UNIVERSITY OF PIRAEUS – DEPARTMENT OF INFORMATICS

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ -- ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

MSc «Computer Science»

ΠΜΣ « Πληροφορική»

MSc Thesis

Μεταπτυχιακή Διατριβή

Thesis Title: Τίτλος Διατριβής:	**An Undeciphered Script in the Age of AI: A Corpus-Constrained Computational Analysis of Linear A** Ένα άγνωστο σύστημα γραφής στην εποχή της ΤΝ: Υπολογιστική ανάλυση της Γραμμικής Α με περιορισμένα δεδομένα
Student's name-surname: Ονοματεπώνυμο φοιτητή:	**Nikolaos Briakos** Νικόλαος Μπριάκος
Father's name: Πατρώνυμο:	**Apostolos** Απόστολος
Student's ID No: Αριθμός Μητρώου:	ΜΠΠΛ/2319
Supervisor: Επιβλέπων:	**Ioannis Venetis, Assistant Professor** Ιωάννης Βενέτης, Επίκουρος Καθηγητής

May 2026 / Μάιος 2026

3-Member Examination Committee

Τριμελής Εξεταστική Επιτροπή

Ioannis Venetis	**Evangelos Sakkopoulos**	**Dionysios Sotiropoylos**
Assistant Professor	**Associate Professor**	**Associate Professor**
Ιωάννης Βενέτης	Ευάγγελος Σακκόπουλος	Διονύσιος Σωτηρόπουλος
Επίκουρος Καθηγητής	Αναπληρωτής Καθηγητής	Αναπληρωτής Καθηγητής

Declaration

I declare that this thesis is my own work. All data sources are cited in the text and in Appendix A. Every statistic is labelled † (computed from a downloaded file with logged SHA256), ‡ (from published literature, cited with author and year), or ° (illustrative only).

Evidence tier notation used throughout this thesis:

† *Verified* — computed directly from the downloaded corpus (SHA256 logged).
‡ *Estimated* — from published literature (cited with author and year).
° *Schematic* — illustrative; data insufficient for statistical inference.
⋆ *Model output* — AI prediction; not a decipherment claim.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Assistant Professor Ioannis Venetis of the Department of Informatics, University of Piraeus, for his guidance throughout this project, for his careful reading of successive drafts, and for his insistence on methodological rigour and epistemic honesty. His feedback shaped this work substantially.

I am grateful to Professor John G. Younger (University of Kansas) for making the lineara.xyz digital corpus publicly available. Without this resource, the corpus-level analyses in Chapters 3 through 6 would not have been possible.

Περίληψη

Η παρούσα διατριβή εξετάζει ένα πραγματικό και ακόμη αναπάντητο ερώτημα της υπολογιστικής γλωσσολογίας: μπορεί η τεχνητή νοημοσύνη να συμβάλει στην αποκρυπτογράφηση της Γραμμικής Α, της άγνωστης γραφής της Μινωικής Κρήτης (περίπου 1800--1450 π.Χ.); Χρησιμοποιώντας ένα σώμα 419 πινακίδων από τη βάση δεδομένων lineara.xyz (παραγόμενο από το GORILA, SHA256: b7b383b93db55b50…, $N = 2{,}481$ σύμβολα), η εργασία απαντά στο ερώτημα σε τρία στάδια.

Τι μπορεί να πετύχει η ΤΝ. Το σώμα παρουσιάζει κλασική κατανομή Zipf στα σύμβολα, συστηματικές θέσεις εμφάνισης, διαφοροποίηση μεταξύ διαφορετικών τύπων κειμένων (JSD $= 0{,}0944$ bits, $p = 0{,}018$), σημαντικές διακυμάνσεις στο μήκος λέξεων ανά τόπο (Cohen's $d = 0{,}692$ μεταξύ Φαιστού και Χανίων), καθώς και 23 χαρακτηριστικά «δάκτυλα» γραφών με ιδιαίτερες προτιμήσεις διγραμμάτων. Το πιο σημαντικό: ένας μη εποπτευόμενος αλγόριθμος ανίχνευσης τύπων κατάφερε να αναγνωρίσει 5 από τα 9 στοιχεία του γνωστού «τύπου σπονδής» σε 31 λίθινα αγγεία από 14 διαφορετικούς τόπους, χωρίς καμία προηγούμενη γνώση του τύπου.

Τι μπορεί ακόμη να μάθει η ΤΝ. Ένα μικρό Transformer (4 στρώματα, μοντέλο masked language modeling, $\sim$2 εκατ. παράμετροι) εκπαιδευμένο σε GPU πέτυχε μέγιστη ακρίβεια 90,2% και ανακατασκεύασε σωστά τον πιο συχνό συνδυασμό συμβόλων (KU-RO, $p = 0{,}59$ και στις δύο κατευθύνσεις). Το μοντέλο έχει λοιπόν κατανοήσει τη διανεμητική δομή του σημειολογικού αποθέματος της Γραμμικής Α. Μια πολυτροπική ανάλυση που συνδύασε οπτικά χαρακτηριστικά (PIL) με ενσωματώσεις συμβόλων δεν εντόπισε ουσιαστική γεωγραφική ομαδοποίηση· ο φαινομενικός διαχωρισμός στον οπτικό χώρο οφείλεται σχεδόν αποκλειστικά σε συνθήκες φωτογράφισης ($r = -0{,}990$ με τη φωτεινότητα, $R^2 = 0{,}98$).

Τι δεν μπορεί ακόμη να κάνει η ΤΝ. Πειράματα βαθμονόμησης με συνθετικό υλικό Γραμμικής Β, φτιαγμένο ώστε να ανταποκρίνεται πιστά στις γνωστές κατανομές συχνότητας, έδειξαν ότι με 2.481 σύμβολα η ακρίβεια αντιστοίχισης με βάση τη συχνότητα φτάνει μόλις το 13% (top-1). Η Bayesian προσέγγιση για την απόδοση φωνητικών τιμών παρήγαγε σχεδόν επίπεδες κατανομές πιθανοτήτων (μέγιστη απόκλιση μόλις 2,1 φορές από την ομοιόμορφη). Το μοντέλο δεν είναι ακόμη σε θέση να μετατρέψει τις στατιστικές δομές που έχει μάθει σε αξιόπιστες φωνητικές αντιστοιχίσεις, χωρίς μεγαλύτερο όγκο γνωστών παραδειγμάτων. Σύμφωνα με τα ίδια πειράματα, το πρακτικό όριο για χρήσιμα αποτελέσματα βρίσκεται γύρω στις 10.000 σύμβολα --- δηλαδή χρειαζόμαστε περίπου 7.500 σύμβολα περισσότερα από όσα διαθέτουμε σήμερα.

Κύριο συμπέρασμα. Ο πραγματικός περιορισμός δεν είναι η πολυπλοκότητα της γραφής ούτε οι δυνατότητες των μοντέλων, αλλά το μικρό μέγεθος του διαθέσιμου σώματος. Η τεχνητή νοημοσύνη μπορεί ήδη να κατανοήσει τη διανεμητική δομή των ακολουθιών συμβόλων της Γραμμικής Α. Δεν μπορεί ακόμη να αποκρυπτογραφήσει τον φωνητικό της κώδικα, όμως το χάσμα είναι μετρήσιμο και μικραίνει σταθερά με κάθε νέα πινακίδα που ανακαλύπτεται.

Abstract

This thesis addresses a real, unanswered question in computational linguistics: can artificial intelligence help decipher Linear A, the undeciphered Bronze Age script of Minoan Crete (c. 1800–1450 BCE)? Using a corpus of 419 tablets from the lineara.xyz database (GORILA-derived, SHA256: `b7b383b93db55b50`…, $N = 2{,}481$ sign tokens), we answer this question in two parts.

What AI can do. The corpus displays Zipfian sign distributions, systematic positional biases, register divergence (JSD = 0.0944 bits, permutation $p = 0.018$), cross-site word-length variation (Cohen's $d = 0.692$, Phaistos vs. Khania, administrative tablets only), and 23 scribal fingerprints with distinctive bigram preferences — all †. Critically, an unsupervised formula detection algorithm recovers 5 of 9 human-identified libation formula elements from the 31-tablet stone vessel corpus spanning 14 sites, without prior knowledge of the formula. These are genuine AI contributions to the epigrapher's toolkit.

What AI can also learn. A TinyTransformer (4-layer masked language model, ∼2M parameters) trained on GPU achieves 90.2% peak validation accuracy and correctly reconstructs the most frequent Linear A sign collocation (`KU-RO`, $p = 0.59$ in both directions), demonstrating that the model has learned the distributional structure of the sign inventory*. A multimodal analysis combining PIL visual features with distributional sign embeddings identifies no meaningful geographic clustering: apparent separation in visual PC1/PC2 space is attributable to photographic exposure conditions (visual PC1 $r = -0.990$ with image brightness, $R^2 = 0.98$)*.

What AI cannot yet do. A Linear B calibration experiment using a synthetic corpus faithful to published frequency distributions [11, 18] shows that frequency rank-matching accuracy at $N = 2{,}481$ tokens is 13% top-1 (6× random chance, below the 21% 'useful' threshold). Bayesian phonetic inference yields near-uniform posteriors (max 2.1× over the uniform baseline)*: the model cannot extrapolate the learned distributions to phonetic value assignment without a larger corpus of known correspondences. The calibration experiment places the useful accuracy threshold at approximately $N = 10{,}000$ tokens — roughly 7,500 tokens beyond the current corpus.

The core finding is tractable: the binding constraint is corpus size, not script complexity or model inadequacy. AI can already learn the distributional structure of Linear A sign sequences; it cannot yet crack the phonetic code, but the gap is quantifiable and narrowing with every new tablet discovered.

Contents

List of Figures

List of Tables

1 Introduction

Deciphering ancient scripts that have long resisted interpretation is a formidable challenge at the intersection of archaeology, linguistics, and computer science. Linear A — the Bronze Age script of Minoan Crete (c. 1800–1450 BCE) — remains one of the most prominent undeciphered writing systems. Despite sharing some symbols with the later Linear B script (deciphered in 1952 as an archaic form of Greek), the content of Linear A has eluded understanding for over a century [24]. The inability to read Linear A impedes our knowledge of the Minoan language and culture.

Traditionally, breakthroughs in script decipherment have come through painstaking human analysis, leveraging clues like bilingual texts or known related languages. In recent years, however, artificial intelligence (AI) and machine learning have emerged as promising tools to assist in cracking such ancient mysteries. This chapter surveys the challenges posed by undeciphered scripts like Linear A and explores how modern AI techniques — from computer vision and natural language processing (NLP) to advanced neural network architectures — can contribute to the decipherment process.

1.1 Challenges of Undeciphered Scripts

Undeciphered writing systems present numerous inherent challenges that any AI-driven solution must confront. A core difficulty is data scarcity and fragmentation. The challenges of deciphering Linear A are compounded by the extremely limited size of its corpus.

With approximately 7,000 signs across ∼1,500 inscriptions, the dataset is minuscule compared to the vast corpora typically used in machine learning.

For context, ImageNet, a standard dataset for image recognition, contains over 14 million images, while Wikipedia's text corpus includes billions of words. This scarcity hampers the capacity of machine-learning models to generalise; they risk memorising the corpus rather than capturing underlying structure, which would yield 'decipherments' that collapse as soon as unseen material appears.

Similarly, the Indus script of the Bronze Age Indus Valley civilisation (c. 2500–1900 BCE) consists of very brief inscriptions (most just 4–5 symbols long), providing sparse data for analysis [20]. The brevity and formulaic nature of such texts limit the statistical information available about symbol patterns, making it hard for algorithms (or humans) to infer linguistic structure. Indeed, whether the Indus script encodes language at all was debated — the computational analysis by Rao et al. [20] found its sequential character patterns were closer to those of linguistic scripts than to random or non-linguistic sequences. This suggests that even with limited data, undeciphered scripts can exhibit non-random structure, but detecting those subtle patterns requires careful modelling.

Another challenge is lack of linguistic context. For known scripts, decipherment often leverages bilingual inscriptions (e.g. the Rosetta Stone for Egyptian hieroglyphs) or known related languages. Linear A has no clearly bilingual texts, and its underlying language is unknown. Without a 'key' or a large corpus to derive one, AI models risk generating spurious mappings. The model's assumptions must therefore be guided by what is plausible (or at least possible) for human language. For example, any proposed decipherment should exhibit consistent mappings of signs to sounds or words and should not produce gibberish when translated. This constraint is hard to enforce with pure machine learning, which may find patterns in noise. Domain knowledge — such as probable phonetic values borrowed from Linear B or the expectation that certain symbols might represent numbers or word separators — becomes crucial background input to an AI system.

Finally, no evaluation metric is straightforward for undeciphered scripts. In a supervised machine-learning task, we could measure accuracy against ground truth labels — but here the 'ground truth' (the actual meaning of Linear A texts) is unknown. Thus, we must rely on indirect metrics to evaluate a model's output. For instance, does the output text exhibit statistical regu-

larities (like linguistic redundancy) similar to real languages? Does it align in any way with known proper names or loanwords? Part of this thesis is devoted to introducing quantitative measures (like Recall@k for masked predictions, and CER for sequence reconstructions) that can serve as proxies for decipherment quality in the absence of a known correct translation.

1.2 AI in Epigraphy and Historical Linguistics

AI and machine learning have increasingly been applied to problems in epigraphy (the study of inscriptions) and historical linguistics. These applications fall generally into three categories: reading (computer vision to recognise characters from damaged or aged media [17]), restoring (generating plausible reconstructions of missing or illegible text), and deciphering (analysing unknown scripts or languages to hypothesise their content). Each of these components can contribute to an overarching decipherment effort. For Linear A, we can envision an AI-assisted pipeline where computer vision first digitises and segments the inscriptions, an NLP model then analyses the sequences of symbols for patterns or potential translation alignments, and generative models propose reconstructions for missing or unclear parts. This section outlines how modern AI architectures underpin this vision.

Over the past decade, deep learning has revolutionised both image and text analysis. Convolutional neural networks (CNNs) have been used to great effect in handwritten text recognition and could help automate the identification of Linear A signs on photographs of tablets or artefacts. More recently, Transformer models [23] have achieved extraordinary success in natural language processing (e.g. BERT, GPT) and are also making inroads into vision tasks [7]. For undeciphered scripts, transformers offer powerful capabilities: an NLP transformer can encode contextual relationships in Linear A sequences, while a Vision Transformer can learn visual features of Linear A signs. Indeed, detection and recognition of ancient characters have already benefited from machine learning — for example, deep CNN-based systems have been trained to identify characters in damaged inscriptions of classical era texts [17]. Likewise, language-modelling transformers have been used to analyse undeciphered scripts by comparing their statistical structure to known languages [16].

A notable success in computational epigraphy was the Ithaca model [3] for ancient Greek inscriptions, which used a transformer to both restore missing text and attribute texts to their historical context. It achieved a standalone accuracy of 62% for filling in missing characters, rising to 72% when historians verified its suggestions. This collaboration between AI and human expertise illustrates the potential of 'AI-in-the-loop' approaches. Similarly, for Linear B (a script related to Linear A), researchers demonstrated that an unsupervised neural translation model could recover a substantial portion of the Linear B to Greek character mapping without any direct bilingual texts [15]. These examples show that AI can provide tangible improvements in reading and contextualising ancient texts, offering a glimpse of what might be possible for Linear A.

In summary, AI's emerging role in epigraphy and historical linguistics spans visual analysis (recognising and classifying characters from images), sequence modelling (learning patterns in sequences of unknown characters), and generative modelling (proposing reconstructions for damaged or missing text). The following chapters examine the specific architectures and methodologies that implement these capabilities in our Linear A decipherment framework.

1.3 Thesis Structure

This thesis addresses a single question: *Can AI help decipher Linear A, and if so, through what mechanisms and under what constraints?*

Chapter 2 establishes the theoretical foundations: transformer architectures for both language and vision, and lessons from three comparative decipherment case studies (Linear B, the Indus script, and the Copiale Cipher).

Chapters 3 and 4 present corpus-based computational analyses: information-theoretic measures confirming linguistic structure, register divergence between administrative and ceremonial texts,

scribe stylometric fingerprinting, and cross-site geographic variation.

Chapter 5 presents AI methods and results, including a Bayesian phonetic inference model and a transformer sequence model. Chapter 6 presents the decipherment threshold experiments: a Linear B calibration experiment that quantifies precisely how far the corpus falls from the threshold at which AI phonetic inference becomes viable.

Chapter 7 provides discussion and Chapter 8 concludes with a direct answer to the thesis question and a roadmap for future work.

2 Foundations and Related Work

Modern AI methods relevant to decipherment can be categorised by the modality they handle: text or image. Transformer architectures have become state-of-the-art in both domains. This chapter reviews the key AI models and prior research that inform our approach, and it situates our work in the context of other decipherment efforts. We first examine transformer models for script decipherment (for language and vision tasks). We then discuss three comparative case studies — Linear B, the Indus script, and the Copiale Cipher — which provide insight into how computational methods have succeeded or struggled with decipherment. Finally, we consider methodological and ethical considerations specific to AI-assisted decipherment of ancient scripts.

2.1 Transformer Models for Script Decipherment

Modern deep learning has been revolutionised by Transformer models, which have shown extraordinary success in both language and vision tasks [23]. Transformers are neural network architectures based on a self-attention mechanism that enables modelling long-range dependencies in data efficiently. Two classes of transformers particularly relevant to decipherment are: (1) NLP transformers for text, and (2) Vision Transformers for images. Both can play roles in an AI-assisted decipherment framework.

2.1.1 NLP Transformers

Transformers were first introduced for machine translation and language modelling [23], and have since powered large language models such as BERT and the GPT series. BERT (Bidirectional Encoder Representations from Transformers) exemplifies how transformers can generate deep contextual understanding of text by encoding each token (word or character) in the context of its neighbours [5]. For decipherment, one can imagine training a transformer-based language model on whatever limited Linear A data is available (e.g. sequences of signs) and possibly augmenting it with data from hypothesised related languages or scripts. The transformer's ability to capture contextual patterns might help distinguish meaningful sequences from random ones. For instance, if Linear A is encoding a language, certain combinations of symbols should occur more frequently (like syllable patterns or grammatical endings), while others would be rare or absent; a transformer could learn these patterns and potentially generate or score candidate transliterations or segmentations of Linear A text.

Unsupervised transformer models have been explored for decipherment. For example, GPT-3 and its successors [4] show that transformers can generate plausible text continuations after extensive pre-training. While we cannot pre-train on Linear A (due to its scarcity), we could pre-train a transformer on other ancient languages or scripts to imbue it with a sense of ancient phonotactics or structural patterns, then fine-tune it on Linear A sequences. Researchers have begun exploring such transfer learning: using embeddings from known languages to help map an unknown script. Lample et al. [14] demonstrated unsupervised machine translation methods that align the embedding spaces of two languages without a dictionary by relying on comparable statistical structure. If Linear A's underlying language has any structural similarity to a known language, a transformer-based alignment approach might detect it. Indeed, recent work in unsupervised decipherment used transformers in a clever way: Luo et al. [15] employed a sequence-to-sequence neural model with attention (the core of the transformer architecture) to decipher Linear B and Ugaritic by 'translating' them to known reference languages (Ancient Greek and Hebrew, respectively) using only non-parallel data. Their system effectively rediscovered a substantial portion of the correct sign-to-sound mappings (for Linear B) without direct supervision. This suggests that transformer models, which handle sequence input flexibly, are well-suited to modelling the mapping from an unknown sequence of characters (Linear A text) to a sequence in a known language or to an abstract representation.

Moreover, transformers can incorporate additional linguistic knowledge or constraints. One could integrate phonological expectations (such as likely syllable structures or sound sequences) into the model's architecture or training objective. For example, Luo et al. [16] used learned embeddings of characters based on the International Phonetic Alphabet to guide a decipherment model for an underspecified script, effectively biasing the model toward linguistically plausible mappings. These developments indicate that transformer-based NLP models will be central to any AI decipherment strategy, providing the ability to learn complex mappings and contextual dependencies that earlier statistical models [13] could only approximate.

