OCR for Isthmus Script

A Computational Approach to Unknown Unknown Recognition

Research Proposal for the Project:

"Exploring an unknown language in an unknown writing system: The Isthmus script."

University of Cologne, Department of Linguistics

Prepared by:

Chem Vatho

12.12.2025

1. Summary

This proposal outlines a computational framework for developing an Optical Character Recognition (OCR) system tailored to the Isthmus script, one of the oldest and still undeciphered writing systems of Mesoamerica (ca. 400 BC – 600 AD). The proposed system employs a Convolutional Recurrent Neural Network (CRNN) architecture optimized for the unique challenges posed by this ancient syllabographic writing system.

The OCR system is designed to complement the historical-comparative linguistic analysis being conducted by Dr. Svenja Bonmann's research team at the University of Cologne, funded by the Volkswagen Foundation's "Pioneering Research" programme. By enabling systematic digitization, searchable indexing, and quantitative analysis of Isthmus inscriptions, this computational tool will facilitate the comparison of reconstructed linguistic material from Mixe-Zoque, Otomanguean, and Huastec language families with sequences identified in the script.

2. Background and Motivation

2.1 The Isthmus Script

The Isthmus script (also known as Epi-Olmec or La Mojarra script) represents one of the earliest known writing systems in Mesoamerica. The script is attested primarily through a small corpus of inscriptions, most notably the La Mojarra Stela 1—a four-ton limestone slab discovered in the Acula River near La Mojarra, Veracruz, Mexico, containing approximately 535 glyphs dating to the second century CE.

The writing system is characterized by:

- **Syllabographic structure:** Most signs represent CV (consonant-vowel) syllables, though some CVC syllables have been identified (e.g., *kak, pak, puk, yaj*)
- **Complex iconography:** Glyphs often incorporate anthropomorphic and zoomorphic elements with distinctive features such as headdresses, sound marks, and facial details
- **Limited corpus:** The small number of surviving inscriptions presents significant challenges for both traditional decipherment and machine learning approaches
- **No bilingual texts:** Unlike the Rosetta Stone for Egyptian hieroglyphs, no bilingual inscriptions have been discovered to aid decipherment

2.2 The Need for Computational Tools

The systematic analysis of Isthmus inscriptions requires tools that can:

1. Accurately identify and segment individual glyphs from inscriptions that may be weathered, damaged, or partially obscured

2. Classify glyphs into consistent categories to enable cross-referencing across different inscriptions
3. Generate machine-readable transcriptions that can be systematically compared with reconstructed phonological sequences from candidate language families
4. Support the iterative refinement of glyph interpretations as the decipherment progresses

3. Proposed Approach: CRNN Architecture

3.1 Architecture Overview

I propose implementing a Convolutional Recurrent Neural Network (CRNN) architecture, which has demonstrated state-of-the-art performance in scene text recognition and historical document OCR. The architecture combines the spatial feature extraction capabilities of Convolutional Neural Networks (CNNs) with the sequential modeling power of Recurrent Neural Networks (RNNs), making it particularly well-suited for recognizing sequences of syllabograms in the Isthmus script.

The complete architecture is illustrated in Figure 1 below:

