Research Proposal

Bayesian Phonotactic Analysis and Historical-Comparative Reconstruction for the Isthmian (Epi-Olmec) Script

1. Research Context and Objectives

The Isthmian (Epi-Olmec) script constitutes one of the most significant unresolved problems in Mesoamerican historical linguistics. Despite several long inscriptions—notably La Mojarra Stela 1 (156 CE) and the Tuxtla Statuette (162 CE)—the linguistic affiliation of the script remains contested. While Justeson and Kaufman (1993) proposed a Mixe-Zoquean decipherment, subsequent criticism by Houston and Coe (2003) demonstrated that their readings failed when applied to new texts.

A fundamental challenge, articulated by Vonk (2020), is the *corpus limitation problem*. Vonk demonstrates that for morpho-phonographic writing systems, a structurally unique decipherment requires that qW (distinct signs / total attestations) satisfy qW ≪ 0.1. The Isthmian corpus, with approximately 160 distinct signs and 700 total attestations, yields qW ≈ 0.23—far exceeding this threshold.

To demonstrate this limitation, Vonk (2020) constructed a complete **Proto-Huastecan decipherment** of the Isthmian texts that is structurally as coherent as Justeson and Kaufman's Mixe-Zoquean proposal. This proves that traditional sign-value decipherment *cannot discriminate between language family hypotheses,* given the current corpus size—both Mixe-Zoquean and Huastecan readings can be made to fit the data.

The present project proposes a **complementary approach**: Bayesian comparison of distributional and phonotactic features that does not require proposed sign readings. This methodology evaluates structural properties of the script, segment lengths, positional entropy, and transition probabilities against typological expectations derived from comparative linguistics.

2. Viable Language Hypotheses

Geographic and historical evidence constrain the viable candidate language families to **two**:

2.1 Proto-Mixe-Zoquean

The Mixe-Zoquean language family is historically documented in the Isthmian region. Campbell and Kaufman (1976) proposed that the Olmec civilization spoke a Mixe-Zoquean language, based on loanword evidence in surrounding languages. The geographic distribution of modern Mixe-Zoquean languages exactly overlaps with the Isthmian inscription sites

(Veracruz, Tabasco, Chiapas). Phonological characteristics include (C)V(ʔ/h)(C) syllable structure with constrained word-finals (Wichmann 1995).

2.2 Proto-Huastecan (Mayan)

The Huastecan branch of Mayan represents a geographic outlier, currently spoken in northern Veracruz and San Luis Potosí. Vonk (2020) argues that if the Maya homeland was southeast of the Isthmus, early Huastecan speakers must have migrated through the region, potentially explaining the inscription distribution's northwest shift over time. Phonological characteristics include CVC syllable structure with ejectives and agglutinative morphology (Kaufman & Norman 1984).

2.3 Excluded: Otomanguean

Otomanguean languages are geographically implausible. The family is concentrated in the Oaxaca highlands and Central Mexico, *not* the Gulf Coast, where inscriptions are found. No archaeological or historical evidence supports Otomanguean speakers in the Isthmian region, and no scholar has proposed this affiliation.

3. Methodological Framework

The proposed methodology employs Bayesian model comparison to evaluate competing language family hypotheses:

3.1 Feature Extraction

Distributional features are extracted from the Isthmian corpus without assuming sign values:

- **Segment statistics:** Mean length, variance, coefficient of variation (using MS20 as boundary marker per Macri 2017b)
- **Positional entropy:** Uncertainty at segment-initial and segment-final positions
- **Bigram entropy:** Predictability of sign transitions
- **Frequency concentration:** Distribution of sign frequencies (Zipf characteristics)

3.2 Phonotactic Priors

For each candidate family, prior distributions are derived from comparative linguistic reconstructions:

- **Proto-Mixe-Zoquean:** Wichmann (1995), Kaufman & Justeson (2001)
- **Proto-Huastecan:** Kaufman & Norman (1984), Kaufman (2003)

Critical note: Proper priors require corpus statistics from reconstructed proto-lexicons. A key objective of the postdoctoral research is to compile such corpora and derive empirically-grounded priors.

3.3 Bayesian Evidence Computation

Model comparison uses marginal likelihood (Bayesian evidence) computed via Normal-Normal conjugacy for continuous features. Bayes factors quantify relative support for each hypothesis, following Kass & Raftery's (1995) interpretation guidelines.