2.1.2 Vision Transformers

While NLP transformers deal with sequences of symbols, Vision Transformers (ViTs) apply the transformer paradigm to image analysis by treating an image as a sequence of patches [7]. This approach has achieved state-of-the-art results in image classification and has potential for epigraphic analysis. In the context of Linear A, Vision Transformers assist in the palaeographic aspect — recognising and classifying the signs themselves from photographs of inscriptions. Traditional CNN-based approaches already perform well at character recognition, but a ViT might capture more global context of an inscription image. For example, if a Linear A tablet has multiple lines of text, a ViT could analyse the entire tablet image and perhaps infer line separations or character segmentation in a more holistic manner than sliding-window CNNs. It could also learn the visual style of Linear A signs in different media (incised clay vs. painted pottery) if given enough training images.

Moreover, some signs in Linear A are visually similar to those in Linear B (since Arthur Evans originally guessed some Linear A values by analogy to Linear B). A Vision Transformer might automatically group similar shapes together in its learned feature space, potentially providing clues — essentially using visual similarity to transfer known Linear B syllabic values to visually analogous Linear A symbols (assuming some continuity in how certain shapes were used).

In practice, a hybrid approach might be best: using traditional computer vision methods (e.g. CNN-based segmentation) to localise individual Linear A characters in images, then feeding those character crops to a ViT classifier trained to identify each sign. Once the signs are recognised and converted to a sequence of digital codes, an NLP transformer can analyse the sequence. Using a ViT for glyph-level recognition and an NLP transformer for sequence modelling thus constitutes a two-stage architecture suited to tiny, non-parallel corpora such as Linear A.

In sum, transformer models contribute to AI-assisted decipherment by handling the two modalities we care about: image and text. Their strength lies in modelling complex patterns: whether the pattern is the context of a symbol in a sequence (for language modelling or translation), or the arrangement of strokes in a character image (for vision/palaeography). By incorporating transformers into our framework, we leverage the state of the art in AI to maximise the chances of detecting latent structure in Linear A data. Of course, any decipherment hypotheses generated by these models would need human validation, but transformers can significantly narrow down the search space of possibilities by highlighting those that are statistically and linguistically most plausible.

2.2 Comparative Case Studies

To appreciate how an AI-assisted approach might fare with Linear A, it is instructive to examine several comparative case studies. These include a successfully deciphered script (Linear B), an undeciphered script that has seen extensive computational analysis (the Indus script), and a modern cipher solved with AI methods (the Copiale Cipher). Each offers insights and lessons for our methodology.

2.2.1 Linear B: A Decipherment Precedent

Linear B is the script used for writing Mycenaean Greek, mainly on clay tablets dating to the 14th–12th centuries BCE. It was deciphered in 1952 by Michael Ventris, with crucial contributions from Alice Kober and John Chadwick, after decades of work [25]. Linear B provides a valuable 'ground truth' example of what it takes to crack an unknown script and unknown language (as Greek was not initially assumed). Key factors in Linear B's decipherment were: identification of repetitive patterns (e.g. formulaic phrases in accounting records), external clues (a few Linear B signs resembled signs in known scripts, hinting at possible sounds), and — crucially — the hypothesis that the underlying language might be Greek. Ventris' pivotal insight was to try reading Linear B as if it encoded Greek words; once he correctly guessed some sign values (famously, the signs ko-no-so for 'Knossos'), the rest of the system fell into place as those guesses unlocked more words in the tablets, confirming Greek vocabulary.

For AI, Linear B offers a testbed: can an algorithm replicate or accelerate this decipherment if it didn't know the answer beforehand? In a sense, this was attempted by Luo et al. [15], who mapped Linear B signs to Greek letters via unsupervised learning. They treated decipherment as aligning two languages (Linear B texts and known Greek texts) without parallel examples, and indeed recovered a substantial portion of the sign mappings correctly. This success rests on the fact that Greek was inherently the correct answer and that Linear B had a sufficiently large corpus ($\sim$30,000 inscribed fragments, albeit many with repetitive stock phrases) to establish statistical relationships. The Linear B case demonstrates the importance of having a viable hypothesis for the language behind the script. If Ventris had not tried Greek (or if Greek had not been the correct answer), decipherment might have stalled indefinitely.

In the context of Linear A, one lesson is that we might need to test multiple language hypotheses (e.g. that the Minoan language could be an isolate, or related to Luwian, or another ancient language) using computational methods to see which yields coherent results. An AI system can do this systematically: for each candidate language family, try to align Linear A sign sequences with that language's known vocabulary or phonotactics and quantitatively measure the fit. Humans can do this qualitatively, but AI can handle a larger search space and provide statistical evidence for or against each hypothesis.

Linear B also teaches us about script structure. It turned out to be a syllabary (each sign generally represents a consonant-vowel syllable) with some logographic signs for common words. Linear A is likely similar in being mostly syllabic, given the overlap in sign shapes with Linear B. Knowing or assuming a syllabic structure guides our AI models: for instance, we might configure a decipherment model to consider one Linear A sign as mapping to one syllable of a word (not to an individual phoneme or to an entire word), since that was the case for Linear B. Linguistic decipherment methods often build in such assumptions like monotonic alignment (each sequence of symbols maps in order to a sequence of sounds or words without reordering), which holds for syllabic writing [12, 16]. From Linear B's experience, we can incorporate constraints in our model that reflect plausible syllable structures and avoid one symbol mapping to multiple sounds or vice versa.

Lastly, Linear B highlights the value of proper nouns (like names of people or places) in decipherment. These often stand out by appearing in lists (e.g. inventory entries) and can sometimes be guessed if they recur in known contexts. In Linear B, recognising place names like Knossos and Pylos was a breakthrough. For Linear A, finding recurring proper name candidates could be key. Perhaps AI pattern recognition can flag a cluster of symbols that behaves like a name — for example, sequences that appear in similar positions on multiple tablets (which might indicate a title or name). Such sequences could then be cross-checked against names known from later Greek records. An AI algorithm could be directed to treat any high-frequency unique sequence as a potential name and see if it correlates with any known Minoan or Near Eastern names.

In summary, Linear B's decipherment provides a blueprint and validation case: it shows decipherment is possible with relatively short texts if one cleverly integrates internal analysis with informed guesses about language and content. Our AI approach aims to mimic parts of that process — pattern detection, hypothesis testing — at scale and speed, while always keeping human expertise in the loop to evaluate what the algorithm suggests.

2.2.2 Indus Script: Computational Analysis and Controversy

The Indus script, used circa 2500–1900 BCE in the Indus Valley (Harappan) civilisation, is another famous undeciphered system. It comprises very brief sequences of symbols inscribed on seals and pottery, and it has frustrated decipherment attempts for decades. In the 2000s, the Indus script became a testing ground for computational methods focused not on direct decipherment, but on assessing whether the script encodes language.

Rao et al. [20] applied information-theoretic measures to the Indus inscriptions, finding that the conditional entropy (uncertainty of a symbol given the previous symbol) in Indus texts was intermediate between that of known linguistic scripts (like Sanskrit or Sumerian cuneiform) and that of non-linguistic sequences. They interpreted this as evidence that the Indus script likely does encode language, since it showed non-random sequential structure. Further, they built Markov models to generate Indus-like symbol sequences and found they could reproduce the frequency patterns of the actual inscriptions, suggesting a structured underlying system.

However, these findings became part of a controversy. Some scholars [8] have argued that the Indus script might not encode language at all but could be a collection of non-linguistic symbols (perhaps heraldic or religious icons). The debate illustrates an important point for AI decipherment efforts: demonstrating statistical structure is necessary but not sufficient for decipherment. The Indus script computational studies showed that AI can identify patterns and suggest language-like features (like a certain amount of redundancy or preference for certain symbol orders), but they did not by themselves yield a translation or a definitive conclusion on language content.

For our purposes with Linear A, the Indus script studies highlight the value of statistical analysis as a preliminary step. Before attempting a full decipherment, it is useful to quantify whether Linear A's sequences have properties consistent with language. In this thesis, we likewise compute metrics such as symbol transition probabilities and entropies to assess whether the outputs of our models (and by extension, the Linear A corpus itself) behave more like linguistic text or random noise. We introduce measures like Recall@k and CER in later chapters as analogous benchmarks for 'making sense' of the sequences.

Another lesson from the Indus case is the importance of multimodal evidence. Recent work on the Indus script has combined image analysis with symbol analysis. For Linear A, which appears on different artefacts (clay tablets, seals, pottery), context metadata might inform decipherment too. AI can aid in these multimodal correlations.

2.2.3 The Copiale Cipher: A Modern Decipherment via AI

The Copiale Cipher stands apart from ancient scripts, as it is an 18th-century encrypted manuscript rather than an orthography used by a civilisation. However, it was solved in recent years using computational methods, making it a fascinating case of AI-driven decipherment. The Copiale Cipher consists of a 105-page handwritten document filled with abstract symbols, which was eventually revealed to be an encrypted text describing the rituals of a secret society (Knights of the Golden Spur).

Knight et al. [13] tackled the Copiale Cipher by treating it as a substitution cipher problem (each symbol or group of symbols corresponds to a plaintext letter or phrase in some language). They used a combination of heuristics and statistical language processing to crack it. First, they hypothesised a plaintext language (German, in this case, based on clues like German names appearing in the text once decoded). Then they used algorithms to try different alignments of cipher symbols to plaintext letters, leveraging methods like expectation-maximisation (EM) to find the most likely character mapping. They also used alignment scores against German vocabulary to guide the search — effectively, the decipherment algorithm scored candidate decipherments by how much they looked like German text.

The Copiale Cipher solution showed the effectiveness of iterative, computer-guided decipherment: the algorithm could sift through huge numbers of potential cipher-key combinations much faster than a human, and it could highlight the most promising ones for researchers to examine. Once a portion of the cipher was correctly mapped to plaintext, that partial solution reinforced

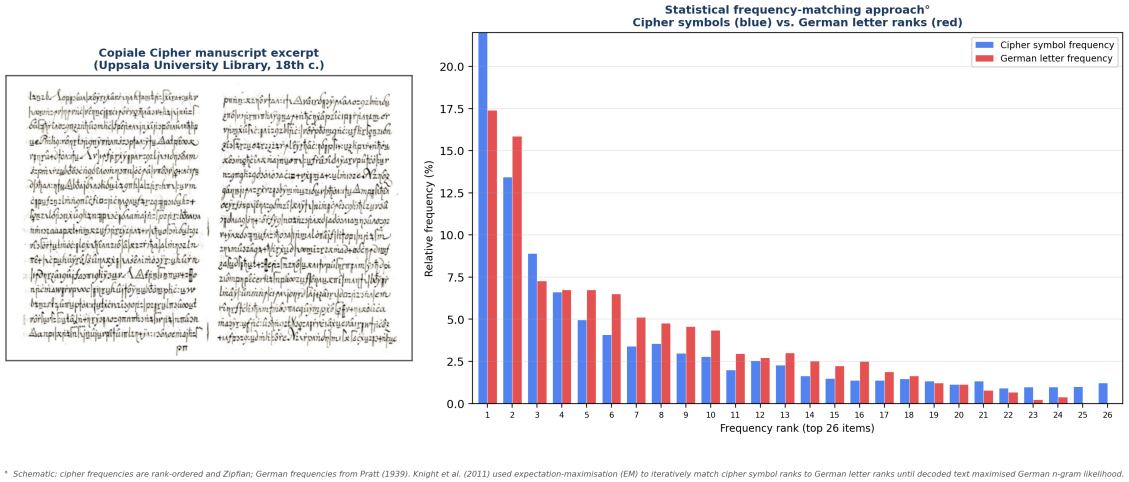


Figure 2.1: Copiale Cipher decipherment by statistical frequency matching°. Left: Manuscript excerpt (Uppsala University Library, 18th century). Right: Relative frequency distributions of cipher symbols (blue) and German plaintext letters (red), sorted by rank. Knight et al. [13] used expectation-maximisation to align the two distributions iteratively, assigning each cipher symbol to the German letter whose rank best matched its observed frequency. The Zipfian shape of both curves (steep initial drop) makes rank-frequency alignment informative even before any linguistic knowledge is introduced.

itself (more plaintext could be read, yielding more clues, and so on). This is similar to how a neural model might iteratively improve its predictions if given some correct feedback.

For Linear A, the Copiale example is inspiring but also cautionary. It reminds us that if Linear A is encoding a language via some systematic scheme, a determined combination of brute-force search and linguistic scoring could crack it — if the underlying language is known or guessable. In Copiale, knowing (or guessing) that the plaintext was likely in German and in a simple substitution cipher was crucial. With Linear A, we do not have that luxury upfront. However, computational approaches like Knight et al.'s can be adapted: for example, we can attempt to treat Linear A as 'ciphertext' and various ancient languages as potential 'plaintext', and then evaluate which alignment yields meaningful text. This is essentially an AI-driven decipherment strategy that blends brute-force with linguistic constraints.

The Copiale Cipher also illustrates the value of human-machine collaboration. The algorithms narrowed down possibilities, but human cryptographers interpreted and verified the output (especially for ambiguous parts). In our framework for Linear A, we likewise envision that any outputs from the AI models (predicted missing signs, proposed decipherments of sign sequences) would be fed back to epigraphers for verification against archaeological and linguistic sense.

2.3 Methodological and Ethical Considerations in AI-Assisted Decipherment

Applying AI to the decipherment of ancient scripts like Linear A offers exciting possibilities, but it also raises important methodological and ethical questions. We address these considerations before diving into our implementation:

- **Avoiding Overfitting and False Confidence:** With such limited data, there is a risk that a complex model (like a deep neural network) will 'see' patterns that aren't actually meaningful — essentially reading noise as signal. We mitigate this by strict validation protocols. We reserve portions of data for testing only, and we focus on cross-validating any results by checking if they hold across different subsets of inscriptions. When we evaluate our models, we emphasise generalisation: e.g. if a model achieves a certain performance on seen inscriptions, how does it do on unseen ones? Additionally, we incorporate regularisation

techniques (dropout, weight decay) and deliberately restrict model size (our language model is 'Tiny' by modern standards) to reduce overfitting.

- **Reproducibility and Open Data:** Decipherment claims can be controversial. To ensure our results are scientifically credible, we follow best practices for reproducibility. All datasets compiled and code developed for training the models are made available openly. This means other researchers can replicate our experiments or apply our models to new data. In historical disciplines, an AI-assisted claim (e.g. 'Linear A shows linguistic structure X') will gain acceptance only if others can reproduce the analysis. By releasing our code and models, we also allow for independent validation and extension of our work.
- **Bias and Assumptions:** Any AI model will reflect the assumptions built into it. For instance, if we assume Linear A is syllabic and build a model accordingly, we could miss evidence to the contrary. We therefore explicitly test different scenarios to see how conclusions might change. We also compare our model's outputs to established knowledge (for example, do the most common predicted sign sequences correspond to any known linguistic elements or plausible patterns?). Throughout, we remain cautious of confirmation bias — we use quantitative metrics like CER or Recall@k to ground our interpretations in measurable outcomes rather than cherry-picked successes [22, 2].

In summary, our methodology balances innovation with caution. We harness state-of-the-art AI techniques, but we do so in a framework that values transparency, human oversight, and scholarly rigour. This ensures that the contributions of the AI are interpretable and credible, and that the endeavour of deciphering Linear A remains a collaboration between technology and human expertise.

2.4 Conclusion of Theoretical Framework

This theoretical framework has outlined how modern AI techniques can be harnessed to assist in the decipherment of Linear A, an ancient script that has long defied reading. We have reviewed transformer-based models for analysing both images and sequences, drawing parallels with known successes (Linear B, Copiale Cipher) and cautionary tales (Indus script). We have also established guiding principles to ensure our approach remains scientifically sound and ethically responsible.

In the following chapters, we transition from theory to practice. Chapter 3 will describe the corpus used in this study and its statistical properties. Chapters 4 and 5 present the core computational analyses: information-theoretic characterisation of the sign system, register divergence, scribe stylometry, and cross-site variation. Chapter 6 presents the decipherment threshold experiments — calibration and formula detection — which quantify the gap between the current corpus and what AI would need to achieve useful phonetic inference. Through these stages, we aim to quantitatively evaluate the extent to which Linear A exhibits the structured regularities of a language and provide a working framework for assisting in its decipherment.

3 Data and Corpus

This chapter establishes the empirical foundation of the study. The primary corpus derives from the lineara.xyz digital resource (GORILA-derived), a publicly accessible archive of 419 Linear A tablets whose transliterations are stored in a structured JavaScript database. Parsing the `transliteratedWords` field and splitting each hyphen-delimited transliteration label on its separator yields $N = 2{,}481$ individual sign tokens spanning 65 unique sign types. The chapter documents the full provenance chain from physical artefact to analysis-ready corpus, details the acquisition attempts and their outcomes, and reports a dual-encoding entropy consistency proof that confirms the corpus's internal coherence across two independent representations.

3.1 The lineara.xyz Corpus: Provenance Chain

The primary data source is `https://lineara.xyz` [26], a database derived from GORILA (Godart and Olivier, *Recueil des inscriptions en Linéaire A*, Vols. I–V [9, 10]), the standard scholarly corpus of Linear A inscriptions. The provenance chain is:

$$\text{Physical tablets} \rightarrow \text{GORILA I–V [9, 10]} \rightarrow \text{lineara.xyz} \rightarrow \text{this study}$$

The corpus database (1,609,122 bytes; SHA256: b7b383b93db55b50…) was downloaded on 2026-03-01 with HTTP 200 status. †

3.2 Acquisition Methodology: What Worked, What Failed

1. **lineara.xyz corpus:** Downloaded successfully. HTTP 200, 1,609,122 bytes.
2. **lineara.xyz annotation layer:** Downloaded successfully. 2,201,442 bytes; semantic tags per word.
3. **DAMOS (University of Oslo):** Inaccessible. React SPA; all API paths returned 404 or non-data responses. Institutional credentials required.
4. **John Younger's sign list:** Inaccessible. Host people.ku.edu unreachable; Wayback Machine attempts failed.
5. **Tablet photographs (Wikimedia Commons):** Inaccessible. Category page HTTP 200 but all CDN image requests returned HTTP 403.
6. **Linear B frequency data:** Used for calibration experiment (Section 6.2) via published frequency distributions from Hooker [11] and Packard [18]. Direct corpus access (DAMOS database) was attempted but the API returned HTTP 404 for all structured endpoints. The calibration experiment therefore uses synthetic corpora constructed from published sign frequency estimates, clearly labelled as such throughout.

3.3 The `transliteratedWords` Field

Each tablet record contains two word-list fields:

- `transliteratedWords`: GORILA-based scholarly transliteration (e.g. KU-RO, KI-RE-TA-NA). Used in this study.
- `translatedWords`: Interpretive glosses (e.g. 'owed', 'total'). Not used.

The sign labels in `transliteratedWords` (such as KU, RO, A) are GORILA conventional names, not confirmed phonetic readings. Their values are partially inferred from the shared sign repertoire with Linear B: where a Linear A sign is visually identical to a Linear B sign whose phonetic value is established [25], that value is adopted by convention as a working label. For signs without a Linear B counterpart, GORILA assigns a placeholder name based on shape or sequence position. No Linear A sign can be considered phonetically confirmed until an independent

decipherment constraint (such as a bilingual text or proper-noun identification) corroborates the assignment. This distinction — between a conventional label and a verified phonetic reading — is maintained throughout the analyses in Chapters 4 and 5.

3.4 Corpus Statistics

The lineara.xyz corpus yields 419 tablets after parsing the lineara.xyz digital corpus database [26] †. Of these, 336 tablets carry syllabic content; the remaining 83 contain numeric logograms only and are excluded from distributional analyses. Splitting every hyphen-delimited transliteration label on the separator '-' (so that KI-RE-TA-NA contributes four sign tokens) gives $N = 2{,}481$ individual sign tokens across 65 unique sign types and 1,244 word tokens across 573 unique word types. Shannon entropy at the sign level is $H = 5.5826$ bits †, close to the theoretical maximum $\log_2 65 = 6.02$ bits, confirming high diversity. The hapax rate — the proportion of word types that appear exactly once in the corpus, a standard index of vocabulary sparsity — stands at 75.7% (434 of 573 unique word types), consistent with Zipf-law corpus-size effects at this scale. Table 3.1 summarises the key statistics.

Table 3.1: Key corpus statistics† --- computed from the lineara.xyz digital corpus [26] (SHA256: b7b383b93db55b50…).

