Machine Translation for a Critically Endangered and Non-Digitalized Language: The Case study of Vedda - Sinhala

Anonymous ACL submission

Abstract

The Vedda language of Sri Lanka is a critically endangered indigenous language with extremely limited written and digital resources. To date, no standardized Vedda–Sinhala bilingual lexicon or computational translation system has been available. This paper presents the first machine-readable Vedda–Sinhala lexical resource together with a statistical machine translation (SMT) prototype designed for extremely low-resource settings. We manually compiled and normalized a parallel corpus of 5,992 Vedda–Sinhala lexical pairs from nine published linguistic and ethnographic sources. Using NLTK, we implemented a cascading n-gram tagging baseline (Trigram → Bigram → Unigram → Default) and further extended it with a morphological post-processing module that performs Vedda suffix decomposition, root re-tagging, and Sinhala case-suffix recovery. Evaluation on a held-out test set of 226 sentences (975 words) shows that the enhanced system improves average sentence BLEU from 43.88 to 45.39, while reducing Word Error Rate from 0.6806 to 0.6634 and Character Error Rate from 0.5333 to 0.4869. The proposed morphological layer also lowers the unknown-word rate from 32.4% to 28.9%, recovering 34 previously untranslated tokens. We further compare our system against three general-purpose large language models, including ChatGPT, DeepSeek, and Gemini, on the same benchmark, where Gemini achieves the strongest aggregate performance. Despite this, our deterministic approach remains competitive while providing transparency, reproducibility, and offline usability. This work contributes the first structured Vedda–Sinhala lexical dataset and establishes a reproducible baseline for low-resource machine translation and indigenous language preservation research.

1 Introduction

The Vedda are an indigenous community of Sri Lanka whose linguistic heritage predates later Indo-Aryan and Dravidian settlements on the island. Historical census records from 1881 estimated the Vedda population at approximately 2,300 individuals; however, by the mid-twentieth century, this number had declined to fewer than one thousand, and today the language is classified as critically endangered (Surendra et al., 2024). Contemporary Vedda survives primarily through limited lexical items, kinship terminology, ritual expressions, and oral traditions rather than as a fully functional language used in everyday communication (Weerasekara, 2020a). As a result, both the linguistic heritage and cultural knowledge embedded within the language face an increasing risk of permanent loss.

From a computational perspective, Vedda represents an extremely low-resource language scenario (Weerasekara, 2020b). Unlike widely studied South Asian languages such as Sinhala or Tamil, Vedda lacks large-scale digitized corpora, standardized orthography, publicly available bilingual dictionaries, and annotated linguistic datasets. Existing lexical material remains scattered across ethnographic publications, anthropological records, and field notes, making systematic computational processing particularly difficult (Ananda, 2020). Furthermore, contemporary neural machine translation (NMT) systems typically require tens of thousands of parallel sentence pairs to achieve reliable generalization and translation quality (Ranathunga et al., 2023), whereas the available Vedda linguistic data are highly fragmented and limited in scale. These constraints make conventional data-intensive neural approaches impractical and motivate the exploration of lightweight, interpretable statistical methods tailored for endangered language preservation.

In response to these challenges, this paper introduces the first machine-readable Vedda–Sinhala lexical resource and a statistical machine translation prototype specifically designed for extremely

low-resource conditions. We manually digitized and normalized 5,992 Vedda–Sinhala lexical pairs collected from nine published linguistic and ethnographic sources, thereby creating the first structured parallel dataset for the language pair. Building on this resource, we implemented a cascading n-gram tagging architecture using NLTK, consisting of Trigram, Bigram, Unigram, and Default backoff taggers, serving as a deterministic baseline system (Version 1). We further extend this baseline with a morphology-aware post-processing module (Version 2) that performs Vedda suffix decomposition, root re-tagging, and Sinhala case-suffix recovery in order to better handle morphologically variant and previously unseen forms. Both systems are evaluated on a held-out test set of 226 sentences under multiple automatic evaluation metrics, including BLEU, Word Error Rate (WER), Character Error Rate (CER), and unknown-word recovery. In addition, we benchmark our approach against three general-purpose large language models, ChatGPT (GPT 4o) (Achiam et al., 2023), DeepSeek-v3 (Liu et al., 2024), and Gemini 2.5 pro (Comanici et al., 2025), using the same evaluation framework. To support accessibility and future research use, the complete prototype is deployed within a three-tier web architecture consisting of a Next.js frontend, a FastAPI backend, and an NLTK-based translation engine.

