provide an overview of missing cuneiform signs, statistics on how many of these signs could be covered by the modification operators, the technical implementation details to apply said cuneiform sign modifiers, and the compatibility of existing (font-) implementations to our proposal.

Finally, we showcase a font implementation that uses the modifiers as a freely accessible prototype on a webpage.

We hope this talk can stimulate a discussion on the inclusion of sign modifiers in Unicode and could potentially result in a Unicode proposal adopting our suggestions.

E. Home

Detecting Scribal Preferences for Classifier Use in Hittite Using TF-IDF

In adapting the cuneiform script for Hittite, the earliest Hittite scribes inherited the system of graphemic classifiers traditionally know as ‘determinatives’. Far from being a fossilised quirk from their Mesopotamian predecessors, the classifier system employed by the Hittite scribes was used systematically and developed independently from Mesopotamian usage, displaying significant innovations and variation across manuscripts. The causes of this variation and the relative impact of variables such as the period or genre of the text on the use of classifiers are still not fully determined. In this paper I will assess the extent to which individual Hittite scribes and/or competing scribal circles have detectable preferences in their use of classifiers, and the significance of scribal habits compared to other variables that effect classifier use.

The corpus will consist of manuscripts with intact scribal signatures, all written by scribes who copied at least two manuscripts. To facilitate comparison across scribal circles, I will use manuscripts copied by scribes supervised by Anuwanza and Walwaziti, as well as the work of scribal team of Palluwuraziti and Pihawalwi. I will take a quantitative digital-based approach, using TF-IDF and the vector space model to measure the similarity of the manuscripts with regard to their use of classifiers. I will assess the extent to which the manuscripts cluster according to scribe and scribal circle, identify which lemmas exhibit significant variation in their classification across the

corpus, and determine whether we can ascribe such variation to differences in scribal habits. Finally, where possible, I will compare significant cases with duplicate manuscripts to assess the role of the individual scribe in classifier choice vs. the influence of the pre-existing manuscript tradition of a text. In doing so we will understand how flexible the classifier system employed by the Hittite scribes was, and the extent to which the use of classifiers was meaningful and intentional for them, rather than a matter of scribal convention.

Harrison Huang, B. Hensley, A. Anderson

Evaluating Online Cuneiform Sign Lists: Challenges and Advances in Digital Representation

The cuneiform writing system, employed extensively across the ancient Near East, has been subject to various digitization efforts aimed at preserving and facilitating scholarly research. This paper critically evaluates existing online cuneiform sign lists, particularly those from the Oracc Sign List (OSL) and the Cuneiform Digital Library Initiative (CDLI). Despite considerable overlap, discrepancies in sign formats and values hinder seamless integration, posing substantial barriers to effective computational analysis and machine learning applications. Originating in 2021 as a UC Berkeley Discovery project, this study sought to automate the transliteration-to-Unicode conversion of cuneiform texts.

After comparing the open source cuneiform sign lists, e.g. Nuolenna, Akkademia, TextFabric, CuneiML, and the OSL, we systematically quantified inconsistencies, ranking the overlapping matches against the OSL, a proven standard in the field. The results of this study were used to establish a harmonized cuneiform ontology, complete with references to the major published sign lists. Our research identified and addressed challenges stemming from differing encoding methods and syllabic values across online resources.

In collaboration with Timo Homburg in Mainz, Germany, our team collaborated in the Wikidata integration of cuneiform signs, their available fonts, and variant readings, cross-referenced with established scholarly publications. This unified approach provides a robust ontology accessible to computational tools, significantly benefiting language modeling applications and digital humanities research. The resultant open-access dataset includes comprehensive sign values and supports transliteration accuracy at both the document and token levels, advancing the study of Assyriology into the modern-day development of language modeling.

He Huang

A Pilot Experiment on Capturing Semantic Shift in Egyptian Languages Using Cross-Lingual Embeddings

This project presents a pilot study investigating diachronic semantic shift across different stages of the Egyptian languages. The Egyptian languages include several distinct stages in their development, including Old, Middle, and Late Egyptian (c. 2600–700 BC), which were written with hieroglyphic and hieratic scripts; demotic (c. 700 BC–400 AD), which used a different script and introduced new vocabulary; and Coptic (after c. 200 AD), which adopted an alphabetic script and borrowed words from Greek.

Although these languages are sometimes considered as separate languages, they belong to a continuous linguistic tradition. Considering the lexical changes in these languages over time, this study analyzes how words from one stage correspond to words with similar meanings in other stages, especially in cases where related concepts are not expressed by cognates or recognizably similar word forms.