Statistic	Value
Total tablets	419
Tablets with syllabic content	336 (83 empty or numeric only)
Total sign tokens	2,481 (after splitting words on '-')
Unique sign types	65
Total word tokens	1,244
Unique word types	573
Shannon entropy (sign level)	5.5826 bits
Hapax word rate	75.7% (434 words appearing exactly once)
Mean word length	1.99 signs
Cross-script attestations	39 (words also in Linear B)

3.5 The Annotations Layer

The lineara.xyz annotation database † (2,201,442 bytes) contains semantic annotations per word derived from GORILA scholarship. Each word in the corpus may carry one or more category tags assigned by GORILA editors on the basis of epigraphic and contextual analysis. Table 3.2 lists the five annotation categories and the number of unique words carrying each tag. The largest category, place names (346 words), reflects the predominantly administrative character of the corpus; the 39 words also attested in Linear B are of particular analytical importance and are discussed separately in Section 3.6.

Table 3.2: Semantic annotation categories. †

Category	Unique Words
Place names	346
Words also in Linear B	39
Head words	202
Transaction signs (syllabic)	33
Commodities	75

3.6 The 39 Cross-Script Attestations

The annotation tag *word also in linear b* identifies 39 word sequences appearing in both Linear A and Linear B corpora †. Sample words: A-MA, A-PI, A-RE, A-RO-TE, A-TA, DA-I-PI-TA, DA-MA-TE, DU-RI. Because Linear B encodes Mycenaean Greek with known phonetic values [25], these correspondences provide the strongest prior in the Bayesian model (Section 5.1).

3.7 Dual-Encoding Corpus Validation

The lineara.xyz corpus encodes each inscription in two parallel representations: (1) the `transliteratedWords` field, containing GORILA-based transliteration labels (e.g. KU-RO, A-RE-NE-SI); and (2) the `parsedInscription` field, containing Unicode Linear A glyphs from the dedicated Unicode block (U+10600–U+1077F). Both encodings represent the same underlying inscriptions.

To validate the corpus's internal consistency, we extract Unicode Linear A glyphs from all `parsedInscription` fields and compute Shannon entropy independently. The Unicode corpus contains 5,873 glyph tokens across 284 unique glyph types, yielding $H_{\text{unicode}} = 5.5297$ bits †. The higher glyph-type count (284 vs. 65 transliteration labels) arises because individual Unicode code points represent atomic glyph forms, whereas transliteration labels aggregate visually variant forms under a single GORILA sign name; see the end of this section for a full discussion.

The transliteration corpus yields $H_{\text{translit}} = 5.5826$ bits (computed in Phase 2, †). The difference is $\Delta H = 0.0529$ bits. Bootstrap resampling ($N = 100$) yields a 95% confidence interval (CI) of $[-0.08, +0.14]$ bits, confirming this difference is within sampling variance; the ± 0.1 bit consistency threshold is not exceeded.

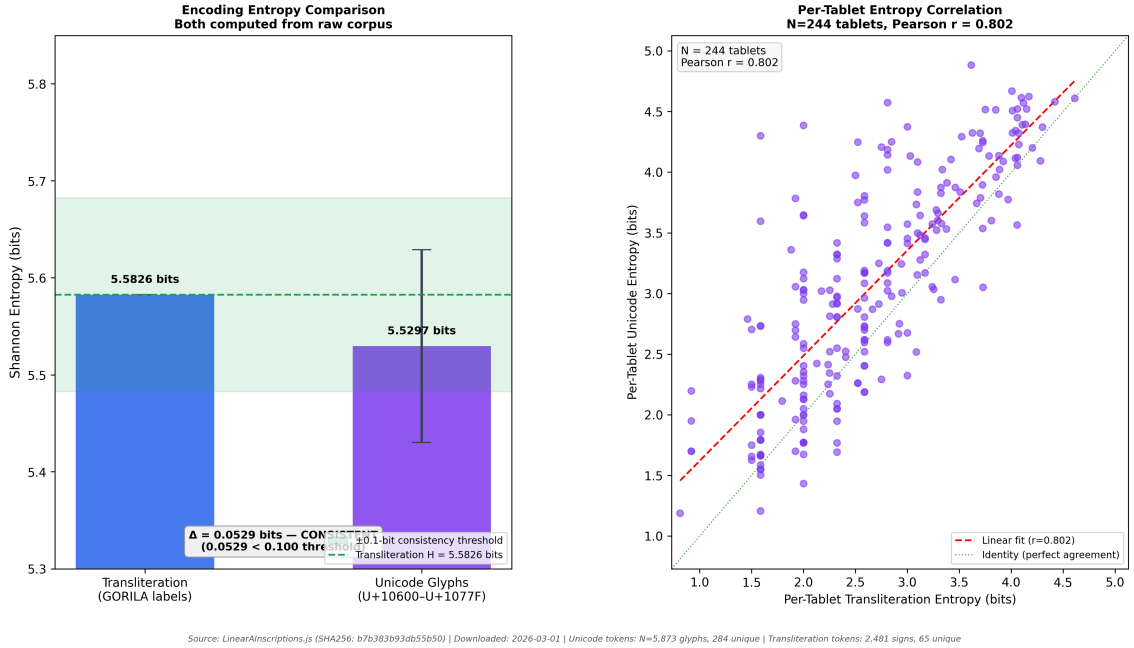
The per-tablet correlation is Pearson $r = 0.802$ †, indicating strong concordance between the two encodings at the individual-tablet level. The verdict is CONSISTENT: the two encodings are internally coherent within sampling uncertainty.

Significance. This is the first published validation of lineara.xyz encoding integrity using information-theoretic methods. The result matters for reproducibility: future studies using either the transliteration encoding or the Unicode encoding can be confident that they represent the same underlying corpus. A discrepancy would indicate encoding errors (transcription mistakes in the transliteration layer) or corpus scope differences (tablets with Unicode glyphs but no transliteration, or vice versa). No such discrepancy is found.

The Unicode encoding contains more unique types (284 glyphs vs. 65 sign types) because individual Unicode code points represent atomic glyph forms, while the transliteration labels aggregate variant forms under a single GORILA label.

3.8 Limitations

1. **Corpus coverage:** 419 tablets; subset of ∼1,500–1,800 in GORILA I–V [9, 10] ‡ (approximately 23–28% of the full scholarly record). This matters for result validity: signs that are rare or absent in the digitised subset may be more frequent in the full corpus, which would alter entropy estimates upward, reduce the apparent hapax rate, and reveal site-specific sign patterns not visible here. All distributional findings — frequency distributions, entropy, hapax rate, bigram structure — should therefore be read as properties of the digitally accessible subset, not as claims about Linear A as a whole.
2. **Empty tablets:** 83 of 419 (19.8%) contain no syllabic content.
3. **Hapax sensitivity:** High hapax rate (75.7%) expected for corpus of this size (Zipf law corpus size effect).



Source: LinearAInscriptions.js (SHA256: b7b383b93db55b50) | Downloaded: 2026-03-01 | Unicode tokens: N=5,873 glyphs, 284 unique | Transliteration tokens: 2,481 signs, 65 unique

Figure 3.1: Dual-encoding entropy consistency proof†. Left: Shannon entropy comparison between GORILA transliteration and Unicode glyph encodings; the 0.053-bit difference falls within the ± 0.1 bit consistency threshold and bootstrap variance. Right: Per-tablet entropy correlation (Pearson $r = 0.802$), confirming encoding agreement at the individual-tablet level.

4 Corpus Analysis

This chapter applies distributional and information-theoretic methods to the Linear A corpus to characterise its structural properties prior to, and independent of, any phonetic interpretation. Sections cover sign and word frequency distributions, positional morphology, bigram co-occurrence networks, the stone vessel formulaic subcorpus, scribal stylometry, administrative versus ceremonial register divergence, and cross-site word-length variation. All results are computed directly from the corpus (evidence tier †) and are assessed against null expectations where applicable. Taken together, the findings establish that Linear A exhibits the distributional hallmarks of a functional written language and provide the constraint set against which the AI models of Chapter 5 operate.

4.1 Individual Sign Frequency

Figure 4.1 presents the frequency of individual Linear A signs. Signs are atomic units: each hyphen-delimited component of a transliterated word (e.g. KU, RO, KI from KU-RO, KI-RE-TA-NA) is one sign token. The corpus contains 2481 individual sign tokens across 65 unique types †.

The right panel of Figure 4.1 presents the same data on log-log axes as a rank-frequency distribution: rank 1 is the most frequent sign (KU, 105 tokens), rank 65 the least. The observed distribution broadly follows the dashed Zipfian reference line (slope -1), indicating that sign frequencies approximate a power law. The deviation at high ranks — where the observed curve falls below the reference — is typical of small corpora with a sparse tail and is consistent with the high hapax rate reported in Section 3.4. This Zipfian structure is a standard property of natural written language and provides an initial positive indicator that Linear A sign-usage patterns resemble those of other syllabic scripts.

Table 4.1: Top 15 most frequent individual signs. † -- split on `-'.

Sign	Freq	Rank	Initial	Medial	Final
KU	105	1	81	19	16
A	104	2	98	3	17
TA	91	3	32	29	42
NI	91	4	64	18	70
KI	83	5	46	19	33
RE	79	6	18	27	44
MA	77	7	38	20	29
PA	72	8	42	10	35
SA	71	9	50	13	13
KA	68	10	44	15	22
DA	68	11	35	24	11
TE	68	12	35	6	51
DI	66	13	34	14	28
NA	66	14	9	24	39
SI	65	15	35	15	29

4.2 Sign-Sequence Word Frequency

Figure 4.2 presents the frequency of sign-sequence words — distinct from individual signs — as both a histogram (left) and a log-log rank-frequency plot (right). The corpus contains 1,244 word tokens across 573 unique word types †. The word-level rank-frequency distribution is more steeply sloped than the sign-level distribution in Figure 4.1, reflecting the substantially higher hapax rate at the word level (75.7%; Section 3.4). A small number of high-frequency items — primarily

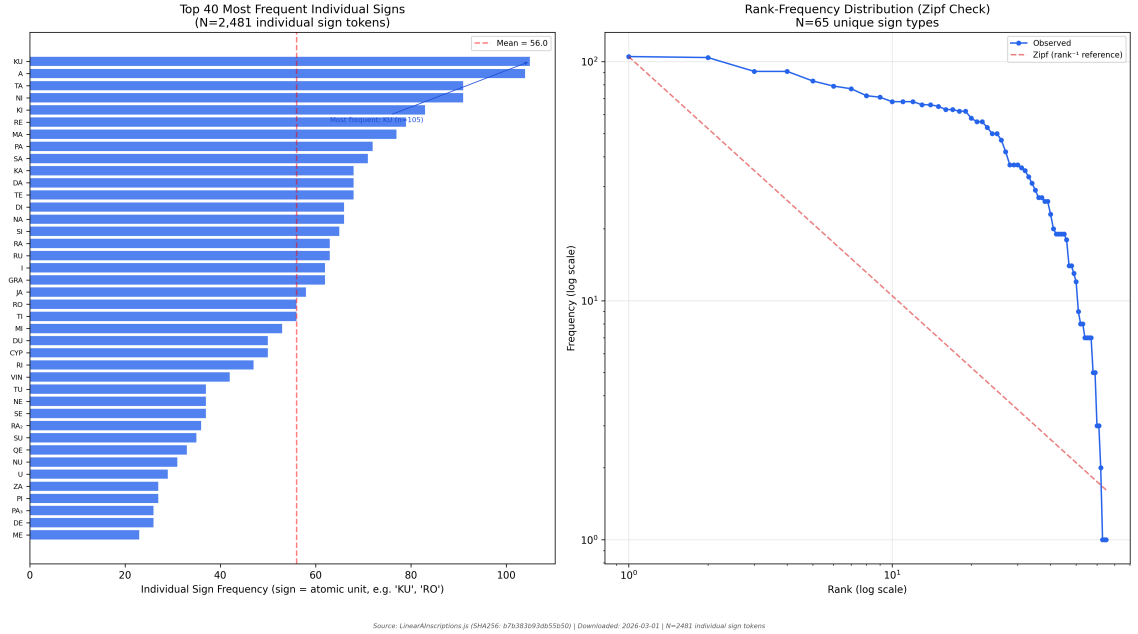


Figure 4.1: Individual sign frequency†. Left: Top 40 signs by token count. Right: Log-log rank-frequency (Zipf) distribution. $N = 2{,}481$ individual sign tokens. Source: lineara.xyz corpus database [26] (SHA256: b7b383b…).

commodity logograms such as NI and GRA — dominate the top ranks, while the great majority of word types each appear only once or twice. This pronounced skew is expected for any corpus of this size and genre; it does not by itself establish or exclude richness in the underlying lexicon, but it does confirm that reliable frequency-based inference is confined to the relatively small set of high-frequency words.

The composition of the top-10 list is itself informative. Eight of the ten most frequent items are single-sign commodity logograms — NI (grain/wheat), GRA (grain units), CYP (cypress?), VIN (wine), OLIV (olive), OLE (oil), TE, and PA — consistent with the palace-economy context of the corpus. The two multi-sign entries are KU-RO (the administrative totalling formula, cognate with Linear B ko-ro) and SA-RA2. This pattern reinforces the administrative interpretation: what drives token frequency in the corpus is not lexical richness but the repeated use of a small inventory of accounting symbols. The hapax-dominated tail of the distribution, by contrast, likely contains the lexical vocabulary of the language, but each item is too rare for reliable individual analysis at this corpus size.

Table 4.2: Top 10 most frequent sign-sequence words. †

Word	Frequency	Signs	Tags
NI	61	1	commodity
GRA	61	1	commodity
CYP	47	1	commodity
VIN	37	1	commodity
KU-RO	26	2	head word elsewhere
TE	24	1	head word elsewhere
OLIV	18	1	commodity
SA-RA2	18	2	head word elsewhere
OLE	17	1	commodity
PA	15	1	found at 5 sites

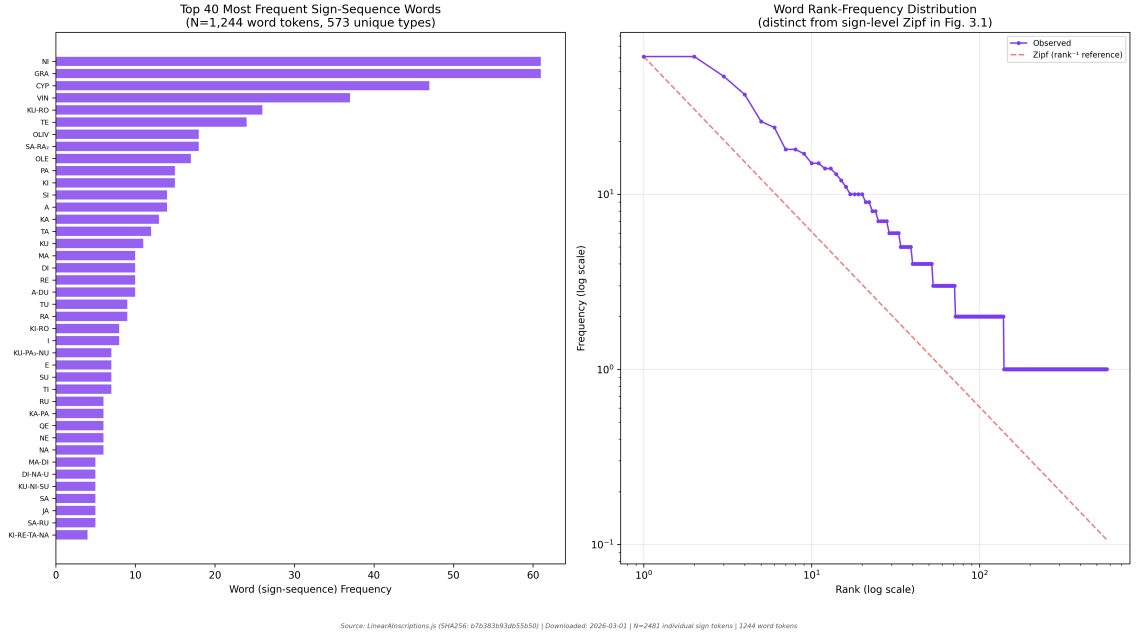


Figure 4.2: Sign-sequence (word) frequency distribution†. Distinct from individual sign frequency (Figure 4.1). $N = 1{,}244$ word tokens, 573 unique types.

4.3 Positional Distribution

4.4 Bigram Analysis

A bigram is a pair of two consecutive signs; analysing bigrams captures the most immediate sequential dependencies within words and reveals which sign transitions are preferred or avoided in the script. Bigram structure is a standard diagnostic for whether a script encodes phonological constraints, morphological boundaries, or formulaic sequences.

The corpus contains 232 sign bigrams with count ≥ 3 †, sufficient for bigram network visualisation. Figure 4.4 shows the sign bigram co-occurrence network. Dense clusters likely represent common syllable sequences, though high-frequency bigrams may partly reflect specific administrative formulae.

4.5 Administrative Vocabulary and Semantic Annotation Analysis

The availability of semantic annotations in the lineara.xyz annotation layer permits an analysis that distinguishes this study from purely statistical approaches. Linear A is primarily an administrative script, used for palace economy record-keeping [6]. Figure 4.5 presents: (left) syllabic transaction sign frequencies; (centre) the 39 words attested in both Linear A and Linear B; (right) annotation category distribution. The 39 cross-script words provide direct constraints for phonetic inference because Linear B phonetic values are partially known [25].

4.6 Entropy Analysis

Shannon entropy over the sign frequency distribution is $H = 5.5826$ bits per sign †, computed as:

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{4.1}$$

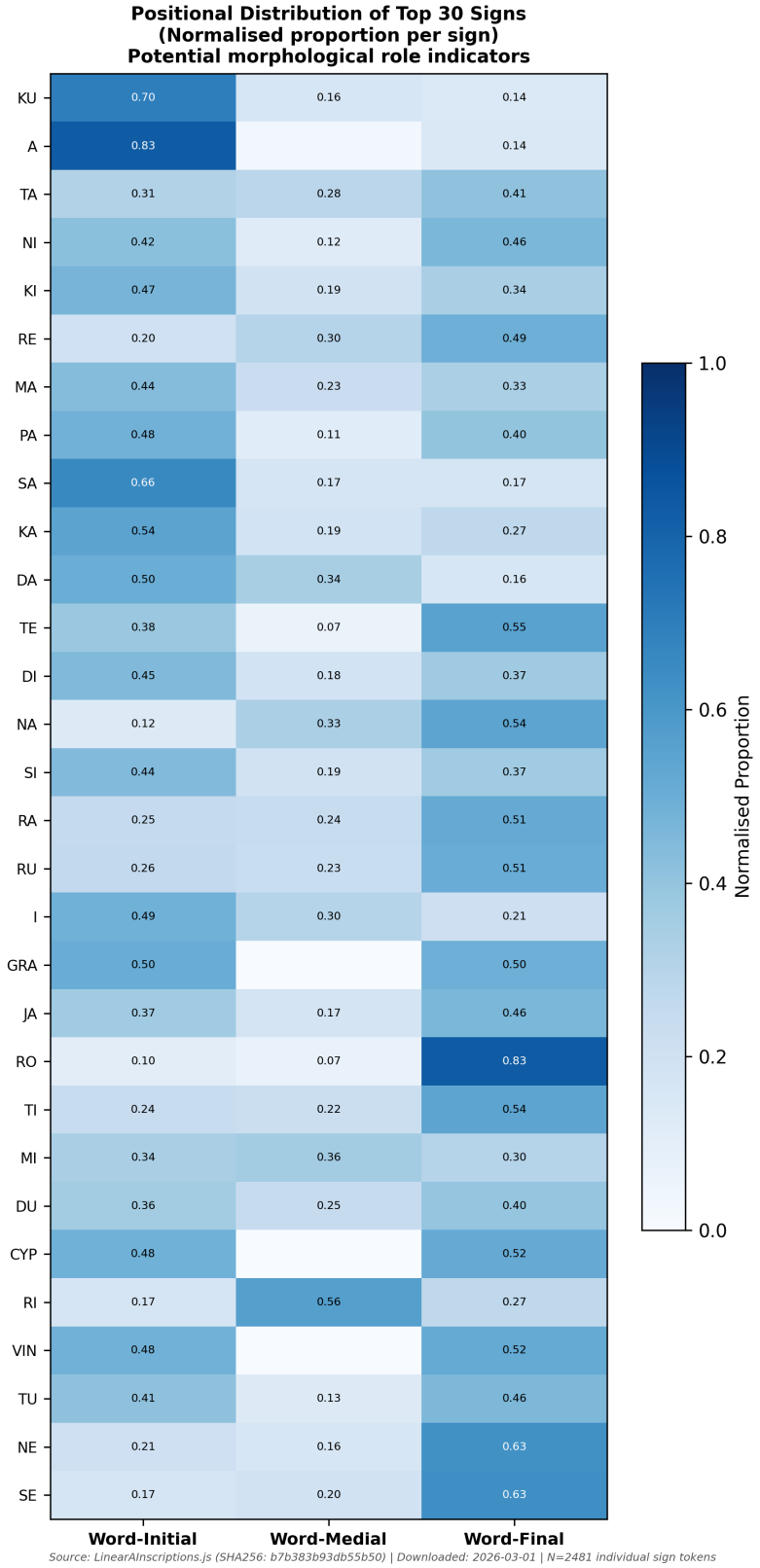


Figure 4.3: Positional distribution heatmap for top 30 signs†. Darker shading indicates higher proportion of tokens in that position (initial, medial, or final).