The main contributions of this work are as follows:

- We construct the first machine-readable Vedda-Sinhala parallel lexical corpus comprising 5,992 manually digitized and normalized lexical pairs collected from multiple published sources.
- We develop a deterministic SMT-based translation framework for Vedda–Sinhala translation using a cascading n-gram architecture and a morphology-aware post-processing module tailored to Vedda suffix patterns.
- We provide the first comparative benchmark for Vedda translation by evaluating our system against modern general-purpose LLMs under identical experimental conditions and evaluation metrics.

The remainder of this paper is organized as follows. Section 2 reviews related work on low-resource and endangered-language machine translation. Section 3 describes the corpus construction and data normalization process. Section 4 presents the proposed translation architecture and morphological enhancement module. Section 5 details the evaluation methodology and the disucssion. Finally, Section 6 concludes the paper and outlines directions for future research.

2 Related Works

2.1 The Vedda Language

The linguistic classification of Vedda remains a subject of scholarly debate. Early linguistic and colonial studies characterized Vedda as a divergent dialect of Sinhala rather than an independent language. Later, (Dharmadasa, 1974) proposed that prolonged contact between Vedda-speaking communities and Sinhala speakers led to the emergence of a creolized linguistic system, consisting of a Vedda substrate significantly influenced by Sinhala lexical and grammatical structures. More recent linguistic studies generally describe Vedda as a heavily Sinhala-influenced mixed language that nevertheless preserves a number of distinctive linguistic features, including unique discourse particles such as pojja and gajja, as well as specialized vocabulary related to flora, fauna, kinship systems, hunting practices, and ritual life (Dharmadasa, 1974; Thimbiripola, 2022). Regardless of its precise classification, the language represents a distinct and culturally significant linguistic variety that remains severely under-documented in digital form. Table 1 illustrates a few sentence-level and word-level translations related to Vedda-Sinhala.

2.2 Low-Resource Machine Translation

Research in low-resource NLP commonly relies on structured lexical resources, rule-based methods, or lightweight statistical models before transitioning to large-scale neural architectures (Ranathunga et al., 2023; Haddow et al., 2022). In severely under-resourced settings, where annotated corpora and parallel datasets are limited or unavailable, interpretable and extensible systems are often preferred over highly data-dependent neural approaches (Hedderich et al., 2021). Such methods offer practical advantages for the preservation of endangered languages, including transparency, reproducibility, and the ability to incrementally incorporate newly collected linguistic data.

Previous computational and documentation efforts involving minority languages in Sri Lanka further demonstrate the value of lexicon-driven

Vedda Language	Sinhala Translation	English Translation
තොපං ඒ ඇත්තන්ට කතා දමනවාද	ඔබ ඒ අයට කතා කරනවද	Do you talk to them.
කැකුලියෙකුට හිතලාපු කැකුලෙක් මස්පොජ්ජ ඇන්න මංගව්චනවා	තරුණියකට ආලය කළ තරුණයෙක් මස් රැගෙන යනවා	A young man who is dating a young woman carries meat.
පරම්පරාපොජ්ජ කුයිනිගොල්ලන්නෙ	පරම්පරාව කුවේණිගෙ	The generation of Kuveni
එළිපොජ්ජ මන්දෝකරන යමෙක	විදුලි පන්දම	Torch
ගිනිකඳයා පටාරිනවා	වෙඩි තියනවා	Shot

Table 1: Example of Vedda-to-Sinhala sentence pairs used for training.

methodologies. For example, the documentation initiative for Sri Lankan Portuguese Creole employed structured lexical and linguistic resources to support preservation and analysis despite severe data scarcity (Cardoso et al., 2019). These studies indicate that meaningful computational support for endangered languages can be achieved even without large digital corpora. To the best of our knowledge, no prior computational translation system, machine-readable corpus, or standardized digital lexical resource currently exists for the Vedda–Sinhala language pair. Consequently, direct comparison with earlier Vedda-specific translation systems is not possible, highlighting both the novelty and necessity of the present work.