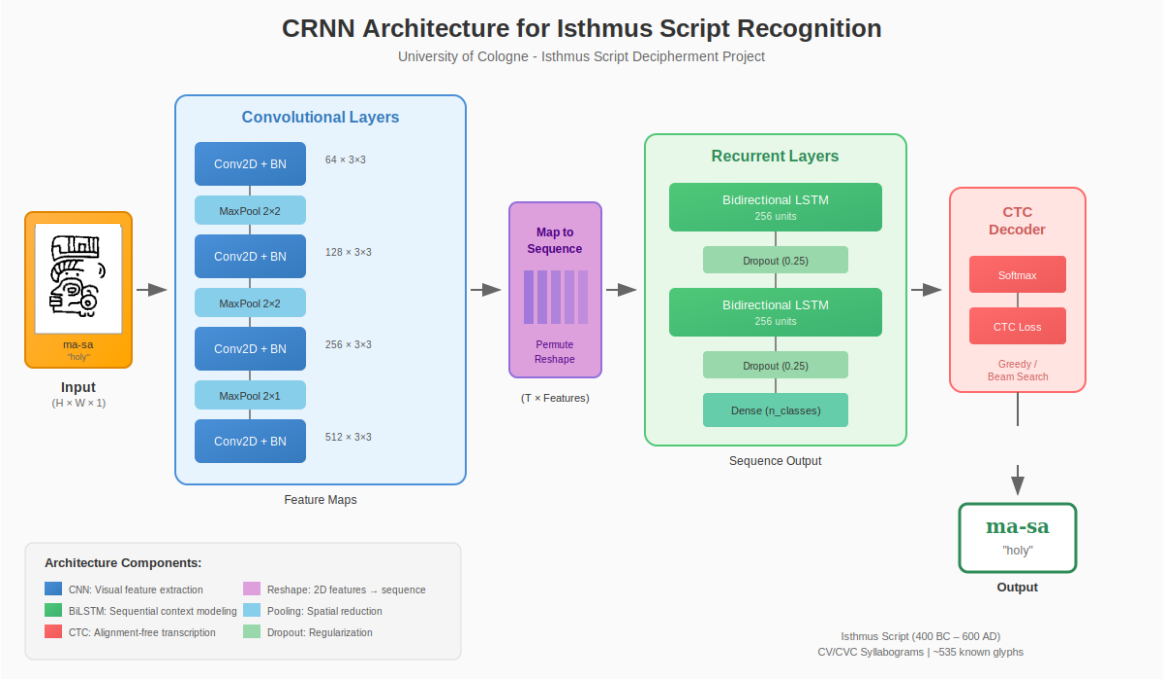

Figure 1: *CRNN Architecture for Isthmus Script Recognition. The pipeline processes glyph images through convolutional layers for feature extraction, reshapes features into sequences, models sequential dependencies with bidirectional LSTMs, and produces transcriptions via CTC decoding.*

3.2 Detailed Component Description

The architecture consists of four main processing stages, each designed to address specific challenges of Isthmus script recognition:

3.2.1 Convolutional Feature Extraction

The CNN backbone serves as a visual feature extractor, learning hierarchical representations of glyph imagery. The network consists of four convolutional blocks, each containing a Conv2D layer with 3×3 kernels, Batch Normalization for training stability, ReLU activation, and MaxPooling for spatial reduction. The filter depth progressively increases (64 → 128 → 256 → 512) to capture increasingly abstract features:

- **Early layers (64–128 filters):** Detect low-level features such as edges, curves, and line intersections that form the basic strokes of Isthmus glyphs
- **Middle layers (256 filters):** Capture mid-level patterns like the distinctive headdress combs, spiral eyes, and sound marks characteristic of the script
- **Deep layers (512 filters):** Encode high-level semantic features representing complete glyph components and their spatial relationships

The pooling strategy is asymmetric (2×2, 2×2, 2×1, 2×1) to preserve more information along the horizontal axis, which is critical for maintaining sequence length in the subsequent recurrent layers.

3.2.2 Map-to-Sequence Transformation

After convolutional processing, the 2D feature maps are transformed into a 1D sequence through a reshape operation. Each column of the feature map becomes a timestep in the sequence, with the feature depth forming the feature vector at each position. This transformation is critical because it:

- Enables variable-width input processing without requiring fixed-size glyph segmentation
- Preserves the left-to-right spatial ordering of visual features
- Creates a natural bridge between the spatial CNN representation and the temporal RNN processing

3.2.3 Bidirectional LSTM Layers

Two stacked Bidirectional Long Short-Term Memory (BiLSTM) layers model sequential dependencies in the feature sequence. Each BiLSTM contains 256 units in both forward and backward directions, producing a 512-dimensional output at each timestep. The bidirectional architecture is essential for Isthmus script recognition because:

- **Contextual disambiguation:** The phonetic value of a glyph may depend on both preceding and following glyphs, similar to how *ma* and *sa* combine to form *masa* ("holy")
- **Long-range dependencies:** LSTM gates enable the network to maintain relevant information across extended sequences
- **Robustness:** Dropout(0.25) between layers prevents overfitting on the limited training data

3.2.4 CTC Decoder

The Connectionist Temporal Classification (CTC) layer enables alignment-free training and inference. At each timestep, the network outputs a probability distribution over the vocabulary of syllabograms plus a special "blank" token. The CTC algorithm:

- **Eliminates segmentation requirements:** No need for pre-segmented character-level annotations, which would be impractical for the limited Isthmus corpus
- **Handles variable-length outputs:** Naturally accommodates CV and CVC syllables of different visual widths
- **Supports flexible decoding:** Both greedy decoding (fast) and beam search (higher accuracy) can be applied at inference time

3.3 Architecture Specifications

Component	Specification
Input	Grayscale glyph images (H × W × 1)
CNN Layers	4 × (Conv2D + BatchNorm + ReLU + MaxPool)
Filter Progression	64 → 128 → 256 → 512 (3×3 kernels)
Pooling Strategy	2×2, 2×2, 2×1, 2×1 (preserve width for sequence)
Recurrent Layers	2 × Bidirectional LSTM (256 units each)
Regularization	Dropout (0.25) after each LSTM layer
Output Layer	Dense (n_classes) + Softmax
Loss Function	CTC Loss
Decoding	Greedy / Beam Search (width=5)

4. Addressing Key Challenges

4.1 Data Scarcity

The limited corpus of approximately 535 known glyphs presents a significant challenge for deep learning approaches that typically require large training datasets. I propose addressing this through:

Extensive Data Augmentation: Implementing transformations that simulate the natural variation found in archaeological inscriptions:

- *Geometric transformations:* rotation (±15°), scaling (0.8–1.2×), perspective warping
- *Degradation simulation:* erosion, dilation, and Gaussian blur to mimic weathering
- *Noise injection:* salt-and-pepper noise, speckle patterns typical of stone inscriptions
- *Elastic deformation:* subtle warping to account for carving irregularities

Transfer Learning: Pre-training on related scripts (Maya hieroglyphs, other Mesoamerican writing systems) before fine-tuning on Isthmus data to leverage shared visual patterns.