A major challenge is that ancient Egyptian and Coptic languages are currently absent from most modern multilingual language models. Even within the Egyptian–Coptic language family, their writing systems are distinct: Egyptian is written without vowels and is transcribed into Latin script by scholars, whereas Coptic includes vocalization and uses Greek-derived letters. These differences make it difficult to represent different stages of this language family within a shared framework in one language model. Even semantically related terms may look entirely unrelated in their written forms and may also appear distant from each other in embedding space. Despite these difficulties, the Egyptian–Coptic language family presents a good opportunity for modeling semantic shift in historical data. Texts are preserved across clearly stratified periods, including the categorization of regionally used dialects in Coptic. These valuable datasets allow for careful analysis of meaning change over time.

This pilot project aims to provide a method for adapting a multilingual transformer-based approach to analyze semantic relationships between the languages, which are historically related but orthographically divergent. While the technical work is still in progress, the broader goal is to provide a case study of integrating ancient, low-resource languages into language models, and to offer a new perspective for future research on other historical languages.

T. Jauhiainen, H. Jauhiainen, K. Lindén

ANEE Idiolect Network Portal

The ANEE Idiolect Network Portal is a collection of web pages linking together more than one hundred thousand texts written initially using the Sumero-Akkadian cuneiform script. When creating the portal, we downloaded all 125,948 transliterated texts available at Oracc. Then, we re-created their cuneiform representations using the Nuolenna program, which was created for the Cuneiform Language Identification shared task in 2019. The Nuolenna sign list was recently updated for text dating experiments presented at RAI 2024. It creates Unicode cuneiform from the transliteration used in Oracc. For the Idiolect Network, we kept the 106,158 texts with at least 20 cuneiform signs after Nuolenna had transformed them.

For each text, we created a separate language model containing the relative frequencies, e.g., probabilities of sign n-grams from one to three. With a sign-n-gram-based Naive Bayes classifier, we used each model to measure the distance between it and each other text. Then, we used a new similarity measure, the Double Mutual Rank, to calculate an actual similarity score between all the texts. The Double Mutual Rank used is the average of the ranks two texts have on each other's list of most similar texts. These scores function as undirected edges in the network. However, due to their sheer number, more than 5,000 million, we distilled them into a subset of c. 33 million of the strongest edges. The final network contains those 105,631 texts, functioning as nodes, connected by this subset of edges. In the portal, each node has its own page, listing the most similar texts along with their corresponding similarity scores. Each list entry

contains a link to the page of the corresponding node, a link to the text entry at Oracc or CDLI, and some additional metadata extracted from the Oracc JSON export. In addition to an online portal, a package containing the scripts and data to create the portal contents will be made available in Zenodo.

E. Jiménez

Image Recognition and the Future of Cuneiform: Insights from the eBL Platform

KEYNOTE

F. Lamsens, G. R. Smidt, L. Foket

Cuneiform Tablets for Everyone. A Digital Scholarly Edition of a Cuneiform Corpus

Our project aims to implement computational approaches to the study of cuneiform tablets. However, prior to that it is paramount that the tablets are published in a manner befitting the use-cases. We believe that the International Image Interoperability Framework (IIIF) is the optimal for displaying our data.

At the Royal Museums of Art and History (RMAH) in Brussels there are c. 70 Old Babylonian tablets related to economical or documentary matters. These tablets were imaged with a Reflectance Transformation Imaging Dome, which produces 2D images where the lighting angles can be manipulated. From this image data, static images with preset lighting angles and visualized depth data have been generated. Accompanying the images, there are transliterations, links between the signs on the images and the transliterations, translations and associated metadata.

To ensure that the corpus is useful for scholars beyond cuneiform studies, we focused on making the data more accessible in a bespoke viewer. On one hand, we simplified the data, on the other we implemented links between textual and image data. At the same time, the rich environment allows for further analyses by specialists. By providing various visualizations with adjustable opacities that utilize the strength of digital images, it is possible to mimic the experience of examining the physical object itself. The simplifications designed to benefit non-specialists also help communicate the research assumptions and processes to cuneiform specialists.

The viewer currently supports the tablets from the RMAH, but other tablets can be imported. To facilitate comparisons, a side-by-side viewer is implemented, allowing users to compare tablets in various formats. When reading imported IIIF manifests, the viewer can also display cuneiform texts in a format commonly used for cuneiform studies, rather than the simpler machine-readable format.

Our presentation will showcase the functionality and interoperability of the viewer, and the reasoning behind our design choices.