3 Dataset Construction

Two datasets were prepared for this study: a training corpus and a held-out evaluation set. The primary training resource contains 5,103 entries representing 5,992 Vedda–Sinhala lexical pairs compiled from nine published sources, including six English-language anthropological and ethnographic studies and Sinhala-language publications (See Appendix A). Due to the absence of existing alignment tools or standardized bilingual resources for the Vedda–Sinhala language pair, all lexical mappings were manually aligned and verified. Each entry was stored in a backslash-delimited format to preserve direct token-level correspondence between source and target forms.

Several preprocessing procedures were necessary to ensure corpus consistency and reliable downstream processing. First, Unicode NFC normalization was applied to standardize equivalent character sequences across the corpus. In addition, zero-width Unicode characters (U+200B, U+200C, and U+200D) were removed to prevent silent token mismatches during lookup and evaluation. Second, corpus parsing was designed to split entries only on sequences of two or more spaces or tab characters rather than on individual whitespace tokens. This step was necessary because many Sinhala translations appear as multi-word expressions; naive whitespace tokenization would therefore incorrectly fragment aligned translation pairs and corrupt the lexical mappings.

Exploratory analysis of the dataset revealed two characteristics that significantly influenced the system design. First, lexical ambiguity was common: approximately 23.4% of unique Vedda tokens had multiple Sinhala translations, indicating that a simple one-to-one dictionary lookup would be insufficient for reliable translation. Second, vocabulary sparsity was severe: 51.8% of the tokens appearing in the held-out evaluation set were unseen during training. This high out-of-vocabulary (OOV) rate established a strong upper bound on token-level translation performance and directly motivated the morphology-aware extension introduced in Section 4.2.

The evaluation dataset initially consisted of 250 sentence pairs collected for testing purposes through Vedda people-based conversations. After removing empty, duplicated, or malformed entries, 226 valid Vedda–Sinhala sentence pairs were retained for evaluation. The resulting test set contains 975 Vedda tokens, with an average sentence length of 4.31 Vedda tokens and 3.86 Sinhala reference tokens per sentence. This held-out dataset was used consistently across all experiments, including comparisons with external large language models.

4 Methodology

4.1 Cascading N-gram Tagger

The baseline translation engine is implemented as a cascading n-gram tagging architecture using the

Natural Language Toolkit (NLTK). During training, each sentence is represented as a sequence of (Vedda token, Sinhala translation) pairs, enabling the model to learn contextual lexical mappings between the source and target languages. The architecture consists of four hierarchical taggers connected through a backoff strategy: a TrigramTagger serves as the primary model, followed by a BigramTagger, a UnigramTagger, and finally a DefaultTagger. The trigram component is queried first because it can exploit a broader local context by considering the current token together with the two preceding tokens. If a trigram pattern is unseen during inference, the system automatically backs off to the bigram level, then to the unigram level, and finally to the default tagger if no matching context is found.

The DefaultTagger assigns the placeholder label NNN to unresolved tokens. This design ensures that the system always produces a complete translation output, even when encountering previously unseen or out-of-vocabulary words. At the same time, unresolved tokens remain explicitly identifiable, allowing the system to transparently indicate translation uncertainty rather than silently omitting unknown items. Once training is completed, the fully trained model is serialized into a pickle file for persistent storage. During deployment, the model is loaded once at server startup, thereby separating the training and inference lifecycles and reducing runtime overhead during translation requests.

