Research Project

Draft Research Report

BrahmiNet: A Neural Pipeline for Translating Ancient Sri Lankan Upper Brahmi into Meaningful Grammatically Correct Sinhala via Syntactic Analysis and Morphological Generation

by

K A Fernando (205023P)

Student	Signature	Date
K A Fernando	……..	……..

Supervisor	Signature	Date
Dr.(Ms.) U Ganegoda	……..	……..

Level 4 - Batch 20 (2025)

Faculty of Information Technology

University of Moratuwa

BrahmiNet: A Neural Pipeline for Translating Ancient Sri Lankan Upper Brahmi into Meaningful Grammatically Correct Sinhala via Syntactic Analysis and Morphological Generation

K A Fernando
Faculty of Information Technology
University of Moratuwa,
Sri Lanka.
fernandoka.20@uom.lk

***Abstract*—**

I. Introduction

Ancient Sri Lankan inscriptions, predominantly etched in stone, were composed in early forms of the Brahmi script, a writing system believed to have originated during the 3rd to 1st centuries BCE. According to the Sri Lankan Department of Archaeology, the majority of these inscriptions are written in Upper Brahmi, which served as the linguistic foundation for modern Sinhala. These inscriptions form a critical part of Sri Lanka's religious, linguistic, and historical heritage. However, interpreting these inscriptions poses significant challenges. The language used in Upper Brahmi inscriptions is often elliptical, morphologically sparse, and syntactically unstructured, making direct interpretation using modern Sinhala infeasible. Thus, the translation of such ancient texts into grammatically correct and semantically meaningful modern Sinhala is not only a linguistic problem but also a cultural necessity, supporting advancements in digital humanities, epigraphy, and heritage preservation.

This paper presents the fourth and final module of an ongoing research project aimed at developing a fully automated translation pipeline from Upper Brahmi inscriptions to Sinhala. While earlier modules address image processing, character recognition, and word segmentation, this module focuses on converting unordered, morphologically incomplete phrase-like translations derived from preceding stages into fluent and grammatically valid Sinhala sentences.

To achieve this, we propose a hybrid neural translation pipeline that combines syntactic analysis with neural reordering and morphological completion. The system begins by applying part-of-speech (POS) tagging and dependency parsing using a custom-trained Stanza model to extract the underlying syntactic structure of the input. Based on this structure, we implement a hybrid reordering strategy using rule-based SOV reordering for syntactically simple cases and a transformer-based neural model for more complex or ambiguous inputs. Finally, a morphological completion step driven by a neural model adds grammatical case markers and corrects verb morphology, producing fluent Sinhala outputs aligned with modern syntactic conventions.

II. Literature Review

Grammatical error detection and correction in Sinhala language processing present a unique challenge due to the language's rich morphological structure and flexible word order. To address these complexities, researchers have employed a combination of rule-based and machine learning techniques to ensure grammatical accuracy and adaptability across varying sentence structures.

One foundational approach utilizes Finite State Transducers (FST) and Context-Free Grammar (CFG) for grammatical correction [ref]. FSTs are primarily employed for morphological analysis, decomposing words into constituent morphemes and identifying grammatical markers. In parallel, CFG is used to verify syntactic correctness by ensuring that sentence structures conform to predefined grammatical rules. This method proves particularly effective in constrained linguistic contexts where formal grammar rules are clearly defined and can be strictly enforced.

To enhance adaptability, a hybrid grammar correction framework combining rule-based techniques with a decision tree classifier has been introduced [ref]. This approach benefits from the precision of linguistic rules while incorporating the flexibility of supervised learning models. By training on

annotated datasets, the decision tree classifier improves the detection and correction of context-sensitive errors, enabling the system to generalize beyond rigid grammatical constructs. The model demonstrates improved accuracy and robustness, particularly in handling variations common in informal or historical Sinhala usage.