E.-S. Lincke, T. B. Paul

Lexical Diversity in the Thesaurus Linguae Aegyptiae: Evaluating Corpus-Linguistic Metrics for Ancient Egyptian Texts

Quantitative approaches to the analysis of ancient Egyptian texts have a comparatively long history. As early as the 1970s, Fritz Hintze applied lexicostatistical methods through analyses of word frequency distributions. His work already addressed many concerns of present-day corpus-based stylistics and anticipated that far more could be achieved with machine-readable texts, yet it was largely ignored outside Egyptology and, within the field, revisited only decades later in a limited number of closely related studies.

The largest electronic Egyptian corpus, the Thesaurus Linguae Aegyptiae (TLA), is the digital successor to the slip archive of the Wörterbuch der aegyptischen Sprache (Dictionary of the Egyptian Language). While continuing that tradition, the TLA has evolved into a digital research infrastructure. It links tokenized and morphologically annotated texts to a set of lexical lists covering all phases of the Egyptian language and to extensive metadata. Despite its rich potential as a machine-readable resource, the TLA remains underutilized for corpus-linguistic purposes.

The TLA's raw data in JSON format can easily be integrated into a Python-based workflow. This opens the door to computational analyses that assess fundamental textual properties, such as lexical diversity—not only at the level of individual texts but at corpus level. In our contribution, we will evaluate established measures of vocabulary richness, such as Herdan's C, Hintze's S*, and contemporary metrics (e. g. vocd-D, MTLD). Our focus will be on their strategies to minimize the influence of text length, a pressing issue for ancient texts, which are typically shorter than contemporary works and often only partially preserved.

To assess the value of these indices, we will also demonstrate how they depend on parameters of the TLA's data model. In particular, we will discuss how definitions of "texts" and the hierarchical structure of the lemma lists (including sub-lemmata) affect the results.

While establishing baseline metrics for lexical diversity is not a novel undertaking, it remains essential, particularly in relation to text length and diachronic variation. Any meaningful corpus-linguistic analysis of ancient or under-resourced languages depends on understanding how their specific properties shape basic statistical behavior.

S. Miyagawa

Advancing Text Reuse Detection in Coptic Literature: A Comparative Study of Large Language Models and Traditional Computational Methods

Text reuse detection in Coptic literature presents unique challenges due to linguistic complexity, manuscript fragmentation, and the fluidity of intertextual practices in early Christian writings. This study evaluates the efficacy of fine-tuned large language models (LLMs) against established histroical text reuse detection tools—TRACER and Passim—in identifying biblical quotations, allusions, and paraphrases in the corpus of Besa, a pivotal 5th-century Coptic monastic leader.

These methods employ n-gram matching, similarity thresholds, and feature-based alignment to detect direct textual parallels. While effective for verbatim

quotations, these approaches often miss nuanced intertextuality, such as thematic echoes, semantic reworkings, or culturally embedded allusions—hallmarks of Besa's epistolary and pedagogical works. To address this gap, we analyze a curated dataset of 30 confirmed biblical intertexts from Besa's writings, manually annotated with four reuse categories: verbatim, near-verbatim, paraphrased, and thematic allusions.

The study benchmarks multiple LLMs, including GPT 4o, GPT o3-mini-high, GPT o1 pro mode, Claude 3.7 Sonnet, Gemini Advanced 2.5 Pro (experimental), Grok 2, and DeepSeek-R1, alongside THOTH AI—a custom model developed by the authors for Coptic and Ancient Egyptian natural language processing tasks based on Claude 3.5 Sonnet through retrieval-augmented generation (RAG) techniques. LLMs' capacity to process contextual, syntactic, and semantic relationships enables superior detection of indirect textual reuse compared to surface-level pattern matching. Preliminary results suggest that LLMs achieve higher recall in identifying paraphrased content and thematic parallels while maintaining precision comparable to classical tools for verbatim detection.

Methodologically, this work introduces a hybrid framework that integrates TRACER/Passim's scalable alignment algorithms with LLMs' contextual analysis, creating a two-stage detection pipeline. This approach reduces computational costs compared to pure LLM implementations while improving nuanced reuse identification. The framework's adaptability is demonstrated through its application to Sahidic and Bohairic Coptic variants, addressing dialectal diversity often overlooked in computational studies.

These findings advance computational philology by reconciling efficiency with interpretative depth, offering transformative potential for analyzing textually fluid traditions like Coptic literature and other ancient corpora. By bridging computational and hermeneutic methods, this research enriches both digital humanities frameworks and historical-critical scholarship, fostering new dialogues between ancient textual practices and modern analytical paradigms.

M. Naaijer, A. Wilson-Wright, B. Sober

An Integrated Methodology of Dating the Book of Jeremiah

International scholarship on the Hebrew Bible shares the unanimous consensus that its books, including the Book of Jeremiah, are literary products of the first millennium BCE. Nearly all else is in dispute, for there are no textual witnesses to the Hebrew Bible that stem from the biblical period itself, as the discovery of biblical manuscripts from the Dead Sea are all post-biblical.