4.2 Morphological Post-Processing

Error analysis of the baseline system (V1) revealed that many tokens labelled as NNN were not entirely unknown lexical items, but rather morphological variants of known Vedda roots. This phenomenon was particularly common among forms containing the Vedda nominalizing suffix පොජ්ජ and its case-inflected derivatives. To address this limitation, the version 2 extends the baseline architecture with a rule-based morphological post-processing layer that is triggered only when the n-gram tagger fails to produce a valid translation.

The post-processing component is implemented through the handling of unknown word routine, which applies a staged recovery pipeline designed to maximize lexical resolution for out-of-vocabulary (OOV) forms. The recovery process proceeds through several ordered steps. First, the system performs an extra-vocabulary lookup targeting frequently occurring kinship terms and manually identified verbal forms from corpus analysis. Second, tokens already valid in Modern Sinhala are passed through unchanged using a predefined Sinhala pass through lexicon. Third, a spelling normalization stage is applied to reduce orthographic inconsistencies present in the source material.

If these stages fail, the system applies Vedda-specific morphological decomposition using a longest-match suffix stripping strategy. Common suffix variants derived from Morphological Post-Processing පොජ්ජ are mapped to their corresponding Sinhala grammatical forms (e.g., පොජ්ජෙන් → න්, පොජ්ජට → ට, and පොජ්ජේ → වල). After suffix removal, the recovered root form is re-submitted to the n-gram translation pipeline for re-tagging. The system additionally performs Sinhala case-suffix stripping for common grammatical endings such as the dative (ට), possessive (ගේ), and ablative (ජෙන්) markers in order to improve root-level matching.

As a final fallback mechanism, unresolved tokens are emitted unchanged rather than discarded, preserving partial Vedda–Sinhala cognates and maintaining translation continuity. An additional hal kirīma normalization step is incorporated to standardize consonant-final forms during lookup. For example, forms such as පේන්න are normalized to the base form පේන prior to vocabulary matching. Collectively, these rule-based recovery procedures improve lexical coverage while maintaining the interpretability and deterministic behaviour of the overall translation framework.

For instance, the dative form ගම්පොජ්ජට (to village) is segmented into ගම් + පොජ්ජ + ට, allowing the system to recover the root ගම while identifying ට as the dative marker. Similarly, the possessive form ගමපොජ්ජේ (in the village) is decomposed into ගම් + පොජ්ජේ, enabling recovery of the base noun ගමේ, ගම්වල with preservation of the possessive marker. In the ablative form ගම්පොජ්ජජෙන් (from the village), the suffix sequence පොජ්ජජෙන් is stripped to recover the root ගමෙන්, while isolating ජෙන් as the ablative marker.

4.3 System Architecture

The prototype is implemented as a three-tier architecture that cleanly separates the presentation layer, application logic, and model components. The presentation layer is built using a Next.js frontend, which accepts Vedda input and displays structured

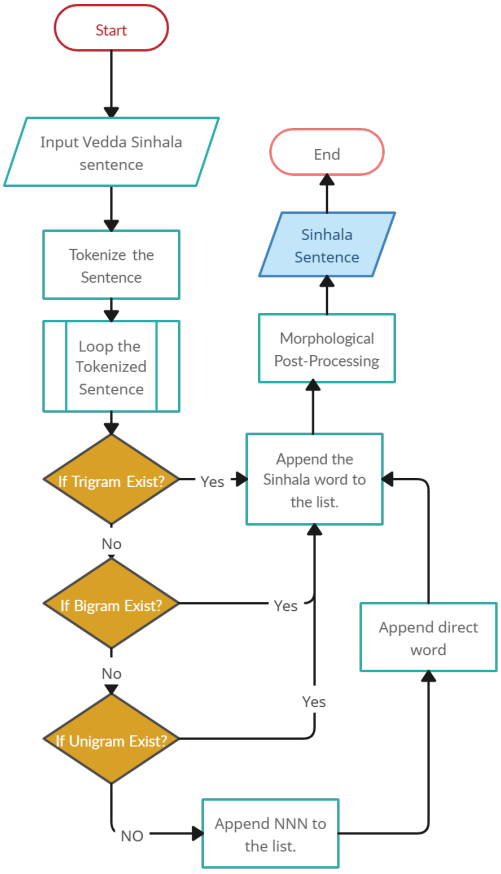

Figure 1: Data flow of the proposed Translator

outputs. These outputs include the full Sinhala translation, a word-by-word alignment with status annotations (e.g., ok, morphological, unknown), and a separate list of unresolved tokens to support error analysis and interpretability. The data flow of the final model is presented in Figure 1.