4. Preliminary Results (Exploratory)

A pilot implementation with *hypothetical priors* (derived from typological generalizations, not actual corpus statistics) yields the following exploratory results:

Language Family	Log Evidence	Rank
Proto-Mixe-Zoquean	−44.41	**1**
Proto-Huastecan	−49.61	2

Feature-by-feature comparison:

Feature	Observed	MZ Prior	Hu Prior	Favors
Segment length	5.06	5.06	6.14	MZ
Final entropy	4.33	2.00	1.80	MZ
Bigram entropy	8.57	4.50	4.00	MZ

Interpretation: With hypothetical priors, Proto-Mixe-Zoquean shows a better fit to observed features. However, *these results are exploratory*—the priors are not derived from actual proto-language corpus statistics. The postdoctoral research would develop proper empirically-derived priors.

5. Research Questions

1. **Prior development:** What are the empirical distributional statistics of reconstructed Proto-Mixe-Zoquean and Proto-Huastecan lexicons, and how do they translate to script-level predictions?
2. **Discrimination power:** Can Bayesian phonotactic analysis discriminate between language family hypotheses given proper priors, or does Vonk's qW limitation extend to distributional methods?
3. **Methodological validation:** Can the methodology be validated against known scripts (e.g., Maya) before application to undeciphered systems?

6. Expected Outcomes

- Compiled Proto-Mixe-Zoquean and Proto-Huastecan lexical databases with distributional statistics
- Empirically-derived phonotactic priors for Bayesian model comparison
- Rigorous assessment of whether distributional methods can discriminate language family hypotheses

- Open-source computational tools applicable to other undeciphered scripts
- Peer-reviewed publications addressing methodology and results

7. Contribution to the Cologne Project

This proposal aligns with the VolkswagenStiftung-funded initiative "Exploring an unknown language in an unknown writing system: The Isthmus script." The project addresses a core methodological challenge: given that traditional sign-value decipherment cannot discriminate between language families (Vonk 2020), what alternative or complementary approaches are available?

The Bayesian phonotactic framework provides one such approach. Even if conclusive discrimination proves impossible (which would itself be a significant finding), the methodology contributes to the project's goal of rigorous hypothesis testing. The postdoctoral research would:

- Support Proto-Mixe-Zoquean or Proto-Huastecan reconstruction efforts
- Provide computational tools for distributional analysis
- Contribute phonetic expertise to phonological reconstruction

8. Timeline (36 months)

1. **Months 1–9:** Compile proto-lexicons; derive empirical priors; validate methodology on Maya
2. **Months 10–21:** Apply framework to Isthmian data; robustness testing; sensitivity analysis
3. **Months 22–30:** Cross-linguistic validation; publication drafting; conference presentations
4. **Months 31–36:** Final publications; tool documentation; project synthesis

9. The Critical Question: What Does MS20 Mark?

The validity of the Bayesian analysis depends entirely on what MS20 represents. The assumption chain underlying the analysis is:

1. **MS20 marks word boundaries** — Could alternatively be phrase boundaries, clause boundaries, or verse markers
2. **Signs before MS20 represent word-final sounds** — Could be logograms (whole words), not syllables
3. **Final entropy reflects word-final phonotactics** — Could reflect scribal conventions instead

4. **Low final entropy indicates CVC language structure** — This is the weakest link in the chain

To evaluate these assumptions, we compare Isthmian to the Maya script—the only fully deciphered Mesoamerican writing system.

10. Maya Script Comparison

10.1 Key Finding: Maya Script Has NO Word Boundary Marker

Scholarly sources confirm that the use of punctuation marks has not been documented for classic Maya hieroglyphic writing. Modern transcriptions compensate for the script's lack of explicit punctuation marks by adding spaces and other aids.

10.2 Comparison: Maya vs. Isthmian Scripts

Key differences between Maya and Isthmian scripts include: Maya has NO word boundary marker while Isthmian has MS20 (disputed function); Maya uses "glyph blocks" as visual word units while Isthmian uses columns; Maya has CVC word structure while Isthmian is unknown; both are logosyllabic scripts; Maya has ~800 signs while Isthmian has ~175.

10.3 How Maya Handles CVC Words Without Boundary Markers

Maya languages have a CVC structure (consonant-vowel-consonant), but the script uses CV syllables. The solution is the "echo vowel" system: for example, the word [kah] "fish fin" (CVC) is written as ka-ha (CV-CV), where the final vowel is silent. The GLYPH BLOCK itself marks word boundaries—no special marker needed.

11. Implications for MS20 Interpretation

11.1 MS20 is Unique to Isthmian

Since Maya (a related Mesoamerican script) has no boundary marker, MS20 in Isthmian is NOT inherited from a shared Maya-Isthmian tradition. It is either a unique Isthmian innovation OR serves a different purpose entirely.