Synthetic Data Generation: Creating synthetic training examples based on the morphological rules identified by the linguistic team, allowing hypothesis-driven expansion of the training set.

4.2 Syllabographic Complexity

The mixed CV and CVC syllabographic structure requires flexible sequence modeling:

- The CTC decoder naturally handles variable-length outputs without requiring explicit syllable boundary annotation
- Bidirectional LSTMs capture both forward and backward contextual dependencies essential for interpreting compound glyphs
- The vocabulary can be structured at the syllable level (CV, CVC) rather than individual phonemes, aligning with the script's fundamental units

4.3 Iterative Refinement

Given the ongoing nature of the decipherment effort, the OCR system must support iterative improvement:

- **Modular vocabulary:** The output layer can be retrained as new glyph identifications are confirmed
- **Confidence scoring:** Probability outputs enable identification of uncertain classifications for expert review
- **Active learning:** The system can prioritize uncertain samples for manual annotation, maximizing the value of expert time

5. Integration with Linguistic Analysis

The OCR system is designed to complement the historical-comparative linguistic methodology central to the project. Specifically, it will enable:

1. **Systematic Cataloguing:** Digital transcriptions of all known inscriptions in a consistent, searchable format
2. **Pattern Detection:** Automated identification of recurring glyph sequences that may correspond to morphological patterns in candidate languages
3. **Hypothesis Testing:** Rapid evaluation of proposed glyph-to-phoneme mappings against the corpus
4. **Cross-linguistic Comparison:** Structured output facilitates comparison with reconstructed phonological and morphological patterns from Mixe-Zoque, Otomanguean, and Huastec

6. Proposed Implementation Timeline

Phase	Activities
Months 1–3	Data collection and preprocessing; Initial glyph segmentation; Augmentation pipeline development
Months 4–6	CRNN architecture implementation; Transfer learning experiments; Initial training and validation
Months 7–9	Model optimization; Integration with linguistic analysis workflow; User interface development
Months 10–12	Iterative refinement based on decipherment progress; Documentation; Publication preparation

7. Expected Outcomes

1. A functional OCR system capable of transcribing Isthmus script inscriptions with measurable accuracy metrics
2. A searchable digital corpus of all known Isthmus inscriptions
3. Open-source tools and trained models for the broader Mesoamerican studies community
4. Peer-reviewed publication(s) documenting the methodology and findings
5. A framework adaptable to other undeciphered or low-resource scripts

8. Relevant Qualifications

This section should be personalized with your specific qualifications. Consider including:

- Experience with OCR systems and deep learning architectures (CRNN, TrOCR, Wav2Vec2)
- Background in computational linguistics and speech/text processing
- Experience working with low-resource languages and scripts
- Familiarity with historical-comparative linguistics methodology
- Track record of interdisciplinary collaboration

9. Conclusion

The development of an OCR system for the Isthmus script represents a unique opportunity to apply modern computational methods to one of archaeology's enduring puzzles. By combining deep learning techniques with the rigorous historical-comparative approach of Dr. Bonmann's project, this work has the potential to make a meaningful contribution to the decipherment effort while advancing the broader field of computational approaches to ancient scripts.

I am enthusiastic about the opportunity to contribute to this pioneering research and look forward to discussing how my computational expertise can support the project's goals.

10. Contact Information

Dr. Chem Vatho

chemvatho@gmail.com

PhD in Phonetics, University of Cologne

Even though I am not sure whether I will be accepted or not, I am still sharing this proposal for consideration. I believe it benefit to your project, why? I am from Cologne, and I want to go back to Cologne, my second home outside Cambodia.

References

Graves, A., Fernández, S., Gómez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning* (pp. 369–376). ACM. https://doi.org/10.1145/1143844.1143891

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Justeson, J. S., & Kaufman, T. (1993). A decipherment of Epi-Olmec hieroglyphic writing. *Science, 259*(5102), 1703–1711. https://doi.org/10.1126/science.259.5102.1703

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), Article 60. https://doi.org/10.1186/s40537-019-0197-0

Shi, B., Bai, X., & Yao, C. (2017). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(11), 2298–2304. https://doi.org/10.1109/TPAMI.2016.2646371