The model tier contains the serialized trigram-based tagging model along with the associated lexical resources used for translation. This strict separation of concerns enables modular development and maintenance of the system. In particular, it allows new components, such as the morphological post-processing module, to be integrated into the pipeline without requiring modifications to the frontend or backend service logic, thereby improving extensibility and system robustness. An accessible version of the system will be released with the camera-ready version of the paper.

5 Results and Discussion

5.1 Setup

Both systems were evaluated on the same held-out test set consisting of 226 sentences (975 Vedda tokens). Sentence-level BLEU scores were computed using add-one smoothing (Chen and Cherry, 2014; Papineni et al., 2002) and averaged across individual sentences rather than computed at the corpus level. This choice is motivated by the relatively small size of the evaluation set, where sentence-level averaging ensures equal contribution from each sentence and reduces sensitivity to sentence-length variation.

Word Error Rate (WER) and Character Error Rate (CER) were computed using standard edit-distance definitions. For CER, all whitespace characters were removed prior to evaluation to ensure that scoring reflects character-level translation accuracy rather than token segmentation artifacts. In addition, we report Unknown-Word Rate (UWR), defined as the proportion of input tokens for which the trigram-based tagging pipeline outputs the placeholder label NNN, indicating failure to resolve the token at any stage of the backoff chain.

Importantly, both system variants share an identical training procedure, the underlying trigram-based tagging model, and the same lexical resources. The only difference lies in the presence or absence of the Morphological Post-Processing module, which is applied after the baseline n-gram tagging process to recover unresolved tokens. This controlled experimental setup ensures that any observed performance differences can be attributed solely to the effect of morphological enhancement.

5.2 Results of Translation system

Table 2 presents the main evaluation results. The system with Morphological Post-Processing consistently outperforms the system without morphological enhancement across all evaluation metrics. BLEU increases from 43.88 to 45.39, corresponding to an absolute gain of 1.51 and a relative improvement of approximately 3.4%. Word Error Rate (WER) shows a modest reduction of about 2.5% relative, while Character Error Rate (CER) decreases more substantially by approximately 8.7% relative. In addition, the Unknown-Word Rate (UWR) drops by 3.5 percentage points (approximately 10.8% relative reduction), indicating improved lexical coverage.

Table 2: Evaluation results on 226 held-out sentences comparing the baseline system (V1) with the system enhanced by morphological post-processing (V2).

Metric	V1	V2 (+Morph)	Δ
Avg Sent. BLEU	43.88	45.39	+1.51
Avg WER	0.6806	0.6634	-0.017
Avg CER	0.5333	0.4869	-0.046
Unknown Word (%)	32.4	28.9	-3.5
Recovered tokens	0	34	+34

Table 3: Five-system benchmark on the same 226-sentence held-out test set. The Unknown-Word Rate (UWR) is defined only for deterministic systems.

System	BLEU	WER	CER	UWR	Det.
V1	43.88	0.681	0.533	32.4	✓
V2	45.39	0.663	0.487	28.9	✓
ChatGPT	44.81	0.614	0.444	–	×
DeepSeek	44.16	0.647	0.494	–	×
Gemini	45.97	0.605	0.428	–	×

A key contribution of the morphological layer is its ability to recover previously unresolved tokens. Specifically, it successfully resolves 34 tokens that were initially assigned the NNN label by the baseline system. This demonstrates that a non-trivial portion of the system's unknown-word cases correspond to morphologically analyzable forms rather than genuinely unseen lexical items.

The comparatively larger improvement observed in CER than in WER suggests that the morphological post-processing module often produces outputs that are closer to the reference at the character level, even when exact word-level matches are not achieved. This indicates partial but meaningful recovery of correct Sinhala morphological structure.