11.2 Three Competing Interpretations

Three interpretations exist: (1) Grammatical suffix (Justeson and Kaufman): MS20 as -wu completive, implying morphology not phonotactics; (2) Phrase boundary (Macri, partial): MS20 as end-of-phrase marker, implying syntax not phonotactics; (3) Graphic marker (Macri, partial): MS20 as column divider, with no linguistic relevance.

11.3 Evidence from Macri (2017)

Key observations from Glyph Dwellers Report 52: MS20 is optional ("Its absence from the second sequence occurs presumably without a change in meaning"); MS20 is absent from Feldspar Mask (over 90 signs), arguing against grammatical suffix hypothesis; MS20 shows dual function with 8 of 22 La Mojarra columns ending with MS20, but MS20 also occurs within columns ~75% of the time.

12. Segment Length Analysis

If MS20 marks word boundaries, segment length should match word length predictions. Current data shows ~5 signs mean segment length (consistent with words), R52 calculation shows ~12.5 tokens (would suggest phrases), and Maya glyph blocks show 1-4 signs (typical word length). The observed mean of ~5 signs per segment is consistent with the word hypothesis.

13. Proposed Validation: Test on Maya Script First

Before trusting results on Isthmian, the methodology should be validated on a known script. The validation experiment involves: (1) Take a Maya inscription (e.g., Palenque Temple of Inscriptions); (2) Remove phonetic values—treat it as "undeciphered"; (3) Apply the same Bayesian analysis using glyph blocks as "segments"; (4) Check if it correctly identifies the language as Mayan.

If the method works, Maya glyph blocks should show patterns consistent with Ch'olan Mayan, and the Bayes factor should favor Mayan. If the method fails, the assumptions about script-phonotactics mapping are invalid, and the results on Isthmian would be meaningless.

14. Summary: MS20 and Maya Comparison

Key findings: Maya has NO boundary markers (MS20 is unique to Isthmian); segment length (~5 signs) is consistent with word-level units; MS20 is absent from Feldspar Mask (may be optional/scribal convention); MS20 occurs within columns (has linguistic, not just graphic, function).

14.1 Revised Conclusion

The comparison with Maya script suggests: MS20 is probably NOT a pure graphic marker (within-column occurrences suggest linguistic function); MS20 is probably NOT a mandatory grammatical suffix (absence from Feldspar Mask argues against this); MS20 most likely marks word or phrase boundaries (segment length supports word hypothesis); the analysis CAN proceed with appropriate caveats (but validation on Maya is essential).

14.2 Final Interpretation

The result "Proto-Mixe-Zoquean is better supported" should be interpreted as: "IF MS20 marks word boundaries (supported by segment length analysis but not proven), AND IF our typological priors are accurate, AND IF the script is primarily syllabic (like Maya), THEN the observed patterns are more consistent with Proto-Mixe-Zoquean than Proto-Huastecan."

The Maya comparison strengthens the word-boundary hypothesis but does not prove it. This remains a conditional conclusion, not an absolute finding.

Acknowledgement

In this research proposal, **Claude Anthropic** is used for code generation and content drafting. For the Isthmian script, **NotebookLM** is used for the literature review and key arguments to compare the Isthmian script with the Maya script. For the English language editor, **Copilot** and **Grammarly** are installed and as add-ined to **Microsoft Word** and the **Google Scholar Lab** for relevant articles on Isthmian and Maya scripts.

Pilot project on GitHub

https://github.com/chemvatho/isthmian-script

https://github.com/chemvatho/isthmian-script/blob/main/isthmian_bayesian_with_maya_comparison.ipynb

References

Campbell, L., & Kaufman, T. (1976). A linguistic look at the Olmecs. American Antiquity, 41(1), 80–89.

Houston, S. D., & Coe, M. D. (2003). Has Isthmian writing been deciphered? Mexicon, 25(6), 151–161.

Justeson, J. S., & Kaufman, T. S. (1993). A decipherment of Epi-Olmec hieroglyphic writing. Science, 259(5102), 1703–1711.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90(430), 773–795.

Kaufman, T. (2003). A preliminary Mayan etymological dictionary. FAMSI.

Kaufman, T., & Norman, W. (1984). An outline of Proto-Cholan phonology, morphology and vocabulary. In Phoneticism in Mayan hieroglyphic writing (pp. 77–166). Institute for Mesoamerican Studies.

Macri, M. J. (2017a). A sign catalog of the Isthmian script (Glyph Dwellers Report No. 51).

Macri, M. J. (2017b). An ending sign in the Isthmian script (Glyph Dwellers Report No. 52).

Vonk, T. (2020). Yet another "decipherment" of the Isthmian writing system. Unpublished manuscript.

Wichmann, S. (1995). The relationship among the Mixe–Zoquean languages. University of Utah Press.