More recently, advanced systems for automated spelling and grammar correction have integrated rule-based, traditional machine learning, and deep learning techniques, including Random Forests (RF), Recurrent Neural Networks (RNNs), and Support Vector Machines (SVMs) [ref]. These systems operate through a structured pipeline involving text preprocessing, syntactic pattern analysis, evaluation of subject-verb agreement, and correction based on verb tense and sentence structure predictions. The use of deep learning models allows for effective handling of complex and ambiguous constructions, especially in morphologically rich languages like Sinhala. Additionally, these approaches offer scalability and continual improvement potential, as performance can be enhanced with more data and further model tuning.

Collectively, these studies highlight a clear trend towards hybrid methodologies that leverage the strengths of both rule-based formalism and statistical or neural learning. Such approaches are particularly relevant to domains like epigraphy and historical linguistics, where language is often unstructured or grammatically incomplete.

III. Methodology

The input to this module consists of segmented Sinhala translations that are directly mapped from Brahmi inscriptions. These translations are typically semantically incoherent, lack grammatical structure, and are often morphologically incomplete, making them unintelligible to a modern Sinhala reader without expert interpretation. Therefore, the objective of this module is to transform these fragmented outputs into grammatically correct and semantically meaningful Sinhala sentences.
The proposed approach consists of three core stages:

1. Syntactic analysis using a custom-trained Stanza model
2. Hybrid sentence reordering combining rule-based logic and mT5-based neural methods
3. Morphological completion using an mT5 model fine-tuned for Sinhala grammar.

The overall system architecture is illustrated in Figure 1.
Due to the absence of publicly available digital corpora for Brahmi inscriptions and their corresponding Sinhala translations, we constructed a custom dataset from scratch. Primary resources include සිංහල සෙල්ලිපි වදන් අකාරාදිය by S. Ranawalla, Inscriptions of Ceylon by S. Paranavithana [ref], and ශිලා ලේකන සංග්‍රහය by Medauyangoda Vimalakeerthi Thero [ref].

A. Syntactic Analysis Using a Custom-Trained Stanza Model

This component uses the Stanza NLP toolkit developed by the Stanford NLP Group, known for its comprehensive pipeline of linguistic analysis tools [ref]. For this work, we customized the Stanza pipeline for Sinhala, utilizing the following processors: Tokenize, Lemma, POS, and Depparse. Since Stanza does not provide pretrained models for Sinhala, we trained models for each processor from scratch, adapting configurations to support non-standard input and label formats.
During processing, the input sentence is first tokenized, followed by lemmatization and POS tagging. Although Sinhala supports lemmatization, its utility in the Brahmi context is limited due to differences in inflection and vocabulary. The POS tagger plays a critical role, assigning universal POS tags to each token. We focused on the most relevant POS categories for inscriptional text: NOUN, N-DAT, PRON, VERB, TITLE, PLACE, ADJ, and CONJ. These annotations provide the foundation for the dependency parser to infer grammatical relations.
The dependency parser, trained on annotated Sinhala inscriptional data, identifies syntactic relationships such as subject (nsubj), object (obj), and root verbs, helping to construct the syntactic skeleton of the sentence. Given the inherently unordered nature of the direct Brahmi-to-Sinhala translations, this step is essential for identifying the underlying sentence structure required for reordering.

B. Hybrid Sentence Reordering

Modern Sinhala typically follows a Subject–Object–Verb (SOV) word order [ref]. To restructure the syntactically disordered sentences, we propose a hybrid reordering approach, leveraging both rule-based methods and a fine-tuned mT5 neural model, depending on sentence complexity.

1.1 Rule-Based Reordering

For short sentences (≤4 tokens) with unambiguous dependency structures (i.e., a single subject and verb), we employ a deterministic rule-based reordering strategy. This approach uses POS tags to classify tokens into syntactic roles:

- Subjects: PROP, TITLE, PRON (e.g., තිස්සා)
- Direct Objects: NOUN, PLACE (e.g., ලෙන), excluding subjects
- Indirect Objects: N-DAT (e.g., සන්සයා, if in dative case)
- Verbs: VERB (e.g., පුජ)

For example, consider the unordered sentence:

Input: තිස්සා පුජ ලෙන සන්සයා
Output (Reordered): තිස්සා ලෙන සන්සයා පුජ

This rule-based method is efficient and interpretable, producing structurally correct outputs. However, it lacks the flexibility to handle longer or syntactically ambiguous inputs and does not address morphological completion.