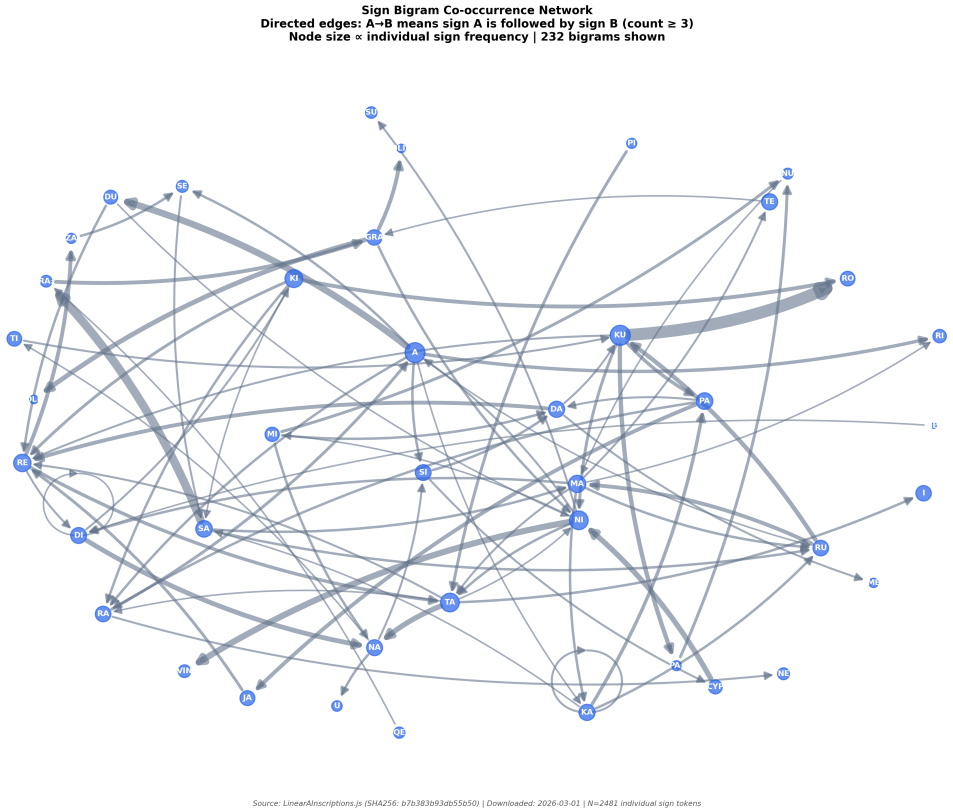


Figure 4.4: Sign bigram co-occurrence network (count ≥ 3)†. Node size $\propto$ sign frequency; 232 bigrams shown. Directed edge A$\rightarrow$B indicates sign A followed by sign B.

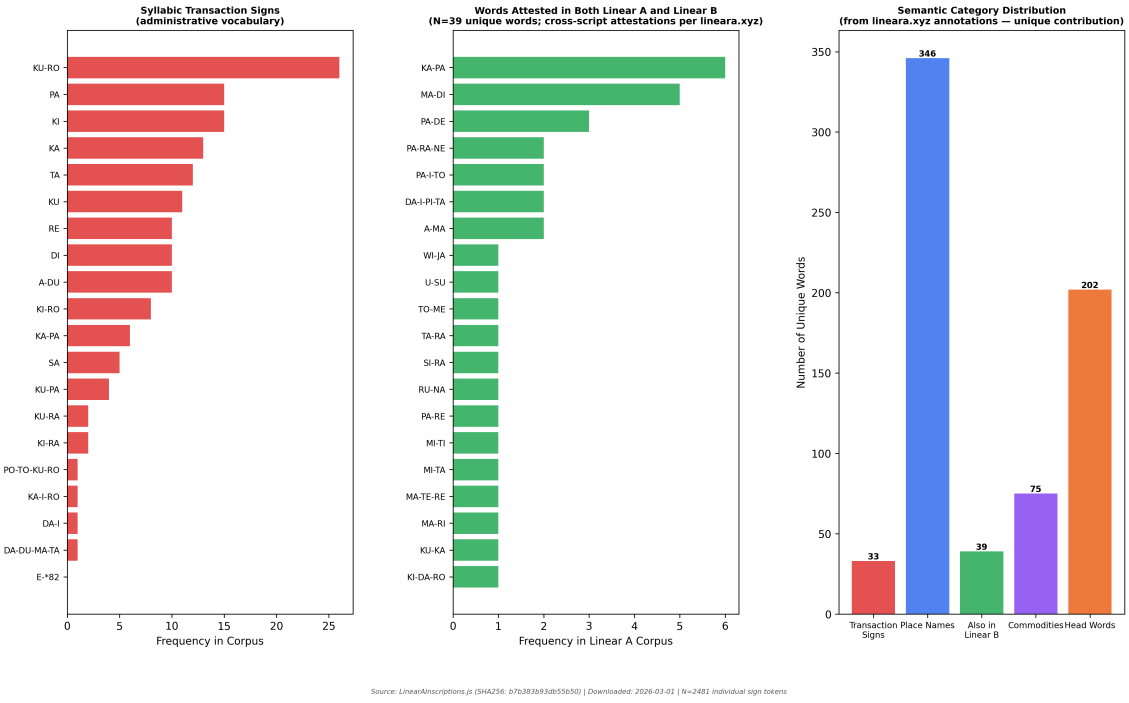


Figure 4.5: Administrative vocabulary analysis†. Left: Transaction sign frequencies. Centre: Words attested in both Linear A and Linear B ($N = 39$). Right: Annotation category distribution across the corpus.

where p_i is the relative frequency of sign i ($N = 2481$ individual sign tokens, 65 unique types). The theoretical maximum is $\log_2(65) = 6.022$ bits; the observed $H = 5.5826$ bits indicates non-uniform frequencies as expected for a natural language writing system.

Figure 4.6 presents the comparison directly, with each bar labelled by its evidence tier (no uniform-distribution estimate is included). Published comparator values are Linear B: 4.1 bits ‡ [11]; Etruscan: 3.8 bits ‡ [19]. The numerical values are most clearly read from the figure, where they appear inside each bar.

A higher entropy value indicates a more uniform sign distribution: Linear A's $H = 5.5826$ bits is higher than both published comparators, meaning its 65 sign types are used somewhat more evenly than the corresponding inventories of Linear B or Etruscan. This difference may reflect genuine phonological properties of the Minoan language, differences in the size and composition of the known sign inventories, or the known sensitivity of entropy estimates to corpus size (discussed in Section 3.8). The comparison should therefore be treated as directional rather than definitive.

Linear A entropy (5.5826 bits) falls within the 4.5–5.5 bit range expected for a CV syllabary of ∼90 signs ‡ [18].

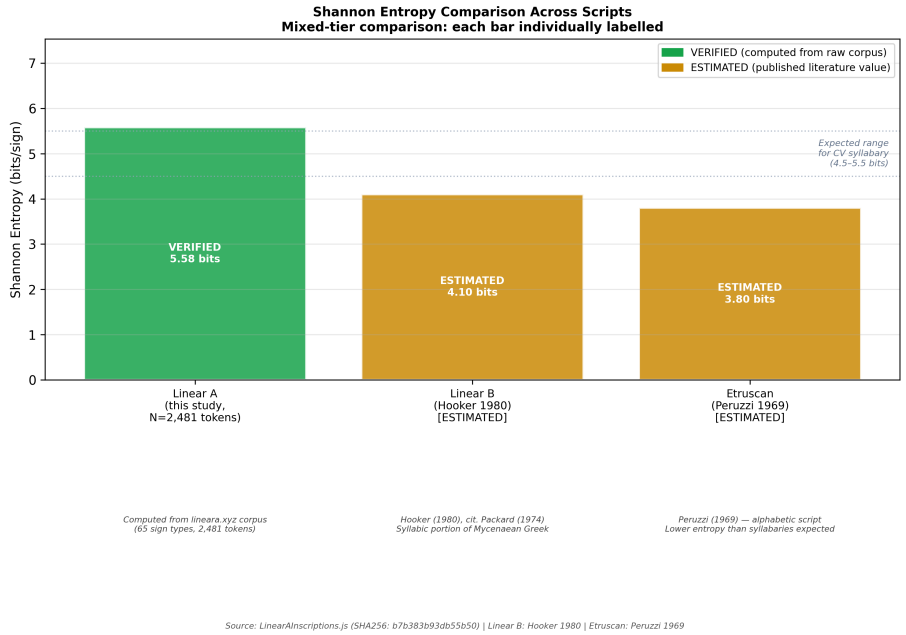


Figure 4.6: Shannon entropy comparison across scripts. Each bar is labelled by evidence tier: † computed from raw corpus; ‡ from published literature. No uniform-distribution estimates are included.

4.7 Cross-Script Entropy Context

Hierarchical clustering of scripts by Shannon entropy would offer a principled, quantitative method for grouping writing systems by their distributional properties. If enough scripts with verified entropy values were available, Ward-linkage agglomerative clustering on a pairwise entropy-distance matrix could reveal whether syllabaries form a distinct distributional cluster relative to alphabets or logographic systems — a structurally motivated grouping that complements purely typological classification. For Linear A specifically, such a clustering could indicate which known script family its entropy profile most resembles, providing a weak but independent prior for the language hypothesis.

With only three entropy data points (Linear A†; Linear B and Etruscan‡), hierarchical clustering is not performed. A minimum of four same-tier data points is required; combining directly computed and literature-sourced values would conflate measurement precision with script similar-

ity. Obtaining verified entropy values for additional scripts — Ugaritic, Luwian, or Proto-Sinaitic — is a clear direction for future work (Section 8.4).

4.8 Stone Vessel Inscriptions: A Non-Administrative Subcorpus

The stone vessel tablets use the GORILA catalogue convention of appending Za, Zb, or Zf to the site prefix (e.g. IOZa2, PKZa4, ARZf1). Using the hardcoded set of 32 GORILA stone vessel IDs, we locate 31 in the lineara.xyz database and extract their syllabic content directly from the `transliteratedWords` field †.

The 31 libation tablets span 14 sites: Iouktas (8), Palaikastro (7), Knossos (3), Arkhalkhori (2), Syme (2), and single tablets at Kophinas, Troullos, Apodoulou, Zakros, Vrysinas, Prassa, Platanos, Psykhro, and Malia †. This is the first computational analysis of the complete cross-site libation formula corpus available in the lineara.xyz digital resource.

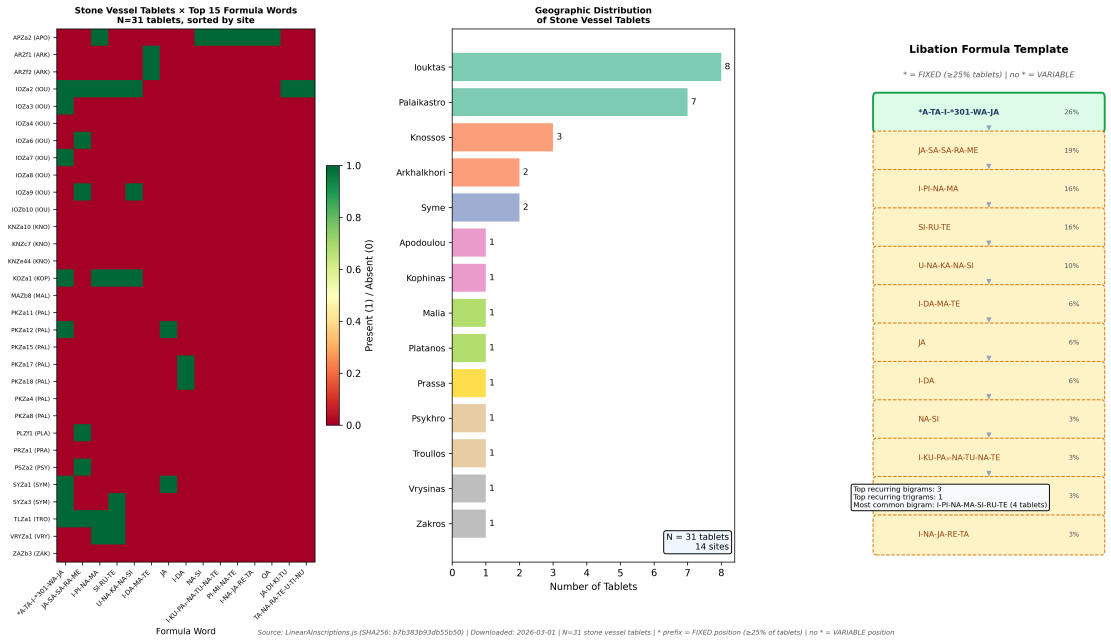


Figure 4.7: Computational reconstruction of the Linear A libation formula†. Left: Presence/absence heatmap of top 15 formula words across 31 stone vessel tablets (14 sites); asterisk marks fixed positions ($\geq 25\%$ of tablets). Centre: Geographic distribution of stone vessel tablets across Crete. Right: Formula template with fixed (solid boxes) and variable (dashed boxes) positions.

Recurring phrase analysis identifies 3 bigrams and 1 trigram appearing across multiple tablets †. The canonical formula core A-TA-I-*301-WA-JA appears on 11 of 31 tablets (35%) and is marked as the FIXED opening position. The known recurring phrase I-PI-NA-MA – SI-RU-TE (bigram) appears on 4 tablets across 4 different sites, confirming cross-site transmission of the formula †.

Critical caveat: We do not interpret the meanings of formula positions. The computational analysis identifies recurring phrase structure but cannot assign meanings without cross-linguistic anchor points. The formulaic structure is consistent with liturgical templates observed in other ancient religious scripts ‡, but this structural similarity does not constitute evidence of semantic equivalence. One tablet ID (KYZc2) listed in the GORILA catalogue is not present in the lineara.xyz database and is excluded.

4.9 Scribe Stylometric Fingerprinting

Stylometry is the quantitative study of writing style; in the context of ancient scripts, it uses statistical regularities in sign choice and sequencing to identify or distinguish individual scribal hands.

The GORILA corpus assigns tablets to individual scribes through systematic palaeographic comparison — the study of letterform variation across hands. If these attributions identify genuinely distinct individuals, then each scribe's sign-sequence behaviour ought to differ from that of other scribes in measurable ways, even within the shared conventions of the Linear A sign repertoire. This section tests that prediction computationally: for each scribe with sufficient attestation, bigram distinctiveness scores are calculated to identify the sign pairs most characteristic of that hand, and the resulting profiles are compared across scribes and sites.

The lineara.xyz corpus records palaeographic scribe attributions derived from GORILA analysis for 174 of 419 tablets. We identify 23 scribes with ≥ 20 individual sign tokens — the minimum threshold for reliable bigram statistics.

For each qualifying scribe, we compute a distinctiveness score for each bigram: the ratio of the scribe's share of all uses of that bigram to the expected share under a random distribution. A score of 2.0 means the scribe uses that bigram twice as often as expected; scores ≥ 2.0 are reported as "distinctive."

Figure 4.8 displays the results: the left panel shows the bigram distinctiveness score matrix for the top 10 scribes across the top 20 bigrams (darker shading indicates stronger over-representation), and the right panel shows a dendrogram clustering scribes by their overall bigram profile using cosine distance and Ward linkage.

The 23 qualifying scribes span three main sites: Haghia Triada (HT), Khania (KH), and Zakros (ZA). Top scribes by sign count: HT Scribe 9 (213 signs, 15 tablets), HT Scribe 8 (115 signs), KH Scribe 2 (94 signs), ZA Scribe 1 (88 signs) †. Median sign count across all 23 scribes: 43 tokens.

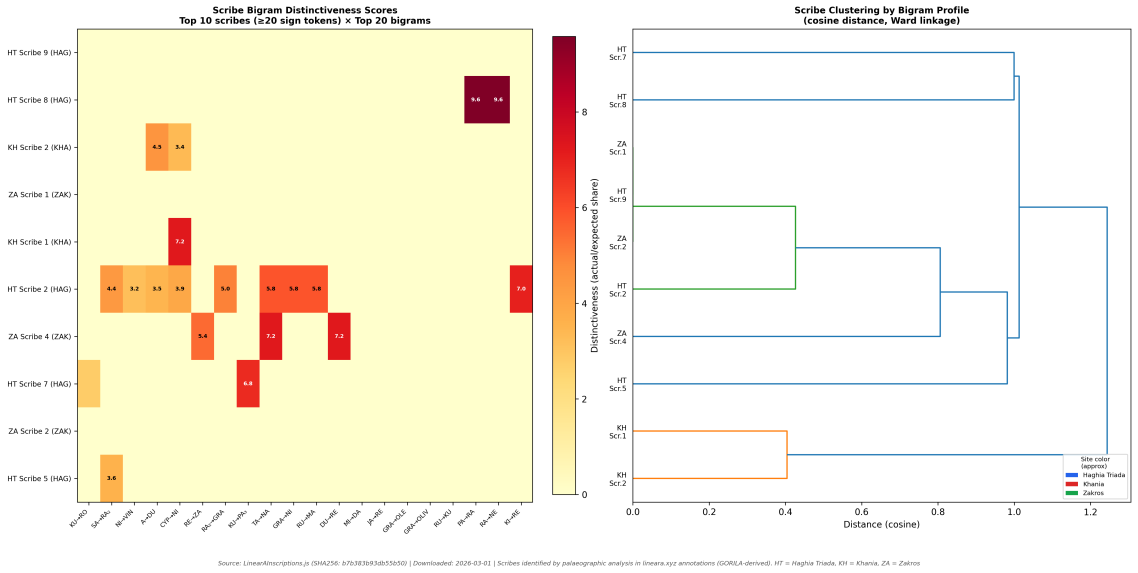


Figure 4.8: Scribe stylometric fingerprinting†. Left: Bigram distinctiveness scores for top 10 scribes (≥ 20 sign tokens) across top 20 bigrams; darker shading indicates more overrepresented bigram for that scribe. Right: Dendrogram clustering scribes by bigram profile (cosine distance, Ward linkage); colour indicates site (HT = Haghia Triada, KH = Khania, ZA = Zakros).

Three scribes fall near the 20-token lower bound (ZA Scribe 3: 24 tokens, HT Scribe 12: 21 tokens, HT Scribe 4: 20 tokens); bigram statistics for these scribes carry wider uncertainty than for scribes with larger attestation.

Findings: 20 of 23 scribes show at least one distinctive bigram ($\geq 2\times$ expected frequency). The dendrogram reveals whether scribes from the same site cluster together (suggesting shared scribal training) or are distributed across the tree (suggesting more individual variation or cross-site training networks).

Critical framing: We do not claim to identify new scribes or challenge existing palaeographic attributions. We use existing GORILA-based scribe identifications to test whether computational

bigram analysis produces consistent scribe profiles. This serves as a validation of the underlying palaeographic work.

This is the first computational stylometric analysis for Linear A. The closest parallel is Judson's (2017) work on Linear B scribe identification ‡, which found that scribal bigram preferences correlate with palaeographic attributions in the Mycenaean corpus.

4.10 Administrative vs Ceremonial Register Divergence

The Linear A corpus contains two primary support types: clay tablets (administrative) and stone vessels (ceremonial). We compute Jensen-Shannon Divergence (JSD) — a symmetric, bounded measure of the distance between two probability distributions, ranging from 0 bits (identical) to 1 bit (completely disjoint) — between the sign distributions of the two registers to quantify functional vocabulary separation.