The Book of Jeremiah has a complex history of composition and editing. It consists of many separate sections that deal with events in the late 7th / early 6th century BCE, but the different sections are not in chronological order and it is difficult to find an overarching structure in the book. There are many different proposed dates for the origin of the different parts of the book, and one of the striking issues is that people working from the perspectives of linguistics and redaction criticism often have different methodologies that lead to different results.

In our research we want to integrate these different perspectives and we expect that this will lead to new results. One approach is to investigate stylistic and linguistic

variation in Jeremiah using language models. Some of these models are freely available on HuggingFace, like HeBert, AlephBERT, DictaBERT and BEREL[1]. These models are trained on Rabbinic and Modern Hebrew, but there is no specialized model for Biblical Hebrew yet. The problem of Biblical Hebrew is that its corpus is small. Therefore, we train models on Biblical Hebrew ourselves and try to improve its performance by augmenting data, adding data from another language and improving the tokenization.

During the presentation we will show the results of the different ways of trying to improve the performance and we will compare our models with the models that are trained on texts from other language phases.

R. Nandy, M. Salve, A. Mitra

Encoded Empires: NLP driven Analysis of Achaemenid Persia in Historical Fiction

Historical Fiction entails the representation of bygone eras by reprising cultural aspects of a select time period. However, this genre of fiction looks at ancient history through a contemporary lens, often clouded by modern ideological biases. This paper proposes to apply Natural Language Processing (NLP) techniques to examine how historical fiction encodes and reconstructs ancient cultural narratives, focusing on The Persian Boy by Mary Renault and Creation by Gore Vidal. Through the use of Named Entity Recognition (NER), Latent Dirichlet Allocation (LDA)-based topic modelling, and sentiment analysis, the research traces linguistic patterns that influence the representation of Achaemenid Persia. The language structure and contextual embedding coupled with understanding character building, these NLP tools will enable a deeper appreciation of the primary texts. The research hopes to connect the findings with narratological study for acquiring data driven comprehension of the texts as historical 're-presentation'. This study hypothesizes that historical fiction embeds modern ideological biases into its linguistic and thematic structures, shaping contemporary perceptions of ancient cultures through selective representation and emotive weighting. By combining computational text analysis with literary criticism, this research demonstrates how NLP provides a systematic framework for interrogating historical fiction's construction and mediation of cultural memory. In doing so, it offers new insights into the intersection of digital humanities, historical representation, and narrative strategy.

A. Narciso, L. Verderame

Ebla 2.0

The paper addresses the use of digital technologies in the analysis of the corpus of ancient texts from the so-called "Royal Archives" of Ebla. The rich and diverse documentation from the archives constitutes the oldest palatine archives of the Near East scientifically excavated and an opportunity for investigating the rise of early states. However, a few aspects of the documents—most notably, the lack of year dates—raise significant concerns for an extensive study of these sources. In this way, the support of digital analysis aids in the partial resolution of these problems.

---

[1] https://huggingface.co/avichr/heBERT, https://huggingface.co/onlplab/alephbert-base, https://huggingface.co/dicta-il/dictabert, https://huggingface.co/dicta-il/BEREL_3.0.

In this paper, we will present the work done in this direction in the last year and a case study. Within the Missione Archeologica Italiana in Siria, in the last year advancements have been reached through a) the digitization of the field notes, b) the photos archive, and c) the catalogue of the unpublished materials. Furthermore, the d) annotation of administrative documents and their e) formalization to be elaborated in network analysis and other software allow the creation of clusters and relationships that integrate a relative chronology. Lastly, e) a computerized glossary of Eblaite, sumerograms, and Akkadograms is fed by the indexing of secondary literature.

As a case study of the results of this preliminary work, the network analysis of texts from the period of the "vizir" Arrugum is presented.

K. V. Nilsson

The Social Network of Iltani

Iltani is a name which appears several times in the Late Old Babylonian period (ca. 1700-1500 BCE). Documentation from this period especially transcribes the name Iltani to both the title of dumu.munus lugal, "daughter of the king," and lukur, "naditu-priestess." There is abundant documentation spread over a large expanse of time where Iltani, the daughter of the king and the naditu, is recorded, so there are various theories to contend the records. While some suggest Iltani is the same person or the same couple of persons throughout the record (Renger 1999), others suggest that Iltani is the name of an Old Babylonian institution, not merely a person (Richardsson, 2017). In order to bring some clarity to these issues, this essay bases itself upon Social Network Analysis (SNA) to unravel the web of actors and relations of the network recorded in 52 documents of Iltani during the Late Old Babylonian Period. Although the focus of this project is the network of Iltani, this project soon came to be a comment and a deeper discussion of the difference between the social networks of women and men. In particular, this essay aims to discuss the physical and geographical restrictions of cloistered women and argue for the cloister's continued use in the Late Old Babylonian period. By creating different visualizations of the social network with Gephi, this essay interconnected Iltani in a network with 185 other individuals. This essay suggests that Iltani and her network should be integrated into a framework of others of her gender and social class (naditu and daughter of the king), not into the networks of men, as the limitations of both the breadth and depth of the relations of Iltani bases upon her restrictions as a woman and a cloistered naditu. Therefore, this essay encourages further research on the social networks of the naditu and other ancient women.