1.2 mT5-Based Neural Reordering

For complex sentences (i.e., >4 tokens or with multiple potential syntactic interpretations), we use a fine-tuned mT5 model (based on google/mt5-small, ~300M parameters) [ref]. This model is trained on a corpus of ~1,000–5,000 synthetic and manually curated sentence pairs, where each pair consists of:

Input: an unordered sentence annotated with POS tags and dependency labels
Output: a correctly ordered sentence in SOV format

The input format is structured as:
reorder: {unordered_tokens} | POS: {pos_tags} | Dep: {dep_labels} → output: {ordered_tokens}

Training involves shuffling grammatically correct Sinhala sentences, annotating them using the Stanza model, and supplementing with 100–200 manually validated examples. Optimization is done using the AdamW optimizer (learning rate = 5e-5), a StepLR scheduler (gamma = 0.9), gradient clipping (max norm = 1.0), and early stopping based on validation loss. During inference, the model uses beam search decoding and the structured prompt format to reorder input tokens while preserving SOV structure and syntactic dependencies.

C. Morphological Completion

The final stage uses a second fine-tuned mT5 model to perform morphological corrections, transforming reordered outputs into fully grammatical and semantically complete Sinhala sentences. Direct translations from Brahmi often lack morphological elements such as case particles, verb conjugations, and possessive suffixes, all of which are critical in Sinhala due to its morphologically rich structure.
The model is trained on a parallel corpus of grammatically correct sentences and their morphologically incomplete versions. Examples of transformations include:

- Possessive: තිස්ස → තිස්සගේ
- Dative: සන්ඝයාව → සන්ඝයාට
- Verb completion: පුජ → පුජා කරන ලදි

Training follows the same hyperparameters as the reordering model. The model learns to add appropriate morphological markers and verb forms based on context, leveraging mT5's multilingual pretraining on the mC4 corpus [ref], which includes Sinhala texts.

Examples of transformations include:

Input: (ordered, morphologically incomplete): තිස්සා ලෙන සන්ඝයා පුජ
output: තිස්සගේ ලෙන සන්ඝයාට පුජා කරන ලදි

In this transformation:

"තිස්සා" becomes "තිස්සගේ" to express possession (Thissa's).
"සන්ඝයා" becomes "සන්ඝයාට" indicating the dative case (to the Sangha).
"පුජ" is converted to "පුජා කරන ලදි", the passive past verb form appropriate for formal expression in inscriptions.

This final output reflects a complete, grammatically accurate Sinhala sentence derived from a flat, unordered input.

IV. Evaluation

V. Conclusion and Future Work

Acknowledgements

We would like to express our heartfelt appreciation to the University of Moratuwa for their kind assistance and funding, both of which were crucial to the success of our work. We would also like to thank the writers of the research papers that have been reviewed for their significant contributions to the field. Without the cooperative efforts of these outstanding institutions and people, this study would not have been feasible.

References

[1] "Comparative study of basic and hybrid filters for the reduction of noise in Estampages | Request PDF," in *ResearchGate*, doi: 10.1109/ICATIECE45860.2019.9063768.

[2] "(PDF) Denoising and Segmentation of Epigraphical Estampages by Multi Scale Template Matching and Connected Component Analysis," *ResearchGate*, doi: 10.1016/j.procs.2020.04.201.

[3] "(PDF) Retracted: Ancient Stone Inscription Image Denoising and Inpainting Methods Based on Deep Neural Networks," *ResearchGate*, Jan. 2025, doi: 10.1155/2024/9842712.