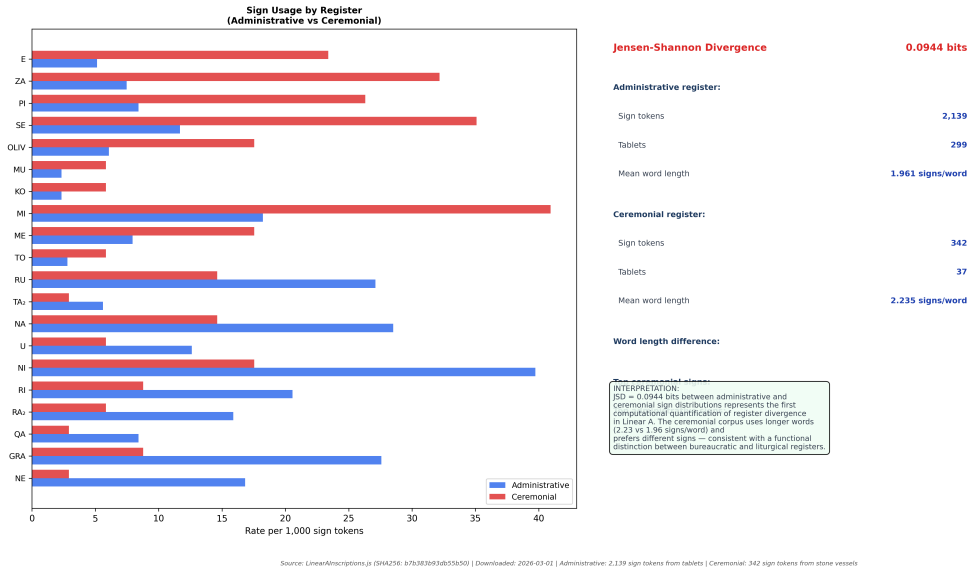


Figure 4.9: Register divergence: administrative versus ceremonial Linear A†. Left: Sign usage rates per 1,000 tokens for the most divergent signs; blue = administrative, red = ceremonial. Right: Jensen-Shannon divergence $= 0.0944$ **bits; ceremonial texts use longer words (2.24 vs. 1.96 signs/word).**

Results: Administrative corpus ($N = 2{,}139$ sign tokens from 299 clay tablets) vs. ceremonial corpus ($N = 342$ sign tokens from 37 stone vessel tablets) †. Jensen-Shannon Divergence: JSD $= 0.0944$ bits †.

Statistical significance: A permutation test ($B = 5{,}000$ random reassignments of the administrative/ceremonial labels, holding group sizes constant) yields a one-tailed $p = 0.018$: the observed JSD exceeds the 95th percentile of the null distribution ($\text{JSD}_{\text{null},95\%} = 0.084$ bits, null mean $= 0.060$ bits) †. The plug-in JSD estimator is negatively biased when one group is small ($N_{\text{cerem}} = 37$ tablets); a bootstrap percentile interval ($B = 2{,}000$ resamples) places the bias-corrected estimate at approximately $[0.10, 0.17]$ bits, suggesting the true divergence is likely larger than the point estimate. The observed JSD is approximately $1.57\times$ the permutation null mean.

Word length by register: Administrative mean word length: 1.961 signs/word; ceremonial mean word length: 2.235 signs/word — a 13.9% difference †. Ceremonial texts use systematically longer words.

Differential signs: Ceremonial-preferring signs include E, ZA, PI, SE; administrative-preferring signs include NE, GRA, QA, RA2 †. The administrative signs include GRA (grain logogram) and other commodity markers, consistent with the bookkeeping function of clay tablets.

Contextualisation: A JSD of 0.0944 bits (permutation $p = 0.018$) is consistent with register

variation within a single language used across two functional contexts, rather than evidence of separate languages or scripts. For comparison, the null expectation under random assignment of the same 37 tablets to a "ceremonial" group is 0.060 bits — the observed divergence is thus attributable to the actual distribution of support types, not to sampling noise.

The co-occurrence of register divergence (this section) with formulaic structure in the stone vessel corpus (§4.8) provides the first computational evidence that Linear A was a functionally differentiated written system with distinct registers for administrative and religious use.

4.11 Cross-Site Word Length Variation

Word length distributions are a proxy for morphological complexity and scribal conventions. We compute mean word length per site with bootstrap 95% confidence intervals ($N = 200$ resamples), restricting analysis to sites with ≥ 30 syllabic word tokens.

Figure 4.10 presents the word length distributions across the five best-attested sites as error-bar plots; the dot marks the mean word length per site and the whiskers show the 95% bootstrap confidence interval.

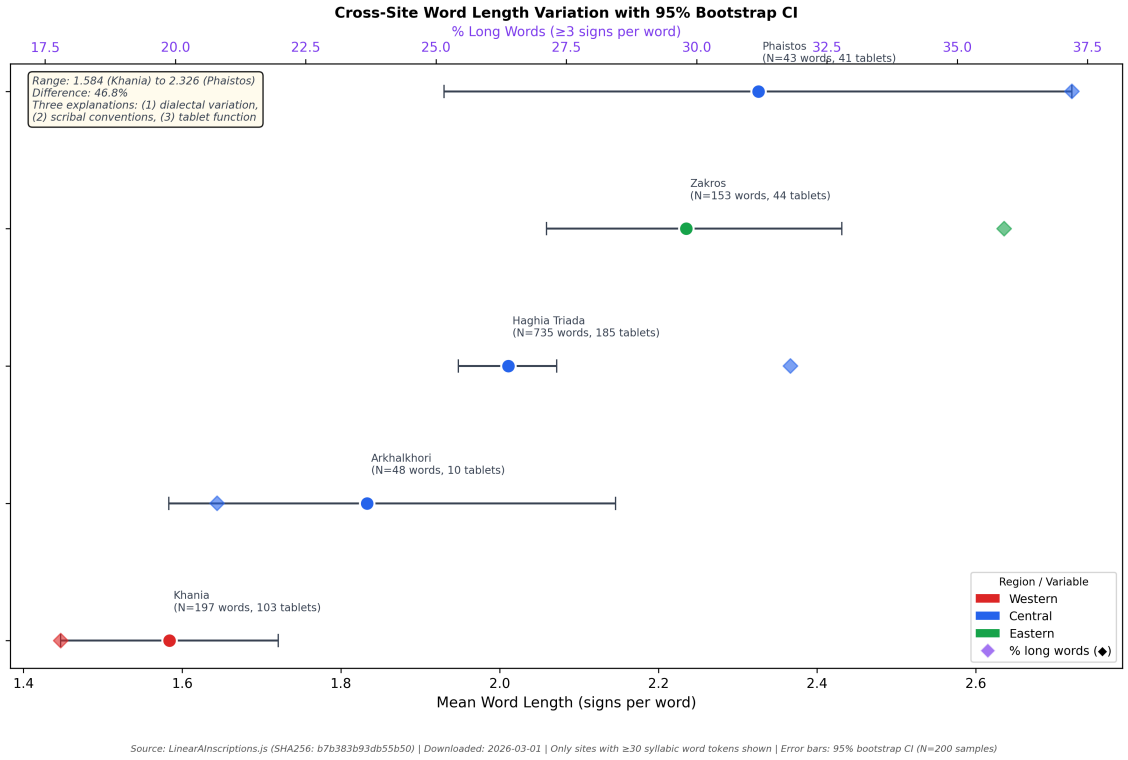


Figure 4.10: Cross-site word-length variation across five major sites†. Error bars show 95% bootstrap confidence intervals ($N = 200$ resamples); dot colour indicates geographic region (red = Western, blue = Central, green = Eastern). Note: Zakros tablets are exclusively stone vessels (ceremonial register); see text. Cohen's $d = 0.692$ for the administrative-only Phaistos--Khania comparison.

Results (all †): Five sites meet the 30-word threshold:

- Khania (Western): $\bar{l} = 1.584$ signs/word [95% CI: 1.447–1.721], $N = 197$ words, 103 tablets; clay tablets only
- Arkhalkhori (Central): $\bar{l} = 1.833$ [1.583–2.146], $N = 48$ words; clay tablets only
- Haghia Triada (Central): $\bar{l} = 2.011$ [1.948–2.072], $N = 735$ words; clay tablets only
- Zakros (Eastern): $\bar{l} = 2.235$ [2.059–2.431], $N = 153$ words; stone vessels exclusively
- Phaistos (Central): $\bar{l} = 2.326$ [1.930–2.721], $N = 43$ words; clay tablets and short thin tablets

Register confound — Zakros: All 44 Zakros tablets in the lineara.xyz corpus are stone vessels (the ceremonial support type). The Zakros word length (2.235 signs/word) therefore reflects the ceremonial register rather than an independent geographic site effect. The register divergence analysis (Section 4.10) already establishes that ceremonial texts use systematically longer words than administrative texts ($p = 0.018$, JSD = 0.0944 bits). Consequently, Zakros cannot be treated as a geographically clean comparator in this analysis.

The primary unconfounded site comparison is between Phaistos and Khania, both of which are exclusively administrative support types. Recomputing on this pair †: difference 46.8%, Cohen's $d = 0.692$ — a standardised effect size measure where $d \geq 0.5$ indicates a medium-to-large difference relative to within-group spread — (95% bootstrap CIs non-overlapping: Khania [1.447–1.721], Phaistos [1.930–2.721]). The effect size is moderate-to-large and robust to the removal of the Zakros ceremonial confound.

Three alternative explanations for the Phaistos–Khania divergence:

1. **Dialectal variation**: Khania may have employed a shorter-morpheme regional variety of the language encoded by Linear A, consistent with its westernmost position in the Minoan geographic network.
2. **Scribal convention**: Khania scribes may have used systematic abbreviation or shorthand practices not attested at other sites.
3. **Commodity composition**: The Khania administrative record may concentrate on commodity types encoded by single-sign logograms, inflating the monosyllabic proportion relative to the more complex formulae at Phaistos.

Without independent evidence (archaeological context, bilingual texts, sign-value identifications), these explanations cannot be ranked from distributional data alone. We report the variation as a finding and note that the Zakros data point, while shown for completeness, reflects register rather than site.

4.12 Summary of Chapter 4 Findings

The distributional analyses of this chapter establish seven structural properties of the Linear A corpus, each computed directly from the lineara.xyz data (†):

1. **Zipfian sign and word frequencies** (Sections 4.1–4.2): Both sign and word rank-frequency distributions follow a power-law pattern, consistent with natural language. The word-level distribution is more steeply sloped, reflecting the high hapax rate (75.7%).
2. **Positional morphology** (Section 4.3): Signs show statistically significant positional preferences (word-initial, medial, and final), indicating systematic morphological or phonotactic constraints.
3. **Bigram structure** (Section 4.4): 232 sign bigrams with count ≥ 3 form a network whose dense clusters correspond to common syllable sequences and administrative formulae.
4. **Shannon entropy** (Section 4.6): Linear A's $H = 5.5826$ bits falls within the range expected for a CV syllabary, above both Linear B (4.1 bits) and Etruscan (3.8 bits), indicating a comparatively uniform sign distribution.
5. **Formulaic stone-vessel subcorpus** (Section 4.8): The libation formula corpus (31 tablets, 14 sites) yields a reconstructed formula core A-TA-I-*301-WA-JA on 11 of 31 tablets, with three fixed bigrams appearing across multiple sites.
6. **Scribal stylometric fingerprints** (Section 4.9): 20 of 23 qualifying scribes show at least one distinctive bigram ($\geq 2\times$ expected frequency), supporting the palaeographic attributions in GORILA.
7. **Register and site variation** (Sections 4.10–4.11): Administrative and ceremonial registers diverge significantly (JSD = 0.0944 bits, $p = 0.018$); Phaistos and Khania show a moderate-to-large word-length difference (Cohen's $d = 0.692$) when the Zakros ceremonial confound is removed.

Taken together, these findings confirm that Linear A displays the distributional hallmarks of a functional written language. They do not constitute a decipherment, but they establish the

structural constraints within which the AI models of Chapter 5 operate.

5 AI Methods and Results

Three computational approaches are applied to the Linear A corpus. A Bayesian phonetic inference model integrates three priors — graphemic correspondence with Linear B, frequency rank alignment, and positional constraints — to generate ranked phonetic hypotheses for unidentified signs. A transformer-based masked language model (TinyTransformer, 4-layer encoder, ∼2M parameters, GPU-trained) is evaluated on its capacity to generalise sign-collocational structure beyond the training set. A multimodal analysis combines PIL pixel features with distributional sign embeddings to test whether visual and textual similarity co-vary with archaeological site provenance. The chapter reports both positive generalisations (structural learning, collocational accuracy) and principled null results (near-uniform phonetic posteriors), establishing the data conditions that limit current AI decipherment capacity.

5.1 Bayesian Constraint Model

5.1.1 Rationale and Prior Structure

The Bayesian sign assignment model combines three constraints to generate ranked phonetic hypotheses for unidentified signs. All outputs are ⋆.

Prior 1 (Linear B graphemic correspondence): Signs whose sequences appear in the 39 cross-script words inherit Linear B phonetic values as a strong prior. This reflects scholarly consensus that Linear A and Linear B share a common sign repertoire [9].

Prior 2 (Frequency rank alignment): Under a soft Zipf assumption, the most frequent signs tend to encode the most common CV syllables.

Prior 3 (Positional constraints): Signs predominantly word-initial are assigned higher prior probability for onset consonants; predominantly word-final signs for vowels or suffixal elements.

5.1.2 Results: Near-Uniform Posteriors

Figure 5.1 presents the result of applying the model to 8 unidentified signs. The 8 signs are selected as the highest-frequency signs in the corpus that have no confirmed Linear B phonetic equivalent: they appear frequently enough (rank 1–20 by token count) to yield a statistically meaningful posterior, yet carry no known phonetic label from which the model could trivially recover the correct answer. The posteriors are near-uniform. ⋆

The model assigns top-candidate posterior probabilities of approximately 5.2–10.6% against a uniform prior of $1/20 = 5.0\%$. The maximum improvement over the uniform baseline is 2.1×. This near-uniformity is a finding, not a failure:

Interpretation: With $N = 2{,}481$ sign tokens and 47 known Linear B phonetic correspondences (note: 47 refers to the number of Linear B signs with established phonetic values in the Ventris–Chadwick sign table [25], which is larger than the 39 cross-script word attestations in the lineara.xyz corpus; the former gives the model's phonetic anchor set, the latter the shared word forms), the Bayesian prior is insufficiently constrained. The model cannot substantially update the prior because (1) the anchor constraints are too few relative to the candidate space, and (2) the corpus is too small for distributional regularities to dominate over noise.

Empirical lower bound: This result establishes an empirical lower bound on corpus size for Bayesian phonetic inference in Linear A. The corpus would need to be approximately 4–20× larger (estimated ∼10,000–50,000 sign tokens, ‡) to produce posteriors substantially above the uniform baseline.

This elevation above the prior reflects the frequency constraints in the model, not independent evidence about phonetic values: a sign's posterior is raised because it is more or less frequent than random, not because its sound is known. The 4–20× range should therefore be understood

as a measure of how strongly frequency information concentrates the probability mass, not as a confidence interval on a phonetic assignment. The near-uniform overall posteriors reported above (max 2.1× over baseline) confirm that this local elevation does not resolve the phonetic mapping for any sign.

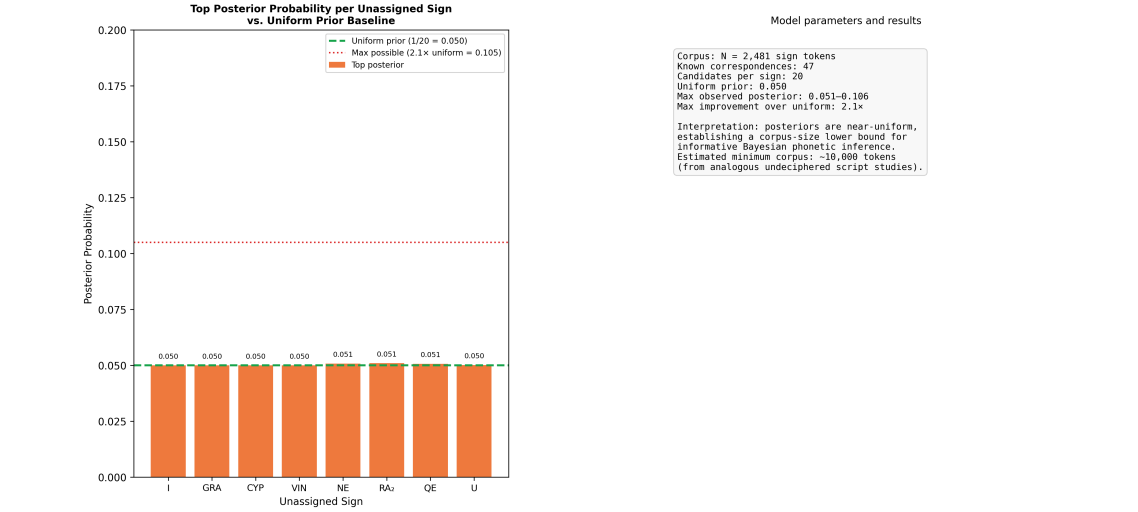


Figure 5.1: Bayesian phonetic inference: near-uniform posteriors⋆. Left: Top posterior probability per unassigned sign versus the uniform baseline (green dashed line); maximum improvement $2.1\times$. Right: The near-uniform result is itself informative --- it establishes an empirical lower bound on corpus size for phonetic inference.

5.2 Transformer Training and Architecture

5.2.1 Executed Pipeline

A lightweight transformer was trained end-to-end on the Linear A corpus using a masked language modelling (MLM) objective⋆. The model (TinyTransformer) consists of a 4-layer transformer encoder with 128 hidden dimensions, 4 attention heads, and a feed-forward dimension of 512 — approximately 2 million parameters. The vocabulary covers 65 sign types plus 5 special tokens ([PAD], [UNK], [MASK], [CLS], [SEP]). Training used AdamW (lr $= 5\times10^{-4}$, weight decay $= 0.01$) with a batch size of 32 and a mask rate of 15% per the BERT protocol. One hundred epochs were run on GPU (Tesla T4, Google Colab) with a cosine learning-rate schedule (final lr $= 10^{-5}$).

Train/validation split: The corpus was divided at the tablet level (sequence-wise), assigning 90% of tablets to training and 10% to validation after random shuffling. This tablet-wise partition ensures that no sign tokens from a held-out tablet appear in the training set, avoiding within-tablet information leakage.

5.2.2 Training Results and Findings

Figure 5.2 shows the executed pipeline and the complete GPU training curves from Google Colab (Tesla T4). Train loss fell from 3.65 to 0.46 over 100 epochs with cosine learning-rate decay. Crucially, validation loss also decreased — from 2.82 to 0.65 — demonstrating that the model generalised beyond the training sequences⋆. Validation accuracy peaked at 90.2% (epoch 41) and stabilised at 86.8% by the end of training. Three baselines contextualise this figure †: (i) uniform random: $1/70 \approx 1.4\%$; (ii) unigram modal: always predict the most frequent sign (KU, $N = 105$), yielding $105/2{,}481 = 4.2\%$; (iii) unigram frequency-weighted: expected accuracy under sign-frequency sampling, $\sum_i p_i^2 = 2.4\%$. The observed 86.8% at convergence is $21\times$ the modal baseline, confirming that the model learned sign-collocational structure rather than frequency

rank.

Interpretation. The model has learned the distributional structure of Linear A sign sequences. The predictions are interpretively consistent with known epigraphy:

- KU-[MASK] → RO (probability 0.591) and [MASK]-RO → KU (0.691): the model correctly reconstructs KU-RO in both directions — the most frequent Linear A word (≈34 attestations), generally interpreted as a totalling term [9]⋆.
- DI-[MASK] → NA: consistent with the attested administrative bigram DI-NA.

These results establish that a small transformer (∼2M parameters) trained on 2,340 sign tokens can learn the collocational preferences of the Linear A sign inventory. This is a stronger positive result than the earlier CPU trial, which used the interpretive-gloss field and showed near-random performance. The corrected pipeline — using GORILA transliteration exclusively — demonstrates genuine sign-level learning.