M. Ong

Digitization and Schematic Representation of Place Names in the RGTC Series

I discuss a recent approach to digitizing the printed volumes of the series Repertoire Geographique Des Textes Cuneiformes (RGTC), which catalogues all known place names in the cuneiform sources. My discussion includes the training of suitable OCR models, recruitment of ancillary tools and resources for digitization, and development of workflow methods that can be implemented by a limited number of individuals. I also

discuss how to extract and schematically represent `geographically relevant information' related to the place names in the volumes, ultimately so that such information can be imported into a Wikibase hosting a linguistically annotated corpus of Akkadian.

Ch. Palladino

Mapping Uncertain Ground: Ancient Spatial Knowledge Across Text, Data, and Materiality

KEYNOTE

A. Romach

Gendered Voices in Cuneiform: A Computational Analysis of the Old Assyrian Letter Corpus

This paper applies stylometric methods to examine gender differences in the Old Assyrian letter corpus. Drawing on 890 transliterated letters from CDLI–76 sent by women, 814 by men, and 140 involving women as senders or recipients–I employ statistical methods to investigate whether women's writing exhibits distinctive patterns in orthography and content as suggested by previous scholarship.

Using text vectorization with TF-IDF scoring and visualization through t-SNE dimensionality reduction, I transform textual features into measurable parameters. The computational evidence reveals that orthographical variation is consistent across genders—both men and women employ diverse spellings of the same words with approximately equal frequency. This observation enriches our understanding of orthographical practices in this period, suggesting that spelling variations represent a broad characteristic of the corpus rather than a gender-specific phenomenon.

Noticeable differences emerge in lexical content and topic distribution in letters in which women are senders or recipients. Women's letters focus more on household matters, food supplies, and interpersonal relationships. Men's correspondence shows greater emphasis on trade-specific terminology, expressed often with logographic writing. Function words, particularly negations and interrogatives (e.g. lā, ula, miššum), prove more distinctive in identifying female authorship, a phenomenon that is also evident from other historical periods (e.g. Verhoeven and Daelemans 2019; De Gussem 2021; Suddaby and Ross 2023).

These findings have significant implications for understanding literacy acquisition in the Old Assyrian period. The orthographical similarities suggest Assyrian children were instructed in writing without a gender distinction. Beyer (2019, 2021), who studied paleography in the Old Assyrian corpus, already identified siblings with similar handwriting that is different from their father's, suggesting either another family member tutored them or they were educating outside of the family. The computational evidence of women's equivalent writing competency can suggest that they may have been taught by their mother.

This research demonstrates how computational methods complement traditional philological approaches to ancient texts, offering quantifiable evidence that both refines and extends previous qualitative observations.

References:

Beyer, Wiebke. 2019. “The Identification of Scribal Hands on the Basis of an Old Assyrian Archive.” Phd, University of Hamburg.

———. 2021. “Teaching in Old Babylonian Nippur, Learning in Old Assyrian Aššur?” In Education Materialised: Reconstructing Teaching and Learning Contexts through Manuscripts, edited by Stefanie Brinkmann, Giovanni Ciotti, Stefano Valente, and Eva Maria Wilden. Studies in Manuscript Cultures. De Gruyter, pp. 15–32. https://doi.org/10.1515/9783110741124-003

De Gussem, Jeroen. 2021. “Larger Than Life? A Stylometric Analysis of the Multi-Authored ‘Vita’ of Hildegard of Bingen.” Interfaces: A Journal of Medieval European Literatures, no. 8 (December): 125–59. https://doi.org/10.54103/interfaces-08-08.

Suddaby, Lee, and Gordon J Ross. 2023. “Did Mary Shelley Write Frankenstein? A Stylometric Analysis.” Digital Scholarship in the Humanities 38 (2): 750–65. https://doi.org/10.1093/llc/fqac061.

Verhoeven, Ben, and Walter Daelemans. 2019. “Discourse Lexicon Induction for Multiple Languages and Its Use for Gender Profiling.” Digital Scholarship in the Humanities 34 (1): 208–20. https://doi.org/10.1093/llc/fqy025.

