Statistical Glyph Clustering in Undeciphered Scripts

Author : Dr Kaelan R. Thorne **Date :** 2006

Title Page

University of Edinburgh — School of History, Classics and Archaeology

Statistical Glyph Clustering in Undeciphered Scripts

Doctoral Thesis submitted by Dr Kaelan Rhys Thorne

Date of Conferral: 2006

Abstract

This thesis presents a quantitative framework for analyzing undeciphered symbol systems using corpus construction, entropy analysis, and clustering techniques. Using several comparative corpora (Linear A, Phaistos Disc facsimiles, proto-Elamite fragments) and a newly assembled corpus of 24 recurring glyphs (hereafter “Ronwa corpus”), I evaluate whether symbol distributions are consistent with linguistic, formulaic, or decorative systems. Results indicate stable positional preferences, statistically significant bigram constraints, and cluster structures consistent with functional sub-alphabets.

Acknowledgements

I thank my supervisors at the University of Edinburgh, the British Museum staff for access to study collections, and peer reviewers who provided early feedback on the statistical pipeline.

Table of Contents

1. Introduction

Undeciphered scripts present two difficulties: small sample sizes and uncertain segmentation. This work proposes rigorous preprocessing and model selection that remain robust under sparse data.

1. Literature Review

- Frequency baselines (Zipf) and constraints in ancient scripts
  - Prior cluster studies of Linear A and proto-writing
  - Risks of overfitting under fragmentary evidence

1. Corpus Construction and Data Hygiene

- High-resolution vector tracing of glyphs
  - Tokenization rules; treatment of damaged signs
  - Ronwa corpus summary: 847 tokens across 47 tablets, 24 glyph classes
  - Inter-annotator agreement (Cohen’s κ = 0.82)

1. Methods

- Shannon entropy H = 4.1 for the Ronwa corpus; comparison across corpora
  - N-gram analysis: bigram and trigram adjacency matrices; PMI
  - Hidden Markov Models for positional likelihood; Viterbi paths on fragments
  - Bootstrapped confidence intervals; permutation tests

1. Clustering

- Feature vectors: position, adjacency counts, curvature, stroke complexity
  - k-means (k=3..8), silhouette and Davies–Bouldin scoring

- Hierarchical clustering (Ward, average linkage) with dendrogram stability
- Spectral clustering on normalized Laplacian of co-occurrence graph

1. Results

- Three stable clusters emerge (functionally: initiators, carriers, terminals)
  - Bigram constraints exclude random decorative usage ($p < 0.001$)
  - Motifs link month/temporal markers to calendrical positions

1. Discussion

- Cross-validation with Linear A positional statistics
  - Implications for bilingual correspondence (Knossos-Ronwa tablet)
  - Limitations from fragmentary state; recommendations for targeted sampling

1. Conclusions

The Ronwa corpus exhibits hallmark structure of a functional writing system with constrained sequencing and role-based glyph classes. Future work: larger corpora, bilingual alignment models, and symbol-phonetic testing.

References

Full bibliographic list (ancient epigraphy, quant linguistics, statistical learning).

Appendices

- A: Corpus schema (CSV fields)
  - B: Co-occurrence tables (top 100 bigrams)
  - C: Figure plates (entropy plots, dendrograms)
  - D: Annotation protocol
  - E: Code notes (reproducibility)

Contact

info@theronwaproject.com