An ablation analysis of the unknown-word handling pipeline further clarifies these effects. Out of 316 tokens initially labeled as NNN, the suffix-stripping and root re-tagging component successfully recovers 34 tokens (10.8% of unknowns). The remaining 282 tokens remain unresolved, primarily because either the resulting stripped root does not exist in the training vocabulary or the input token does not conform to any of the predefined suffix patterns covered by the morphological rules.

5.3 Comparison with General-Purpose LLMs

To situate the proposed system within a broader computational context, we evaluate three general-purpose large language models, including ChatGPT, DeepSeek, and Gemini AI, on the same held-out test set using identical implementations of BLEU, WER, and CER (see Table 3). The "Det." column indicates whether a system is deterministic and can be executed offline without an external API dependency. Among the evaluated systems, Gemini AI achieves the strongest aggregate performance, with a BLEU score of 45.97, WER of 0.605, and CER of 0.428, narrowly outperforming the morphological-enhanced system. ChatGPT ranks second in error-based metrics, while DeepSeek performs comparatively worse across most measures.

However, the performance gap between the morphological-enhanced system and the best-performing LLM is relatively small in absolute terms and should be interpreted cautiously. First, these general-purpose LLMs are not explicitly trained on Vedda data and likely have minimal or no exposure to Vedda in their pre-training corpora. Second, the evaluation is based on single-prompt, single-output generation per sentence against a single human reference, which limits statistical robustness and increases variance in measured performance. In contrast, the proposed morphological-enhanced system is explicitly designed for Vedda–Sinhala translation using a curated lexical resource and deterministic inference.

Despite slightly lower performance compared to the best LLM in some metrics, the morphological-enhanced system offers important practical advantages. It is fully deterministic, transparent, and reproducible; it operates offline without reliance on external APIs, and it is grounded in an inspectable lexical database that can be systematically extended. These properties make it particularly suitable for endangered-language settings where interpretability, resource control, and long-term sustainability are critical.

5.4 Error Analysis

Three main error classes account for the majority of the remaining failures.

- Out-of-vocabulary tokens: items that remain tagged as NNN after morphological recovery are preserved in the output unchanged. While this design maintains transparency, it negatively affects surface-level accuracy.
- Lexical ambiguity in culturally specific kinship terms, where the appropriate Sinhala rendering is context-dependent and cannot be resolved within a limited trigram context win-

dow. example: කොදුයි → අඩුයි / නෑ / බෑ / බැරිවුණා /එපා

- Multi-word or idiomatic compression, in which a single Vedda expression corresponds to a shorter Sinhala equivalent that cannot be recovered through surface n-gram matching alone. for a instance කබලෙ කරුකුරු ගච්චන ඇටපොජ්ජ → ඉරිඟු

6 Conclusion

This paper presents the first structured machine-readable Vedda–Sinhala lexical resource, consisting of 5,992 word pairs, together with a working statistical machine translation (SMT) prototype based on a cascading n-gram tagger. A morphological post-processing extension improves all four primary evaluation metrics over the baseline and recovers 34 previously untranslatable tokens. A controlled benchmark against three general-purpose large language models on the same test set positions the proposed system as a transparent, deterministic, and offline-capable alternative, with performance that is competitive with Gemini AI in absolute terms.

Overall, this work provides a reproducible baseline and a reusable digital resource that can support future research in lexicon expansion, morphological modeling, and hybrid neural–statistical machine translation approaches.

Limitations

Despite its contributions, this work is subject to several important limitations. The dataset used in this study, although the largest known machine-readable Vedda–Sinhala lexical resource, remains relatively small, consisting of 5,992 word pairs. It was compiled and annotated by a single annotator and therefore does not include inter-annotator agreement measurements, limiting the ability to quantify annotation consistency and potential subjectivity in lexical alignment. The evaluation set consists of only 226 sentences, each paired with a single human reference translation, which limits the statistical reliability and robustness of the reported automatic evaluation metrics, including BLEU, WER, and CER.