E. Roßberger, S. Hageneuer

KIŠIB. Digital Corpus of Ancient West Asian Seals and Sealings

The long-term project KIŠIB. Digital Corpus of Ancient West Asian Seals and Sealings started in 2025 in Munich and Berlin and receives funding from the Bavarian Academy of Sciences and Humanities and the Berlin-Brandenburg Academy of Sciences and Humanities.

More than a hundred years of archaeological, philological, and art historical research on West Asian seals and sealings have contributed significantly to our knowledge about the region’s past. However, access to this knowledge remains limited with the original artefacts dispersed across numerous collections worldwide.

KIŠIB seeks to overcome these challenges by establishing a representative digital corpus of c. 80,000 seals and sealings. It will aggregate and curate seal-related data from various sources to facilitate high-quality knowledge exchange in close collaboration with seal-holding institutions, other projects active in digital ancient Near Eastern studies, and with colleagues from West Asian countries.

The project presentation will explain the objectives and strategies envisioned for the project with a focus on its digital infrastructure.

F. Shamsian

Indirect Translation of Thucydides into Persian: Comparisons with Mediating Texts Through Translation Alignment

This paper reports on the methodological challenges in evaluating indirect translations from Ancient Greek into Persian using word-level translation alignment. Here, I examine the Persian translation of Thucydides's "The Peloponnesian War," which was created through German (1991 & Heilmann, 1760) and English (1916) mediating texts by Mohamad Hassan Lotfi (1998), as well as two other indirect translations from 1952 and 1954.

Indirect translation is a common practice in Persian, particularly for Classical Greek texts, in which the Greek text is translated into Persian based on translations in other languages such as English, French, or German. In contrast, direct Ancient Greek-Persian translations remain exceptionally rare. The 1998 translation of Thucydides exemplifies the compilative approach (Ivaska & Paloposki, 2018) frequently employed in Persian indirect translations, where multiple mediating texts, rather than a single mediating translation, shape the final translation. This complex translation process creates unique challenges for traditional translation assessment methods. The ultimate goal of this evaluation is to assess the feasibility of creating a parallel corpus using indirect translations.

By mapping relationships between the Greek source text, the German and English mediating texts, and the target text using the Ugarit translation alignment tool (Yousef et al., 2022), this study examines how semantic content, stylistic features, and linguistic structures were transformed in the translation process. Both quantitative metrics and qualitative analysis are employed to assess the relationships between texts and evaluate the final translation's fidelity to the Greek source text.

I conclude with observations on the broader implications of this methodology for building parallel corpora that incorporate indirect translations and address whether a multilingual corpus can be constructed despite the absence of direct source-target relationships, identifying specific linguistic and semantic challenges that emerge in this process.

References:

Heilmann, J. D. (1760). Des Thucydides acht Bücher der Geschichte aus dem griechischen mit vielen kritischen Anmerkungen übersezt.

Ivaska, L., & Paloposki, O. (2018). Attitudes towards indirect translation in Finland and translators' strategies: Compilative and collaborative translation. Translation Studies, 11(1), 33–46. https://doi.org/10.1080/14781700.2017.1399819.

Thucydides. (1916). The Peloponnesian War (R. Warner, Trans.). Penguin Books. http://archive.org/details/in.ernet.dli.2015.461316.

Thucydides. (1991). Geschichte des Peloponnesischen Krieges (M. Fuhrmann, Ed.; G. P. Landmann, Trans.). Deutscher Taschenbuch Verlag.

Thucydides. (1998). Tārikh-e Jang-e Peloponezi (M. H. Lotfi, Trans.). Khārazmi.

Yousef, T., Palladino, C., Shamsian, F., & Foradi, M. (2022). Translation Alignment with Ugarit. Information, 13(2), 65. https://doi.org/10.3390/info13020065.

Sendyka, D. Robinson, J. Tenney, N. Lawrence

Grounded Theory and the Extended Mind as a Method for Managing Assyriological Legacy Data

Digital approaches to cuneiform tablets have focused on linguistic and lexical content—in particular machine translations of poetry and myth, lexical investigations and annotation, and 2D photography. Together, they increased access and developed a "Digital Assyriology" aligned with the lexical, philological, and prosopographical goals of the field.

We report on a pilot project conducted jointly by AI@cams and the Department of Archaeology at Cambridge meant to augment and externalise the reasoning processes of Assyriologists. We conducted a field study with the observational methods of Grounded Theory and the Three Esses Framework to analyse the workflow of an Assyriologist who is interested in historical, rather than philological, research questions. We identified an opportunity to develop a suite of machine learning tools tailored to his research questions and environment, thus adding our own contribution to the meaning of "Digital Assyriology". In addition, we consider the conditions under which an Assyriologist might be able to leverage machine learning and AI without any specialized training and/or the assistance of data science experts.