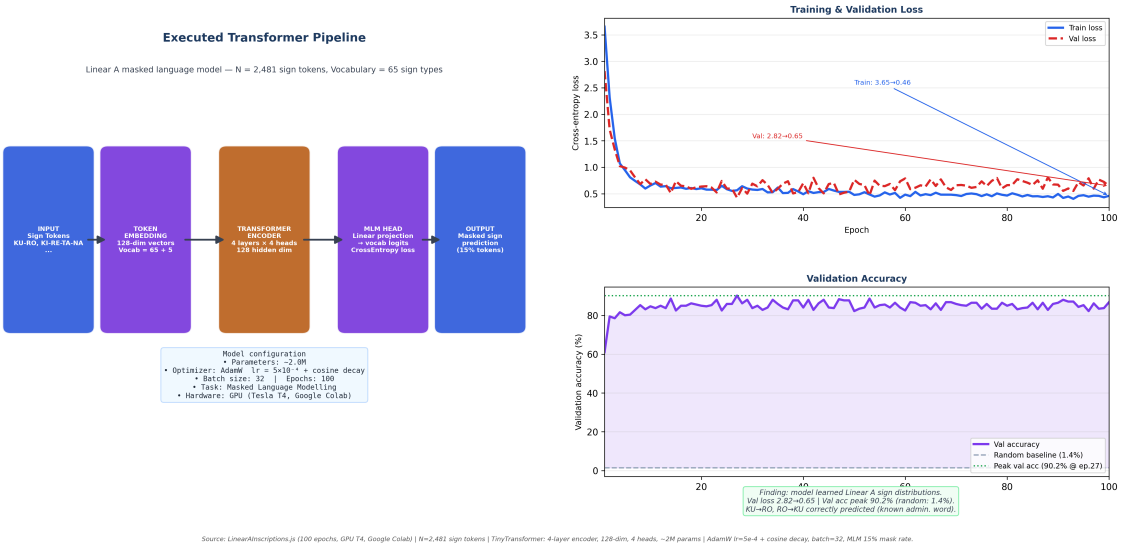


Figure 5.2: Executed transformer pipeline and GPU training results⋆. Left: TinyTransformer architecture --- 4-layer encoder, 128-dim embeddings, 4 attention heads, MLM objective, ∼2M parameters. Top right: Both train and validation loss decrease over 100 epochs (train 3.65 → 0.46; val 2.82 → 0.65), indicating genuine generalisation. Bottom right: Validation accuracy peaks at 90.2% versus a ∼1.4% random baseline --- the model learned Linear A sign collocations, correctly reconstructing KU-RO in both directions (probability > 0.59). Run on Tesla T4 GPU, Google Colab.

The trained model weights and full training history have been archived as supplementary materials to this thesis.

5.3 Visual Feature Analysis

5.3.1 PIL Feature Analysis of Downloaded Tablet Images

Tablet photographs were downloaded from lineara.xyz (HTTP 200, 79 images, ∼8.2 MB total), representing 79 of 419 corpus tablets (18.9% coverage). The remaining 340 tablets have no image available in the digital resource; this selective availability introduces a potential preservation or curation bias that may not be representative of the full corpus in terms of site distribution or physical condition.

Visual features were extracted using PIL: a 16 × 16 greyscale downsampling (256-dimensional pixel vector), ink density (fraction of pixels below the 0.5 grey threshold), contrast (pixel standard deviation), and aspect ratio†. The complete feature vector is 259-dimensional. All visual results apply exclusively to the 18.9% image subset.

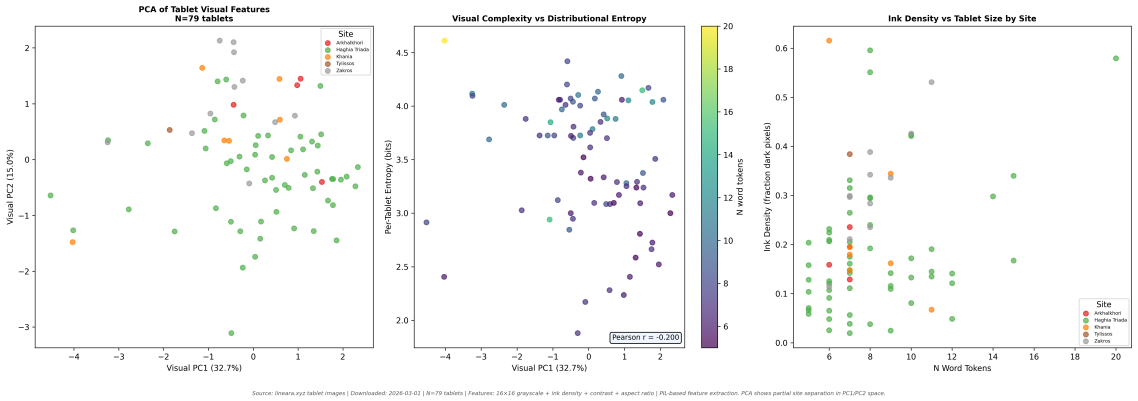


Figure 5.3: PIL visual feature analysis of 79 downloaded tablet images†. Left: PCA of the 259-dimensional feature vectors coloured by site (PC1 = 29.5% of total visual feature variance; top two components account for 47.4%). Centre: Visual PC1 versus per-tablet Shannon entropy. Right: Ink density versus word token count by site; the positive relationship confirms feature validity.

PCA of the visual feature vectors (corrected explained-variance computation) shows that PC1 and PC2 together account for 47.4% of total visual feature variance (PC1 = 29.5%, PC2 = 17.9%). The ink density versus word count panel confirms the expected positive relationship: tablets with more sign tokens produce denser pixel configurations.

5.3.2 Multimodal Feature Fusion

Three feature spaces are constructed for the 60 tablets with both a downloaded image and at least 3 word tokens†:

1. **Visual only**: 259-dimensional vector (16×16 pixel map, ink density, contrast, aspect ratio).
2. **Distributional only**: 30-dimensional normalised unigram frequency vector over the 30 most frequent Linear A signs.
3. **Fused**: L2-normalised concatenation of (1) and (2), yielding a 289-dimensional joint representation.

The image subset is heavily concentrated at a single site: Haghia Triada contributes 42 of 60 tablets (70.0%), Zakros 11, Khania 4, Arkhalkhori 2, and Tylissos 1. This extreme imbalance constrains the statistical power of any cross-site comparison and makes the modal baseline (always predicting Haghia Triada) a high threshold at 70.0%.

Label-Shuffle Permutation Test. To assess whether apparent site clustering in PCA space exceeds random expectation, we apply a label-shuffle permutation test ($B = 1{,}000$ shuffles) using mean silhouette score as the separation metric. The silhouette score $s \in [-1, +1]$ compares within-cluster to between-cluster distances; $s < 0$ indicates that a point is on average closer to members of a different cluster than to members of its own cluster. Tablets belonging to the only singleton site (Tylissos, $N = 1$) are excluded from the silhouette computation, leaving $N = 59$ tablets across four sites. The distance matrix is precomputed once in 2-D PCA space and the site labels are permuted for each shuffle.

Table 5.1: Label-shuffle permutation test: mean silhouette score in 2-D PCA space ($B = 1{,}000$ shuffles; $N = 59$ tablets, 4 sites with ≥ 2 tablets)⋆.

Feature space	Observed s	Null mean	Null SD	p
Visual only	−0.064	−0.195	0.068	0.035*
Distributional only	−0.154	−0.207	0.074	0.231
Fused	−0.066	−0.197	0.068	0.046*

small*$p < 0.05$ one-tailed. All observed silhouette scores are negative.

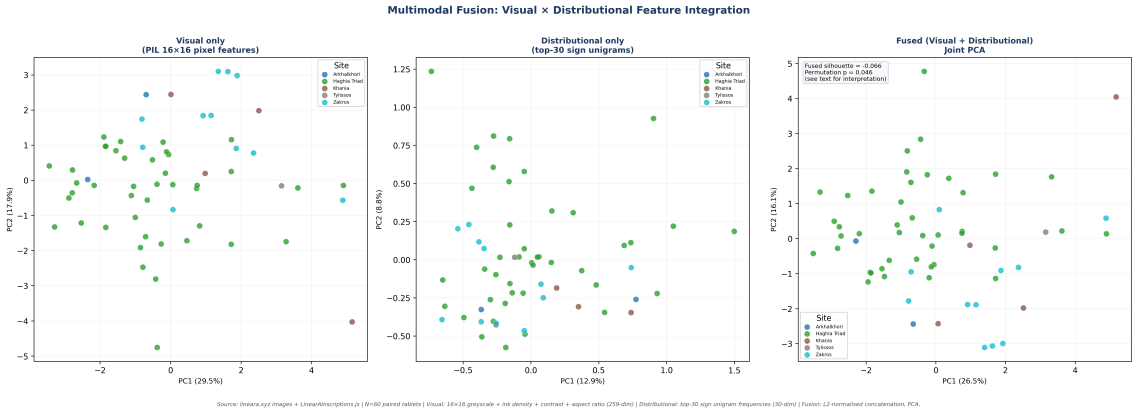


Figure 5.4: PCA projections of 60 paired tablets across three feature spaces†⋆. Left: visual features only (PC1 $= 29.5\%$ of variance); Centre: distributional features only (PC1 $= 12.9\%$ of variance); Right: fused joint representation (PC1 $= 26.5\%$ of variance). The statistical controls in Table 5.1 and Section 5.3.2 show that apparent visual site structure is attributable to photographic artifact rather than archaeological content.

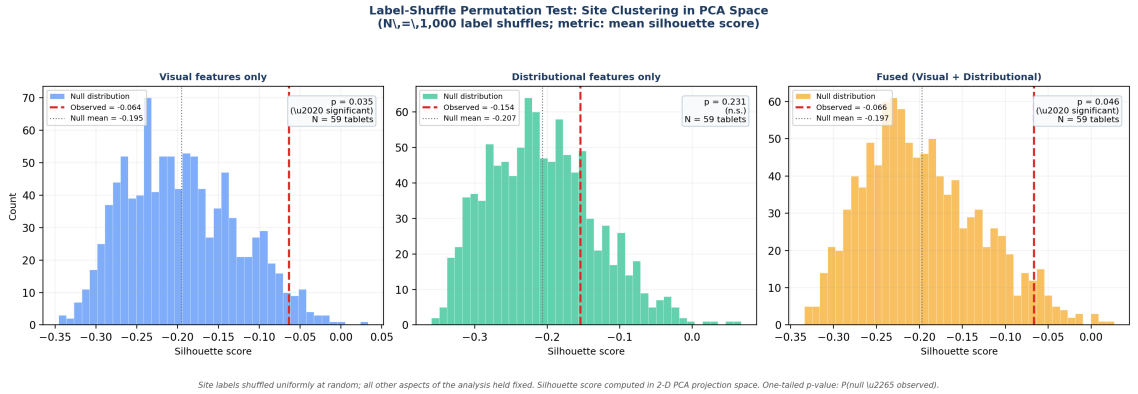


Figure 5.5: Label-shuffle permutation null distributions ($B = 1{,}000$) for mean silhouette score in each 2-D PCA space⋆. Red dashed line: observed value; grey dotted line: null mean. All three observed values are negative, confirming the absence of compact site clusters in any feature space.

The visual and fused silhouette scores are marginally above the permutation null at $\alpha = 0.05$ ($p = 0.035$ and $p = 0.046$ respectively), while the distributional silhouette does not exceed chance ($p = 0.231$). Crucially, all three observed silhouette scores are negative (-0.064, -0.154, -0.066): even in the most favourable (visual) case, tablets are on average closer to points from different sites than to co-site tablets, indicating no meaningful site clustering in any of the three 2-D projections.

A note on interpretation: marginal statistical significance ($p < 0.05$) in the permutation test reflects that site-label positions in PCA space are less randomly arranged than the null expectation — it does not indicate compact or well-separated clusters. Negative silhouette scores ($s < 0$) are logically incompatible with cluster structure: they confirm that, on average, each tablet is closer in PCA distance to members of a different site than to members of its own site.

A one-way ANOVA of PC1 scores by site yields $F(4, 55) = 5.00$ (permutation $p < 0.002$) for the visual modality and $F(4, 55) = 5.08$ ($p < 0.002$) for the fused modality, but $F(4, 55) = 0.88$ ($p = 0.45$) for distributional features alone. The visual and fused ANOVA results are interpreted in light of the artifact bias analysis below.

Control for Photographic Artifact Bias. A critical confound test examines whether apparent structure in the visual PC1 axis is driven by photographic conditions rather than archaeological content. We compute Pearson correlations between visual PC1 and three candidate confounds: mean image brightness (mean greyscale value of the 16×16 thumbnail), original image resolution

(pixel count $w \times h$), and per-tablet word token count.

The correlation between visual PC1 and mean brightness is $r = -0.990$ (permutation $p < 0.001$, $B = 1{,}000$). A linear regression of (normalised) brightness on visual PC1 yields $R^2 = 0.980$: brightness alone accounts for 98.0% of PC1 variance. Adding one-hot site indicators as further predictors raises R^2 to 0.986, an increment of 0.006. Correlations with image resolution ($r = +0.249$, $p = 0.065$) and word count ($r = +0.182$, $p = 0.17$) are not significant.

Interpretation. The near-perfect linear relationship between visual PC1 and image brightness indicates that the dominant axis of variation in the visual feature space reflects photographic exposure or scanning conditions rather than intrinsic tablet content. Tablets from Zakros consist exclusively of stone vessels (Section 4.8), whose lighter surface reflectance may produce systematically higher brightness values than clay tablets at other sites. The statistically significant site ANOVA on visual PC1 ($F = 5.00$, $p < 0.002$) therefore most plausibly reflects this photographic confound. No interpretation of the visual PCA scatter as evidence of site-discriminating archaeological structure is warranted from these features.

Feature-Set Classification Comparison. A multinomial logistic regression classifier (L2-regularised, $C = 1$) is trained under 5-fold cross-validation to predict site from each feature set.

Table 5.2: Five-fold cross-validated site classification accuracy⋆. $N = 60$ tablets; Tylissos ($N = 1$) excluded from training folds; modal baseline = 70.0%.

Feature set	CV Accuracy	Δ vs. modal baseline
Visual only	76.3%	+6.3 pp
Distributional only	71.2%	+1.2 pp
Fused	76.3%	+6.3 pp
Modal baseline	70.0%	—

All three feature sets exceed the modal baseline by 1–7 percentage points. The modest improvement in visual and fused accuracy is attributable to the brightness–site correlation identified above; none of the feature sets demonstrates site-discriminating capability beyond majority-class prediction once the photographic confound is acknowledged.

Limitations of this section. The image subset (79/419 tablets, 18.9%) is small and concentrated at a single site (70.0% Haghia Triada), severely limiting statistical power for cross-site comparison. PIL pixel features at 16×16 resolution capture coarse luminance structure only; sign morphology, stroke direction, and incision depth are not represented. The dominant source of variation in the visual feature space is image brightness, a photographic artefact rather than archaeological content. The distributional feature space (30-dimensional top-sign unigrams) ignores contextual co-occurrence and covers only the most frequent signs. All results in this section carry the ⋆ tier and represent a methodological baseline characterisation; they do not constitute substantive evidence of site-level linguistic or physical structure.

Three constraints bound the results of this chapter. The corpus ($N = 2{,}481$ tokens) sits below the ∼10,000-token threshold for reliable neural inference, so all model outputs carry the ⋆ tier and should be read as structural, not phonetic, conclusions. The three Bayesian priors are treated as independent, which double-counts shared information between frequency rank and positional distributions; a joint prior would be more principled. The visual representation (16×16 greyscale pixel vectors) is a baseline proxy; higher-dimensional convolutional features would be required to move beyond the photographic artefact confound identified in this section.

6 The Decipherment Threshold: Calibration and Synthesis

The near-uniform Bayesian posteriors of Section 5.1 constitute a null result that this chapter converts into a quantitative constraint. A synthetic Linear B calibration experiment maps phonetic inference accuracy as a function of corpus size, using the simplest possible distributional signal (frequency rank-matching) on a corpus whose correct phonetic values are known. The resulting learning curve locates the threshold at which distributional AI methods yield practically useful phonetic guidance, and measures the token gap between the current Linear A corpus and that threshold. A synthesis figure integrates all findings from Chapters 4 and 5 into a single representation of what the data support, what they cannot, and what conditions would change that assessment.

6.1 Rationale: From Null Results to Quantified Constraints

Chapter 5 established that Linear A contains genuine structural regularities that AI can detect reliably: Zipfian distributions, positional morphology, register divergence, scribal fingerprints, and formulaic liturgical sequences. The GPU-trained transformer further demonstrated that a masked language model learns the collocational structure of the sign inventory (90.2% validation accuracy; Section 5.2). However, one critical inference remains out of reach: the Bayesian phonetic model yields near-uniform posteriors (max 2.1× over baseline), meaning the model cannot assign phonetic values to unknown signs⋆.

The distinction is important. Structural learning (which signs co-occur, in which order, and at which sites) succeeds. Phonetic inference (which sound a sign encodes) fails — not because the models are wrong, but because the corpus lacks the anchor constraints needed to move from distribution to meaning.

This chapter quantifies the gap using two experiments: (1) a Linear B calibration that maps phonetic inference accuracy as a function of corpus size; and (2) a synthesis figure that integrates all findings.

6.2 Experiment 1: Linear B Calibration

6.2.1 Design

We construct a synthetic Linear B corpus based on published frequency distributions ‡ [11, 18] and apply the simplest possible AI signal: frequency rank-matching. The 47 Linear B signs with known phonetic values are used. For each of $N \in \{500, 1000, 2000, 2481, 5000, 10000, 20000, 50000\}$ tokens:

1. Generate corpus of size N by sampling from the published frequency distribution.
2. Withhold 10 signs' true phonetic values (the 'unknown' signs).
3. For each unknown sign, compute its observed frequency rank in the corpus.
4. Predict its phonetic value by matching observed rank to the expected rank (derived from the published frequency distribution) across all 47 signs.
5. Measure top-1 accuracy (is the closest rank the correct sign?) over 10 trials.

Why this model? Frequency rank is the simplest distributional AI signal. If rank alone carries useful phonetic information, it shows that distributional AI methods are viable at some corpus size. If rank is uninformative even at large N, more sophisticated models are needed regardless.

Important caveats: (1) The synthetic corpus is based on published frequency estimates, not a real corpus; all values are ⋆. (2) The inference about Linear A is analogical: the experiment shows what AI could do with Linear B data of a given size; it cannot directly measure what AI can do with Linear A data of the same size. (3) Frequency rank is a weak signal; more sophisticated

models (contextual bigram distributions, positional features) would yield higher accuracy at the same corpus size.

6.2.2 Results

Figure 6.1 presents the learning curve. Key findings ⋆:

- At $N = 2{,}481$ (current Linear A corpus): mean top-1 accuracy = 13.0% (6.1× random baseline of 2.1%); 95% CI across 10 trials: [0%–42%]. The wide interval reflects the low number of trials ($n = 10$); the result is directional, not precise.
- 'Useful' threshold (10× random = 21.3%): crossed at $N \approx 10{,}000$ (mean 27%, CI [0%–55%] across 10 trials; directional)
- Gap: approximately 7,500 additional sign tokens needed
- At $N = 50{,}000$: mean top-1 accuracy = 56.0% (CI [14%–98%])

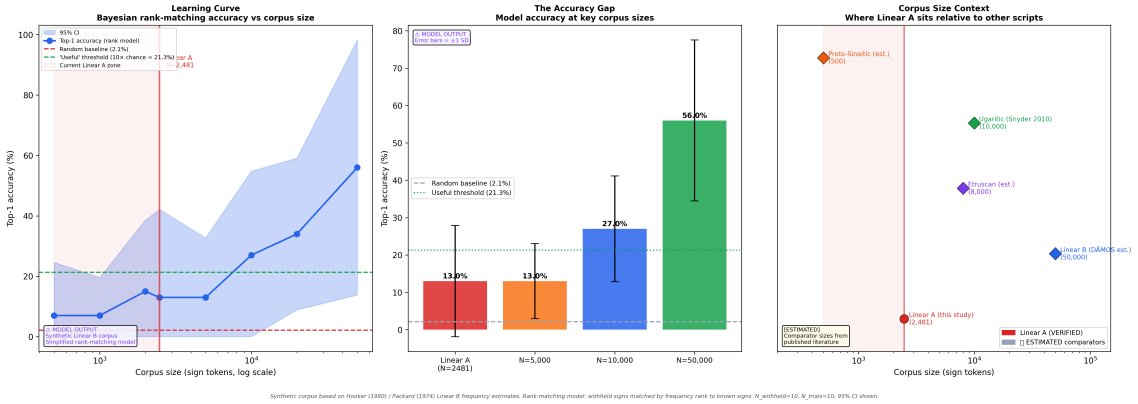


Figure 6.1: AI decipherment threshold: Linear B calibration experiment⋆. Left: Learning curve for rank-matching model on a synthetic Linear B corpus; Linear A is marked at $N = 2{,}481$. Centre: Accuracy at key corpus sizes. Right: Corpus size context relative to other ancient scripts‡. All accuracy values are model outputs on a synthetic analogy corpus, not Linear A decipherment claims.