Due to these dataset size constraints, sequence-to-sequence neural models were not evaluated in this study. In particular, modern neural machine translation approaches typically require substantially larger parallel corpora to avoid severe overfitting and to achieve meaningful generalisation, making them unsuitable for reliable training under the current data conditions. As a result, this work focuses on a deterministic statistical baseline augmented with rule-based enhancements rather than data-intensive neural architectures.

Additionally, the Morphological Post-Processing module is inherently pattern-specific and covers only a subset of observed Vedda affixation phenomena. It does not constitute a full morphological analyzer and cannot handle unrestricted or previously unseen morphological constructions. The trigram-based context model is also limited to short-range dependencies, and therefore cannot capture broader discourse-level structure, cultural knowledge, or deeper semantic disambiguation required for more fluent translation.

The comparison with large language models is based on single-run prompting across three commercial systems, whose training data and exposure to Vedda-related linguistic material are not controlled, introducing potential variability in their performance.

Ethical Statement

This work involves an endangered Indigenous language and is conducted in accordance with the ACL Ethics Policy. All Vedda–Sinhala material was sourced from previously published linguistic and ethnographic works; no living informants were consulted in creating the dataset. The system explicitly surfaces unknown tokens rather than substituting uncertain guesses, thereby reducing the risk of misleading outputs. We acknowledge that any digital resource involving the Vedda language must ultimately be developed in collaboration with, and with the consent of, the Vedda community. Accordingly, we view the present prototype as a foundation for such future collaboration rather than as a replacement for it.

Acknowledgments

We gratefully acknowledge the contributions of all linguists, ethnographers, and authors whose foundational work has enabled the documentation and preservation of the Vedda language. In particular, we recognize the researchers and book authors whose published descriptions, wordlists, and ethnographic studies provided the essential material for constructing the Vedda–Sinhala lexical re-

source used in this work. Their efforts to document and preserve an endangered linguistic heritage form the backbone of this study. Without their careful fieldwork and scholarship, this resource would not have been possible.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

MGL Ananda. 2020. The syntax of the vedda language. *International Journal of Multidisciplinary Studies*, 7(2).

Hugo C Cardoso, Mahesh Radhakrishnan, Patrícia Costa, and Rui Pereira. 2019. Documenting modern sri lanka portuguese.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the ninth workshop on statistical machine translation*, pages 362–367.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

KNO Dharmadasa. 1974. The creolization of an aboriginal language: The case of vedda in sri lanka (ceylon). *Anthropological Linguistics*, pages 79–106.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli-Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Suvee Vathma Surendra, Swati Priyadarshini, and Shantipriya Parida. 2024. Preservation of vedda's language in sri lanka. In *Applying AI-Based Tools and Technologies Towards Revitalization of Indigenous and Endangered Languages*, pages 35–44. Springer.

Rohan Thimbiripola. 2022. Discursive socio-cultural identity: A study of cultural discourse trajectories in shaping the socio-cultural identity of veddas in sri lanka.

RAD Priyanka Weerasekara. 2020a. The linguistic study of the contemporary context of vedda language with special reference to dambana, sri lanka. *International Journal of Research and Innovation in Social Science (IJRISS)*, 4(8):737–741.

RADP Weerasekara. 2020b. The value of protecting endangered language in culture: With special reference to examine the vedda language in sri lanka. *International Journal of Research and Innovation in Social Science*, 4(10):429–434.

A Used artefacts for creating the dataset

For the dataset creation process, we used the following books as the foundation to gather the Vedda languages.

- බින්තැන්නේ සිංහල, Binthanne Sinhala by Sandaruwan Lokuhewa.
- සින්හලයේ පරිනාමය, Sinhalaye parinamaya by Somanandha Himi.
- වැද්දෝ, Vaddo by Chandra Sri Ranasinghe.
- වැද්දන්ගේ විකාශන ක්‍රමය Veddange vikashana kramaya by Nandadewa Wijesekara
- ලංකාවේ වැද්දෝ lankawe Veddo by E.M. Rathnapala
- දඹානේ වැදිබස Dhabane Wadhibase by Sugathadasa de Silva