G. R. Smidt

Inline or in the Line: Quantitative analyses of layout in Old Babylonian economical texts using coordinates from image annotations

Old Babylonian cuneiform tablets are handwritten objects, especially those pertaining to economical and documentary matters. They are predominately made out of clay, which makes them plastic and malleable. One result is that they require the manipulation of lighting angles to be deciphered, but it also results in freedom of textual formatting. An element of formatting that is commonly used, but poorly studied, is the use of margins and indentations. In a contract it can indicate family relations or a verb, where in lexical lists it is often used to mark the introduction of a new word group.

My project has more than 300 high-quality images (2D+) available of various Old Babylonian corpora. The signs of these tablets are annotated on the images, which provides rich analytical material for image recognition and precise positional data for formatting analyses. By combining textual information with positional data, I have developed a methodology that can display a quantitative understanding of the link between the two. Understanding layout of tablets can answer questions of meaning embedded in formatting and help with the reconstruction of texts.

This poster will present a methodology that can be used to explore the links

between positional and linguistic data. I will showcase examples of how the use of space in Old Babylonian cuneiform tablets can have specific meanings.

C. Swaelens, I. De Vos, E. Lefever

Between the Lines: Exploring Similarity in Byzantine Metrical Paratexts

Book epigrams, otherwise known as metrical paratexts (Genette, 1987), furnish the reader of a manuscript with information regarding the text and/or the manuscript itself. These book epigrams are part of a literary tradition that has been passed down through various cultures and languages, including Arabic (Quiring-Zoche, 2013), Greek (Ricceri et al., 2023), and Latin (Meesters, 2022). The present research primarily focuses on the Byzantine Greek tradition of book epigrams, with the possibility for expansion into a multilingual context.

The objective of our research is to identify analogous verses within the corpus of the Database of Byzantine Book Epigrams (Demoen et al., 2022), thereby offering novel insights into their written tradition. To detect related verses, however, it is important to note that the concept of similarity is inherently open to interpretation. To address this, we are developing a range of similarity metrics, from a strict orthographic edit distance (Deforche et al., 2024), to edit distances based on parts-of-speech or lemmas, as well as a vector-based similarity metric.

In order to integrate linguistic information such as part-of-speech, morphology, or lemmas, new annotation methods had to be developed, as existing taggers did not perform well on this corpus of unedited Greek data (Anonymous, XXXX). A gold standard of unedited Greek and a linguistic annotation pipeline have already been developed, yielding state-of-the-art results. The integration of linguistic information has been demonstrated to relax the strict orthographic distance, thereby providing more desirable similarity clusters of related verses/epigrams (Anonymous, XXXX).

The present study focuses on the detection of semantic similarity based on word embeddings. The initial benchmark for the evaluation of this subjective task was created by annotating 300 verse pairs using pairwise comparison (David, 1988). The inter-annotator agreement of 0.65 demonstrates the reliability of this benchmark for future research. The development of the similarity detection algorithm involves the exploration of various embedding models and distance metrics to ascertain their effectiveness in capturing semantic similarity.

E. Verwimp, G. R. Smidt, H. Hameeuw, K. De Graef

Cun-AI-form: OCR Algorithms for Cuneiform Writing: An Analysis of the Challenges and Opportunities on Old Babylonian Clay Tablets

Current algorithms for Optical Character Recognition (OCR) are successful in the detection of many handwritten texts, from historical documents [1] to handwritten notes in various writing systems (e.g. [2]). In this work, the potential for these algorithms to locate and identify cuneiform signs inscribed on Old Babylonian clay tablets is evaluated. Earlier work describes the creation of a cuneiform sign corpus from 2D+ images with various lighting angles. It lists the challenges particular to the detection and

identification of cuneiform signs (e.g. the 3D nature of cuneiform writing, overlapping signs, signs in multiple directions etc.) [3]. These challenges are based on human expert opinion, but there does not yet exist a qualitative analysis of trained OCR models on Old Babylonian cuneiform writing. This paper analyzes the mistakes made by such OCR models. In particular, it evaluates how the types of errors a machine learning model makes coincide with those listed in [3] and whether more, machine-specific challenges, exist. The current corpus consists of high-quality images from 12 different light sources, made on tablets that originate from two nearby cities, both belonging to the same scribal tradition. The cuneiform corpus is extensive, the tablets are from different geographical areas and periods, and of various text genres. Many of these have previously been imaged using only one lighting condition, e.g. flatbed scans. The second part of this work explores the transferability of a model trained on a set of cuneiform tablets to an out of domain set of tablets, with a particular focus on lightning conditions and geographical variation. Both the error analysis and the transferability study of cuneiform-OCR models, guide the development of a semi-automatic cuneiform sign annotation tool. This will assist experts in the study of cuneiform tablets. The paper concludes with initial results on the automated annotation process.