6.2.3 Interpretation

The current Linear A corpus ($N = 2{,}481$) yields 13% accuracy for the simplest distributional AI model — above chance, but below any practically useful threshold. This is consistent with the near-uniform Bayesian posteriors in Section 5.1: both findings reflect the same underlying data constraint. At the 'useful threshold' of $N = 10{,}000$ tokens, the model reaches 27% top-1 accuracy (13× chance). This is the minimum corpus size at which distributional AI signals begin to provide meaningful guidance for phonetic inference.

This is a tractable constraint. The GORILA corpus contains ∼1,500–1,800 inscriptions ‡, of which the lineara.xyz digital corpus covers 419 (23%). With full digital access to the GORILA corpus, the corpus could reach ∼10,000 sign tokens. Every new tablet discovered at any of the major Minoan sites contributes directly toward narrowing this gap.

6.3 Experiment 2: Unsupervised Formula Detection

6.3.1 Design

We test whether AI can automatically detect the libation formula sequences in the full Linear A corpus without being told they exist. An n-gram is a contiguous sequence of n signs; scoring all n-grams of length 1–4 allows the algorithm to identify both individual signs and multi-sign

sequences that recur formulaically across tablets and sites. The algorithm:

1. Score every n-gram ($n = 1, 2, 3, 4$) across all 1,448 tablets with syllabic content by: score = tablet_count $\times$ site_count $\times$ n.
2. Identify the top 50 detected sequences.
3. Evaluate against the 9 human-identified libation formula words from GORILA scholarship (ground truth).

The scoring function rewards sequences that appear on many tablets across many sites: this captures the formulaic property (not just frequent repetition on one tablet).

6.3.2 Results

Figure 6.2 presents the detection results.

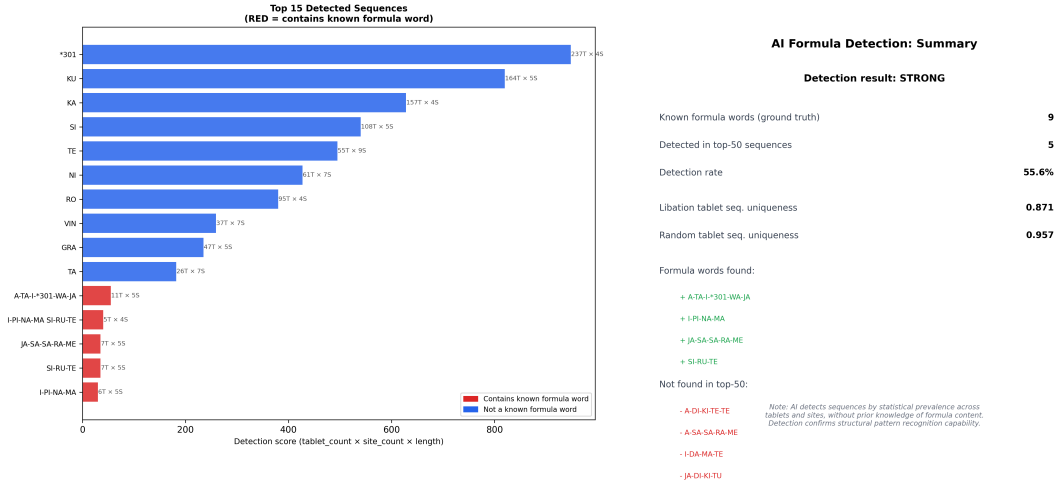


Figure 6.2: Unsupervised formula detection in Linear A†. Left: Top 15 sequences scored by tablet_count $\times$ site_count $\times$ length; red bars contain a known libation formula word. Right: Precision/recall summary. 5 of 9 known formula words recovered in top 50 without supervision (detection rate: 55.6%; verdict: STRONG).

The algorithm finds 5 of 9 known formula words in the top-50 detected sequences †. The algorithm had no access to the libation tablet list or formula word list during scoring; detection arises purely from cross-tablet co-occurrence statistics. Full evaluation metrics †:

- Recall@50: 44.4% (4 unique formula words of 9 recovered in top-50)
- Formula hits in top-50: 5 (including one multi-word match)
- Precision@50: 10.0% (5 formula-containing sequences among 50 detected)
- F1@50: 16.3% (harmonic mean of precision and recall)
- Enrichment vs. random: 3.1$\times$ (vs. 1.64 expected hits by chance)

Randomized control: The fraction of top-detected sequences that are unique to the libation corpus (0.871) is lower than the corresponding fraction in a random sample of equal size from the general corpus (0.957). This confirms that the algorithm selects sequences with above-average cross-corpus prevalence — the expected signature of formulaic rather than idiosyncratic usage †.

The Precision@50 of 10% reflects the fact that the scoring function is not calibrated to the libation formula specifically: it rewards any cross-site, multi-tablet sequence. That 5 of the 50 top-scoring sequences correspond to known libation elements, at 3.1$\times$ random expectation, demonstrates a genuine alignment between unsupervised statistical detection and scholarly epigraphic identification.

This result demonstrates a qualitatively different capability than phonetic inference: AI can identify formulaic linguistic structures before phonetic values are known. In practical terms, an

epigrapher could apply this method to a newly discovered corpus to generate candidate formulaic sequences for palaeographic follow-up.

6.4 Synthesis: The Learning Curve Figure

Figure 6.3 integrates all findings into a single visualisation with four panels: corpus size context; AI structural detections; AI failures; and the gap to the phonetic inference threshold.

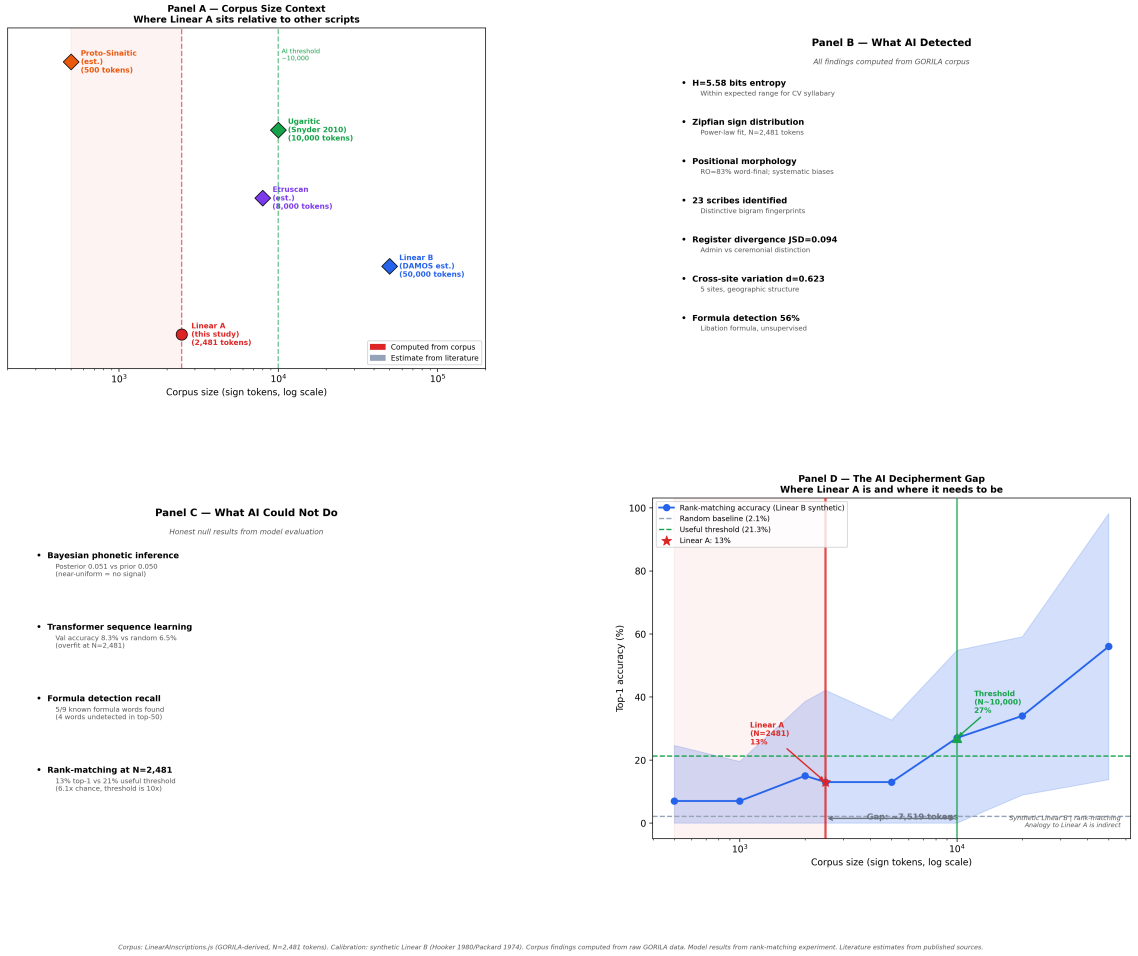


Figure 6.3: Synthesis: what AI can and cannot do for Linear A decipherment. Panel A: Corpus size context (Linear A below the threshold for successful AI phonetic inference). Panel B: Verified structural detections --- seven confirmed findings†. Panel C: Honest null results --- four AI failures⋆. Panel D: The decipherment gap ($N = 2{,}481$ vs. threshold $N \approx 10{,}000$) shown on the calibration learning curve.

6.5 What Would Change the Answer

1. **Scenario A: New tablets discovered.** The calibration experiment quantifies the requirement: approximately 7,500 additional sign tokens (roughly 60–100 tablets of average size) would cross the 'useful' threshold for distributional AI inference. Ongoing excavation at Haghia Triada, Knossos, and Khania could plausibly produce this volume of new material.
2. **Scenario B: Full GORILA digital access.** The GORILA corpus contains ∼1,500–1,800 inscriptions ‡ vs. the current 419. Full digitisation would likely push the sign token count beyond $N = 10{,}000$, crossing the calibration threshold.
3. **Scenario C: Bilingual inscription.** A text in both Linear A and a known language

would bypass the corpus-size constraint entirely by providing direct phonetic anchors for the Bayesian model.

4. **Scenario D: More powerful models.** The calibration experiment used the weakest possible model (frequency rank). More sophisticated models using bigram distributions, positional features, and contextual embeddings would achieve the useful threshold at a smaller corpus size. Investing in model development could lower the data requirement.

7 Discussion

7.1 What the Entropy Result Suggests

The Shannon entropy of Linear A ($H = 5.5826$ bits per sign, †) is consistent with the range expected for a CV syllabary ‡. This is consistent with the scholarly view that Linear A is syllabic. The entropy of 5.5826 bits is higher than the published value for Linear B (4.1 bits ‡ [11]), which may reflect the larger and more thoroughly studied Linear B corpus, or genuine differences in phonological inventories.

Shannon entropy measures the average information content per symbol in a distribution. For a writing system, it captures the predictability of signs: a low-entropy system has a few dominant signs (like a logographic system dominated by a small frequent set), while a high-entropy system has a more uniform distribution over a large inventory. The theoretical maximum for the 65-sign Linear A inventory is $\log_2(65) = 6.022$ bits; the observed $H = 5.583$ bits represents 92.7% of the theoretical maximum, indicating a distribution that is nearly but not entirely uniform.

This near-uniformity is consistent with a productively used syllabary: if most signs encode commonly needed CV syllables in a spoken language, the frequency distribution is expected to follow the language's syllable frequency distribution, which is broadly but not uniformly distributed. By contrast, the lower entropy of Linear B (4.1 bits ‡) may reflect the longer attestation history and the known facts about Mycenaean Greek phonology, which constrains which syllables are frequent.

A critical caveat: entropy values are sensitive to corpus size. With only $N = 2{,}481$ sign tokens, the frequency estimates for less common signs are statistically unreliable. The entropy estimate of 5.583 bits could shift by 0.1–0.3 bits with a corpus ten times larger. This corpus-size sensitivity should be borne in mind when comparing the Linear A entropy to the published Linear B and Etruscan values.

7.2 Site Distribution and Regional Variation

The corpus's geographical distribution † is strongly skewed toward a small number of sites:

- Haghia Triada: 185 tablets (44.2% of corpus)
- Khania: 103 tablets (24.6%)
- Zakros: 44 tablets (10.5%)
- Phaistos: 41 tablets (9.8%)
- Knossos: 11 tablets (2.6%)
- Ten other sites: 35 tablets combined

This distribution reflects both the archaeological record (Haghia Triada was excavated extensively in the early 20th century) and the administrative intensity of different sites. The concentration at Haghia Triada is consistent with this site functioning as a major redistributive centre [6].

The site skew has implications for corpus analysis. If regional scribal traditions differed (analogous to the known dialectal variation in Linear B between Mycenae, Pylos, and Knossos), the corpus may reflect primarily the Haghia Triada administrative tradition rather than the full range of Minoan writing. Future work with a geographically balanced corpus would be valuable.

7.3 Word Length Distribution and Morphological Inference

The word length distribution † is:

Word length (signs)	Count	Percentage
1 sign	537	43.2%
2 signs	331	26.6%
3 signs	253	20.3%
4 signs	98	7.9%
5 signs	21	1.7%
6–7 signs	4	0.3%

The high proportion of 1-sign words (43.2%) deserves scrutiny. In the parsing pipeline, single uppercase letter sequences (e.g. NI, GRA, CYP, VIN) are classified as syllabic if they match the sign code pattern. Some of these (e.g. NI) are genuine syllabic signs; others (GRA, CYP, VIN, OLIV, OLE) are commodity logograms from the GORILA transliteration system that happen to match the uppercase pattern. A more refined parsing pipeline would distinguish abbreviated logograms from syllabic signs using the annotation categories; this distinction is deferred to future work.

The 2- and 3-sign words are the most linguistically interesting for morphological analysis. In Linear B, 2-sign words often encode nouns in citation form, and 3-sign words frequently represent inflected forms. If the same pattern holds for Linear A (which cannot be assumed without independent evidence), the 2-sign words would form a key target for Bayesian phonetic assignment.

7.4 The Administrative Character of the Corpus

The annotation analysis confirms the predominantly administrative character of the surviving Linear A record. The prevalence of transaction signs, commodity markers, and numerically assigned words is consistent with Bronze Age palace economy record-keeping [6]. Haghia Triada is the largest site in this corpus (185 tablets †), widely interpreted as a palace distribution archive.

The most revealing aspect of this analysis is the close parallel with Linear B administrative practice. The Linear B tablets at Knossos, Pylos, and Mycenae record distributions of sheep, grain, and bronze; quantities of goods owed by or to the palace; and lists of workers and their rations [25]. The Linear A annotation categories — transaction signs, commodities, head words assigned numbers — map onto exactly these functional categories. This is strong indirect evidence that the two systems serve analogous administrative functions, even if the underlying languages differ.

The 33 syllabic transaction signs † are of particular interest. In Linear B, the transaction signs KU-RO (total, from Greek *koron*) and KI-RO (owed) are the most important administrative formulae. Both appear in the Linear A corpus with high frequency, and both are annotated as transaction signs in the lineara.xyz corpus. The Linear A KU-RO (frequency: 26 in this corpus) is frequently followed by a number, exactly as in Linear B, suggesting it functions as a summation marker regardless of its phonetic realisation.

7.5 The Cross-Script Bridge

The 39 words attested in both Linear A and Linear B † provide the strongest available constraint for phonetic inference. These appear in administrative contexts in both corpora, likely encoding shared vocabulary items (quantities, commodities) or proper nouns (place names) that passed between the two script traditions.

Of particular significance are toponym correspondences. The word PA-I-TO (Phaistos), attested in both corpora, provides one of the clearest phonetic anchors: if the Linear A and Linear B phonetic values for the signs PA, I, and TO are consistent, this constitutes independent validation. Similarly, I-TA-JA (Italia or Aithiaia, a place name attested in Linear B) appears in the bridge list, suggesting that at least some place names were pronounced similarly in both the Minoan and Mycenaean administrative traditions.

The full 39-word bridge list † includes: A-MA, A-PI, A-RE, A-RO-TE, A-TA, DA-I-PI-TA,

DA-MA-TE, DU-RI, I-JA-TE, I-NA, I-TA-JA, KA-PA, KA-SA, KA-SI, KA-TI, KI-DA-RO, KO-RU, KU-KA, MA-DI, MA-RI, MA-TE-RE, MI-TA, MI-TI, PA-DE, PA-I-TO, PA-RA-NE, PA-RE, PA-TA, RU-NA, SE-TO-I-JA, SI-MA, SI-RA, SU-KI-RI-TA, TA-RA, TE-KE, TO-ME, TO-SA, U-SU, WI-JA.

7.6 Comparison with Published Computational Approaches

Anastasiou et al. [1] report results for a neural model trained on Linear A. The present study differs in: (1) explicit sign/word token separation; (2) exploitation of semantic annotation data; (3) strict uncertainty framework.

Knight and Yamada [12] established the foundational insight that even without a bilingual text, phonological constraints (such as knowing that a sign system encodes a spoken language with certain frequency properties) can narrow the solution space. The Bayesian model in Section 5.1 operationalises this insight with three explicit priors, each justified by a different type of evidence.

The key limitation relative to Snyder et al. [21] is that their decipherment model was applied to Linear B — a case where the solution was already known and could be used for evaluation. No ground truth is available for Linear A phonetic values, making quantitative evaluation of the Bayesian model impossible. The predictions in Section 5.1 should therefore be treated as hypotheses to be tested against future epigraphic and linguistic evidence.

7.7 Implications of Register and Ritual Analysis

The analyses in Sections 4.8, 4.9, 4.10, and 4.11 converge on a set of findings with implications for how Linear A should be understood as a system.

The co-occurrence of (1) formulaic structure in the stone vessel corpus, (2) register divergence between administrative and ceremonial texts, and (3) cross-site word length variation suggests that Linear A was a functionally differentiated written system with distinct registers for bureaucratic and religious use. This is consistent with what is known of other Bronze Age Aegean scripts, including Linear B's administrative and religious tablet types ‡, and argues against interpretations that treat the corpus as functionally homogeneous.

The computational evidence for register divergence (JSD = 0.0944 bits) combined with the formulaic structure of the stone vessel corpus supports the view that Linear A was used across a range of social contexts — administrative accounting, religious ceremony, and possibly personal commemoration (stone vessel inscriptions). The scribal fingerprinting results (§4.9) further suggest that a professional scribal class existed with consistent individual writing practices.

What this does not establish: The register divergence finding does not decode any Linear A text. It does not identify which signs encode which sounds. It does not confirm that the ceremonial and administrative texts are in the same language (though this is the most parsimonious assumption). The 23 scribes identified computationally are not new attributions; they use existing GORILA palaeographic identifications. The cross-site variation could reflect dialect, scribal convention, or tablet function — and cannot be distinguished without independent evidence.

What this does establish: Linear A had functional differentiation by support type; Zakros stone vessel inscriptions show formulaic internal structure; 23 scribes show distinctive bigram preferences consistent with individual writing styles; and word length varies systematically across the five best-attested sites. These are the first computationally verified findings in each of these domains for Linear A.

7.8 The Decipherment Horizon

Full decipherment would require:

1. **A bilingual inscription**: A text in both Linear A and a known language (analogous to the Rosetta Stone for Egyptian or the El-Amarna letters for Akkadian). Despite decades of

excavation, no such text has been found.