References:

[1] Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., ... & Puppe, F. (2019). OCR4all—An open-source tool providing a (semi-) automatic OCR workflow for historical printings. Applied Sciences, 9(22), 4853.

[2] Huang, J., Pang, G., Kovvuri, R., Toh, M., Liang, K. J., Krishnan, P., ... & Hassner, T. (2021). A multiplexed network for end-to-end, multilingual OCR. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4547-4557).

[3] Hameeuw, H., De Graef, K., Smidt, G. R., Goddeeris, A., Homburg, T., & Kumar Thirukokaranam Chandrasekar, K. (2024). Preparing multi-layered visualisations of Old Babylonian cuneiform tablets for a machine learning OCR training model towards automated sign recognition. it-Information Technology, 65(6), 229-242.

E. Yavasan, M. Molina

Universal Dependencies in Hittite

This paper introduces a project focused on developing Universal Dependencies (UD) for Hittite that maintains cross-linguistic comparability and at the same time accurately represents language-specific phenomena. To date, only one small Hittite UD treebank exists, based solely on textbook examples (Andersen, Rozonoyer 2020). A UD treebank based on critical editions of original Hittite texts has yet to be published.

The development of advanced frameworks, such as Universal Dependencies (UD), has brought standardisation into the analysis of ancient languages. In the field of ancient Near East (ANE) languages, UD standards have been applied first and foremost to Akkadian (Luukko et al. 2020; Ong 2024), which benefits from a widely digitised corpus. UD also exist for Ancient Greek, Latin, Biblical Hebrew, Classical Armenian. Hittite, as

one of the prominent languages of the second millennium's Asia Minor, deserves a place in this row.

Our work builds upon a prior pilot project within the PROIEL framework (Inglese et al. 2018) with addition of the text-mining solutions by E.Yavasan (Yavasan, Gordin forthcoming). The corpus includes Hittite letters, as in Hoffner (2009), as well as Edel (1994) and Hagenbuhner (1989) not covered by Hoffner (2009). The texts are annotated in CONLL-U format, with planned conversion to CONLL-U+. Visualization is carried out using the PALMYRA 2.4 platform, and all annotated files are available via GitHub.

The annotation of Hittite presents language-specific challenges not encountered in modern or more extensively annotated ancient languages. One key issue is the fragmentary nature of the texts, addressed by tagging missing portions with the 'FRGM' label. Additional complexities include multilingual lemmatization—due to the presence of Hittite, Akkadian, and Sumerian heterograms—as well as the annotation of clitic chains, determiners, and correlatives.

Lemmatization employs three distinct fields (Hittite, Akkadian, Sumerian), at least one of which must be filled. These are merged to comply with the CONLL-U format. Clitics are annotated as separate tokens with the XPOS tag CL, functioning as objects, predicates, or discourse particles. Determiners are marked as dependents, given their semantic role in clarifying noun phrases. Hittite relative clauses, inherently correlative, are treated as components of complex sentences.

M. Wamposzyc

Exploration and Development of Pictograms for Cuneiform Artefacts

The proposed poster elaborates on the development of and conclusions from three years of educational practice in leading a university level Visual Information Design module, exemplified through work and process books from three consecutive cohorts of Year 2 Scottish students. The particular interest within the topic will relate to the epistemic issues of 2D/3D data representation and development of visual language for pictographic sets for cuneiform artefacts stored in the National Museum of Scotland.

W. Wendrich

Recording Intangible Cultural Heritage Through Movement and Touch

KEYNOTE

P. Zadworny, Sh. Gordin

An Automated Approach to Administration at the Dawn of Writing: Labelling Inventories and Assignments among Archaic Accounts

Proto-cuneiform texts are our earliest evidence for written administration in history. After 100 years, we have a decent understanding of their content, yet our current typologies ignore an important aspect: their use in administration.

In this paper, we further develop our earlier experiments with computational

approaches to labeling archaic accounts (Zadworny & Gordin 2025). Our new method will use support vector machine (SVM) algorithms to include the administrative use of archaic texts in the labeling process. We distinguish two categories: assignments and classifications – as they contain entirely different information, and require different methods of study, traditional or automated.

In addition to being an infrastructural improvement, this paper is part of a series of experiments to apply computational methods to an edge case: proto-cuneiform is a non-linguistic writing system, which offers us a chance to get unique insight about our understanding of text types and genres.