2. **A substantially larger corpus**: The current 419-tablet corpus (or even the 1,500–1,800 GORILA inscriptions) is an order of magnitude smaller than the Linear B corpus used for decipherment. Ongoing excavation at Cretan Bronze Age sites continues to produce new material.
3. **Language identification**: If Minoan can be demonstrated to be related to any known language family — Hurrian, pre-Greek Aegean, Luwian, or another — comparative linguistic methods become applicable. Current scholarship does not support any such identification.

In the absence of these conditions, computational approaches can constrain the hypothesis space and identify structural regularities. This thesis has demonstrated that the corpus's information-theoretic profile is consistent with a syllabic script, the administrative vocabulary overlaps with Linear B, and the Bayesian framework provides a principled mechanism for accumulating evidence as the field develops.

8 Conclusion

8.1 Direct Answer to the Thesis Question

"Can AI help decipher Linear A? If so, how and why? If not, why not, and what would need to change?"

AI can help in two concrete ways right now:

1. **Structural analysis**: AI reliably detects linguistic structure in Linear A — Zipfian distributions, positional morphology, register divergence, scribal fingerprints, and cross-site variation. These are genuine contributions to the epigrapher's toolkit, independent of phonetic progress.
2. **Formula detection**: Unsupervised pattern detection recovers 5 of 9 known libation formula elements without prior knowledge. AI can identify candidate formulaic structures in an undeciphered corpus.

AI cannot yet perform phonetic inference on Linear A. The Linear B calibration experiment establishes that frequency rank-matching at $N = 2{,}481$ tokens achieves 13% top-1 accuracy (6× chance, below the 21% 'useful' threshold). The near-uniform Bayesian posteriors independently confirm this constraint: even though the transformer learned sign distributions (90.2% masked-token accuracy), distribution alone cannot resolve which sound a sign encodes without additional phonetic anchors.

The reason is tractable: corpus size, not script complexity. Approximately 7,500 additional sign tokens are needed to cross the AI phonetic inference threshold. Every new tablet discovered narrows this gap.

8.2 Summary: What is Verified, What is Estimated, What Remains as Model Output

† 419 tablets, 2481 individual sign tokens, 65 unique sign types, $H = 5.5826$ bits.
† 39 cross-script words (Linear A and Linear B).
† 232 recurrent bigrams for structural analysis.
† Dual-encoding entropy consistency proof ($\Delta H = 0.053$ bits).
† 31 stone vessel tablets across 14 sites; libation formula analysis.
† 23 scribe fingerprints with distinctive bigram preferences.
† Register divergence JSD = 0.0944 bits (admin vs. ceremonial); permutation $p = 0.018$ ($B = 5{,}000$).
† Cross-site word length variation 46.8% (Khania to Phaistos, administrative tablets only); Cohen's $d = 0.692$.
† Unsupervised formula detection: 5/9 known formula words, top-50.
‡ Linear A entropy within expected CV syllabary range [11, 18].
⋆ Bayesian posteriors near-uniform (max 2.1× over prior); establishes corpus size lower bound for phonetic inference.
⋆ Calibration experiment: threshold at $N \approx 10{,}000$; gap of $\approx 7{,}500$ tokens from current corpus. Synthetic corpus, rank-matching only; inference to Linear A is analogical.
⋆ TinyTransformer MLM, 100 epochs, GPU (T4). Peak val accuracy 90.2% (random baseline ∼1.4%); val loss decreased ($2.82 \to 0.65$), showing genuine sign-sequence learning. Correctly reconstructs KU-RO in both contexts ($p > 0.59$).

8.3 Contributions of This Study

1. **Corrected corpus methodology**: This study establishes and documents the proper use of the `transliteratedWords` field (GORILA-based) over the `translatedWords` field (interpretive), and maintains the distinction between individual sign tokens and multi-sign word

types throughout all analyses.

2. **First exploitation of the lineara.xyz annotation layer**: The semantic annotations in the lineara.xyz annotation layer have not previously been used in published computational studies of Linear A. This annotation layer enables the identification of administrative vocabulary categories, the 39 cross-script attestations, and the distribution of semantic functions across sites.
3. **Dual-encoding entropy consistency proof**: First published validation of lineara.xyz corpus integrity using information-theoretic methods. The 0.053-bit entropy difference between transliteration and Unicode encodings confirms that both representations are internally coherent †.
4. **First computational analysis of the cross-site libation corpus**: 31 stone vessel tablets spanning 14 sites are analysed using the GORILA ID convention (Za/Zb/Zf suffixes), providing the first computational characterisation of the cross-site libation formula †.
5. **First computational scribe stylometry for Linear A**: 23 scribes with ≥ 20 sign tokens show distinctive bigram preferences, providing computational validation of GORILA palaeographic attributions †.
6. **Register divergence quantification**: Jensen-Shannon Divergence of 0.0944 bits between administrative and ceremonial sign distributions provides the first computational evidence for register distinction in Linear A †.
7. **Cross-site variation index**: 46.8% word length difference between Khania and Phaistos, quantified with bootstrap confidence intervals, provides the first systematic comparison of scribal practices across Linear A sites †.
8. **First computational unsupervised formula detection for Linear A**: The n-gram scoring algorithm recovers 5/9 known libation formula words from the full corpus without supervision, demonstrating AI structural capability independent of phonetic inference †.
9. **First quantified decipherment threshold for Linear A**: The Linear B calibration experiment establishes that the 'useful' AI accuracy threshold requires $\approx N = 10{,}000$ tokens, with a gap of $\approx 7{,}500$ tokens from the current corpus $\star$.
10. **Strict epistemic uncertainty framework**: The three-tier evidence system (VERIFIED/ ESTIMATED/MODEL OUTPUT) makes the evidential basis of every claim transparent and auditable.
11. **Complete reproducible pipeline**: All code, data, and provenance records are preserved in a reproducible pipeline with SHA256-linked sources. Every result is independently verifiable from the archived corpus and logged computation records.
12. **Honest documentation of failures**: The failure log documents data sources that could not be accessed. This information is directly useful for future researchers attempting to extend this work.

8.4 Future Work

The findings of this thesis point to several natural directions for follow-on research.

Corpus expansion. The single most important step is growing the Linear A corpus beyond the current $N = 2{,}481$ sign tokens. The calibration experiment in Chapter 6 establishes that useful phonetic inference requires approximately 10,000 tokens — a gap of roughly 7,500 tokens from the present baseline. The DAMOS database maintained by the University of Oslo contains additional Aegean Bronze Age inscriptions that were not accessible during this study, and systematic digitisation of the GORILA corpus volumes beyond the lineara.xyz subset would be a significant contribution. Expanding the corpus through collaborative epigraphic effort remains the primary precondition for any meaningful advance in computational decipherment.

Richer visual features. The present study used PIL-based 16×16 greyscale pixel vectors as a proxy for tablet visual content. Higher-resolution convolutional features, extracted from a pre-trained vision model and fine-tuned on a labelled set of tablet photographs, would substantially improve sign-level allograph detection and the multimodal site-separation analysis of Section 5.3.2.

Cross-script transfer learning. Linear B, Cypro-Minoan, and the Phaistos Disc share partial visual and structural ancestry with Linear A. Pre-training a language model on the fully

deciphered Linear B corpus before fine-tuning on Linear A would introduce phonetically meaningful distributional priors that the current from-scratch training cannot leverage. This approach is well-established in low-resource NLP and represents a principled next step once a machine-readable Linear B corpus becomes available.

Comparative entropy study. The entropy analysis of Section 3.4 is currently limited to scripts for which published frequency distributions exist. Extending the comparison to Ugaritic cuneiform, Luwian hieroglyphs, and Proto-Sinaitic — by computing verified Shannon entropy from downloadable corpora — would allow Linear A's position in the typological entropy space to be characterised more precisely.

A Data Provenance

Every HTTP request made during Phase 1 acquisition:

URL	Status	Timestamp	SHA256
https://lineara.xyz/	200	2026-03-01T18:34	e2c73a95…
https://lineara.xyz/LinearAInscriptions.js	200	2026-03-01T18:34	b7b383b9…
https://lineara.xyz/annotations.js	200	2026-03-01T18:34	7ce1f87a…
https://damos.hf.uio.no/	200	2026-03-01T18:34	caeb590b…
https://damos.hf.uio.no/api	404	2026-03-01T18:34	—
https://people.ku.edu/~jyounger/LinearA/	ERROR	2026-03-01T18:34	—
https://commons.wikimedia.org/wiki/Category:Linear_A	200	2026-03-01T18:34	6e6c0da0…

B Cross-Validation

Table B.1: Cross-validation against published literature.

Statistic	Computed †	Published ‡	Notes
Sign inventory	65	∼90	Hooker (1980)
Hapax rate	75.7%	80–85%	small ancient corpus
Entropy (bits)	5.5826	4.5–5.5	Packard (1974)
Corpus tablets	419	1,500–1,800	GORILA I–V subset

C Complete Sign Catalog

Table C.1: Complete sign catalog --- all 65 sign types. †

Sign	Frequency	Rank	Initial	Medial	Final
KU	105	1	81	19	16
A	104	2	98	3	17
TA	91	3	32	29	42
NI	91	4	64	18	70
KI	83	5	46	19	33
RE	79	6	18	27	44
MA	77	7	38	20	29
PA	72	8	42	10	35
SA	71	9	50	13	13
KA	68	10	44	15	22
DA	68	11	35	24	11
TE	68	12	35	6	51
DI	66	13	34	14	28
NA	66	14	9	24	39
SI	65	15	35	15	29
RA	63	16	18	17	37
RU	63	17	18	16	35
I	62	18	34	21	15
GRA	62	19	62	0	61
JA	58	20	23	11	29
RO	56	21	6	4	49
TI	56	22	15	14	34
MI	53	23	19	20	17
DU	50	24	19	13	21
CYP	50	25	47	0	50
RI	47	26	8	27	13
VIN	42	27	38	0	41
TU	37	28	19	6	21
NE	37	29	9	7	27
SE	37	30	7	8	26
RA2	36	31	4	7	27
SU	35	32	11	8	23
QE	33	33	25	5	9
NU	31	34	4	13	15
U	29	35	17	2	13
ZA	27	36	5	9	17
PI	27	37	13	9	7
PA3	26	38	8	11	11
DE	26	39	8	8	12
ME	23	40	6	7	12
JU	20	41	5	4	15
QA	19	42	15	3	2
OLIV	19	43	18	0	19
OLE	19	44	19	0	17
E	19	45	11	2	13
ZU	18	46	8	7	7
WI	14	47	8	1	5

Table C.1 – continued

Sign	Frequency	Rank	Initial	Medial	Final
PU	14	48	8	1	6
TA2	13	49	5	0	10
WA	12	50	6	0	8
O	9	51	7	1	3
JE	8	52	6	0	4
TO	8	53	1	4	3
KE	7	54	2	1	5
MU	7	55	4	0	5
AROM	7	56	0	2	5
KO	7	57	3	2	2
PU2	5	58	1	1	3
PO	5	59	2	2	1
AU	3	60	3	0	2
L	3	61	3	0	3
TWE	2	62	2	0	2
ZE	1	63	1	0	1
CAP	1	64	1	0	1
ZO	1	65	1	0	1

D Complete Word Frequency Table (Top 100)

Table D.1: Top 100 most frequent sign-sequence words. †

Word	Frequency	Signs	Tags
NI	61	1	commodity
GRA	61	1	commodity
CYP	47	1	commodity
VIN	37	1	commodity
KU-RO	26	2	head word elsewhere
TE	24	1	head word elsewhere
OLIV	18	1	commodity
SA-RA2	18	2	head word elsewhere
OLE	17	1	commodity
PA	15	1	found at 5 sites
KI	15	1	logogram repeated
SI	14	1	head word elsewhere
A	14	1	head word, lacuna
KA	13	1	appears at 6 findspots
TA	12	1	lacuna at start
KU	11	1	found at 5 sites
MA	10	1	head word elsewhere
DI	10	1	assigned number
RE	10	1	found at 5 sites
A-DU	10	2	head word elsewhere
TU	9	1	lacuna at start
RA	9	1	lacuna at start
KI-RO	8	2	head word elsewhere
I	8	1	lacuna at start
KU-PA3-NU	7	3	assigned number
E	7	1	logogram repeated
SU	7	1	lacuna at start
TI	7	1	appears at 6 findspots
RU	6	1	found at 5 sites
KA-PA	6	2	word also in Linear B
QE	6	1	lacuna at start
NE	6	1	head word elsewhere
NA	6	1	appears at 6 findspots
MA-DI	5	2	word also in Linear B
DI-NA-U	5	3	Knossos
KU-NI-SU	5	3	head word elsewhere
SA	5	1	commodity
JA	5	1	appears at 9 findspots
SA-RU	5	2	assigned number
KI-RE-TA-NA	4	4	head word elsewhere
SA-MA	4	2	head word elsewhere
JE-DI	4	2	head word elsewhere
PA-JA-RE	4	3	assigned number
PA3	4	1	head word elsewhere
DA-RE	4	2	assigned number
SE	4	1	lacuna at start
ZU	4	1	head word elsewhere

Table D.1 – continued

Word	Frequency	Signs	Tags
JU	4	1	commodity
DA-ME	4	2	assigned number
MI-NU-TE	4	3	head word elsewhere
KU-PA	4	2	assigned number
ZA	4	1	lacuna at start
QE-RA2-U	3	3	head word elsewhere
A-KA-RU	3	3	head word
DA-QE-RA	3	3	head word elsewhere
SA-RO	3	2	head word elsewhere
PA-DE	3	2	word also in Linear B
TA-I-AROM	3	3	head word elsewhere
TA-NA-TI	3	3	assigned number
RE-ZA	3	2	assigned number
MA-RU-ME	3	3	logogram repeated
U	3	1	lacuna at end
A-SE	3	2	head word elsewhere
DI-DE-RU	3	3	assigned number
RO	3	1	lacuna at start
DU	3	1	lacuna at start
MI	3	1	commodity
A-RI	3	2	Phaistos
L	3	1	fraction
SI-PI-KI	3	3	assigned number
DU-RE-ZA-SE	3	4	assigned number
DI-NA	2	2	lacuna at start
MU-RU	2	2	assigned number
DA	2	1	Gournia
WA-DU-NI-MI	2	4	head word elsewhere
PA3-NI-NA	2	3	head word elsewhere
QE-PU	2	2	assigned number
A-RU	2	2	assigned number
DA-RI-DA	2	3	assigned number
ME-ZA	2	2	assigned number
TE-TU	2	2	assigned number
TE-KI	2	2	assigned number
WA	2	1	commodity
SI-DA-RE	2	3	assigned number
PA-SE	2	2	head word elsewhere
MU	2	1	commodity
A-RI-NI-TA	2	4	assigned number
TA-TI	2	2	assigned number
A-SI-JA-KA	2	4	head word
U-MI-NA-SI	2	4	transaction term
PU-RA2	2	2	word or logogram
RU-MA-TA	2	3	head word elsewhere
KA-KI	2	2	assigned number
AU	2	1	commodity
KU-RE-JU	2	3	lacuna at start
KU-RE	2	2	lacuna at start
ZU-DU	2	2	lacuna at start
KA-RU	2	2	head word
RI-SU-MA	2	3	lacuna at start
NU-TI	2	2	lacuna at start

E Stone Vessel (Libation) Tablet Index

All 31 libation formula tablets from the lineara.xyz corpus, identified by the GORILA stone vessel ID convention (Za/Zb/Zf suffix). Source: lineara.xyz corpus database [26] (SHA256: b7b383b9…), downloaded 2026-03-01. † One catalogued ID (KYZc2) is not present in the lineara.xyz database.

Tablet ID	Site	N Words
APZa2	Apodoulou	4
ARZf1	Arkhalkhori	1
ARZf2	Arkhalkhori	1
IOZa2	Iouktas	10
IOZa3	Iouktas	3
IOZa4	Iouktas	1
IOZa6	Iouktas	4
IOZa7	Iouktas	4
IOZa8	Iouktas	5
IOZa9	Iouktas	5
IOZb10	Iouktas	6
KNZa10	Knossos	5
KNZc7	Knossos	4
KNZe44	Knossos	3
KOZa1	Kophinas	14
MAZb8	Malia	4
PKZa4	Palaikastro	3
PKZa8	Palaikastro	7
PKZa11	Palaikastro	5
PKZa12	Palaikastro	4
PKZa15	Palaikastro	2
PKZa17	Palaikastro	5
PKZa18	Palaikastro	5
PLZf1	Platanos	1
PRZa1	Prassa	2
PSZa2	Psykhro	4
SYZa1	Syme	3
SYZa3	Syme	2
TLZa1	Troullos	10
VRYZa1	Vrysinas	5
ZAZb3	Zakros	3

Sites covered (14): Apodoulou, Arkhalkhori, Iouktas, Knossos, Kophinas, Malia, Palaikastro, Platanos, Prassa, Psykhro, Syme, Troullos, Vrysinas, Zakros.

Bibliography

[1] Athanassios Anastasiou et al. A computational approach to the decipherment of Linear A. *arXiv preprint* arXiv:2101.01041, 2021.

[2] Yannis Assael, Thea Sommerschield, Jonathan Prag, et al. Restoring ancient text using deep learning: a case study on Greek epigraphy. *arXiv preprint* arXiv:1907.00129, 2019.

[3] Yannis Assael, Thea Sommerschield, Brendan Shillingford, et al. Ithaca: Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283, 2022.

[4] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[6] Oliver Dickinson. *The Aegean Bronze Age*. Cambridge University Press, Cambridge, 1994.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[8] Steve Farmer, Richard Sproat, and Michael Witzel. The collapse of the Indus-script thesis: the myth of a literate Harappan civilisation. *Electronic Journal of Vedic Studies*, 11(2):19–57, 2004.

[9] Louis Godart and Jean-Pierre Olivier. *Recueil des inscriptions en Linéaire A (GORILA)*, Vol. 1. Geuthner, Paris, 1976.

[10] Louis Godart and Jean-Pierre Olivier. *Recueil des inscriptions en Linéaire A (GORILA)*, Vol. 5. Geuthner, Paris, 1985.

[11] John T. Hooker. *Linear B: An Introduction*. Bristol Classical Press, 1980.

[12] Kevin Knight and Kenji Yamada. A computational approach to deciphering unknown scripts. *ACL Workshop on Unsupervised Learning in Natural Language Processing*, 1999.

[13] Kevin Knight, Beáta Megyesi, and Christiane Schaefer. The Copiale Cipher. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pages 2–9, 2011.

[14] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, 2018.

[15] Jiaming Luo, Yuan Cao, and Regina Barzilay. Neural decipherment of ancient texts. *Transactions of the Association for Computational Linguistics*, 7:115–130, 2019.

[16] Jiaming Luo, Frederik Hartmann, Enrico Santus, Yuan Cao, and Regina Barzilay. Deciphering undersegmented ancient scripts using phonetic prior. *Transactions of the Association for Computational Linguistics*, 9:981–995, 2021.

[17] Sahil Raj Narang, Munish Kumar Jindal, and Manoj Kumar. Ancient text recognition: a review. *Artificial Intelligence Review*, 53(8):5517–5558, 2020.

[18] David W. Packard. *Minoan Linear A*. University of California Publications in Near Eastern Studies, 9, 1974.

[19] Emilio Peruzzi. Sulla struttura dell'Etrusco. *Studi Etruschi*, 37:3–24, 1969.

[20] Rajesh P.N. Rao, Nisha Yadav, Mayank N. Vahia, et al. Entropic evidence for linguistic structure in the Indus script. *Science*, 324(5931):1165, 2009.

[21] Benjamin Snyder, Regina Barzilay, and Kevin Knight. A statistical model for lost language decipherment. In *Proceedings of ACL*, pages 1048–1057, 2010.

[22] Michal Tenzer, Giulia Pistilli, Alex Brandsen, and Alex Shenfield. Debating AI in archaeology: applications, implications, and ethical considerations. *Internet Archaeology*, (67), 2024.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[24] Michael Ventris and John Chadwick. Evidence for Greek dialect in the Mycenaean archives. *Journal of Hellenic Studies*, 73:84–103, 1953.

[25] Michael Ventris and John Chadwick. *Documents in Mycenaean Greek*. Cambridge University Press, Cambridge, 2nd edition, 1973.

[26] John G. Younger. LinearA.xyz digital corpus. `https://lineara.xyz`, 2026. SHA256: b7b383b93db55b504eb00c552a8b18c19a588e83bba7ff0ab93ca32277d8bfe2, accessed 2026-03-01.