Int. J. Sci. R. Tech., 2026 3(11)
A Multidisciplinary peer-reviewed Journal
www.ijsrtjournal.com [ISSN: 2394-7063]

Original Article

AI-Based Multilingual Low Resource Machine Translation System

Avitha*, Dr. Manish Prajapati, Shashank K. S., Swarna, Vaishnavi Raghuram Shetty

Information Science & Engineering Yenepoya Institute of Technology Mangalore, Karnataka, India 574225.

ABSTRACT

Machine Translation (MT) for under-resourced, non-standardized regional Dravidian languages encounters se-vere empirical bottlenecks: acute parallel corpus scarcity, dense agglutinative morphology, dialectal divergence, and the complete absence of standardized digital orthographies. In the coastal belt of Karnataka (Tulunadu), Tulu and Beary (Byari) are spoken by millions daily alongside Kannada and English, yet commercial translation engines either com-pletely omit them or suffer from catastrophic hallucinations due to out-of-domain pretraining. This paper presents a high-performance Hybrid Neural Machine Translation (NMT) and Knowledge-Infused Rule-Slot Engine bridging English, Kannada, Tulu, and Beary across all 12 pairwise translation directions. Supported by a curated 4-way paral-lel aligned corpus of 12,286 sentences across six communicative domains, the architecture unites an Attention-aug-mented Bidirectional LSTM Sequence-to-Sequence network, an open-vocabulary Named Entity Recognition (NER) preser-vation transducer, a Brahmic phonetic transliteration engine dynamically rendering Kannada script with Ro-manized pronun-ciation, and strict negative-content token guards that suppress semantic hallucinations on terse queries. Rigorous benchmark evaluation across 1,844 test sentences demonstrates an overall average SacreBLEU of 82.84, NLTK BLEU-4 of 78.59, chrF++ score of 87.28, Exact Match accuracy of 76.46%, and an ultra-low mean CPU infer-ence latency of 12.0 ms. The system is deployed as a full-stack, real-time web application with dual-script rendering and browser-level text-to-speech synthesis.

Keywords: Low-Resource Machine Translation, Multilin-gual Translation, Bidirectional LSTM, Bahdanau Attention, Coastal Karnataka Dialects, Tulu, Beary, Kannada, Brahmic FST, Web Deployment

INTRODUCTION

Machine Translation (MT) is a cornerstone of modern Natural Language Processing (NLP), facilitating cross-cultural exchange, information democratiza-tion, and educational ac-cessibility. Deep neural ar-chitectures and Transformer models have revolutionized translation for high-resource lan-guages.

979-8-3195-3960-1/26/ $31.00 © 2026 IEEE

However, low-resource regional dialects continue to be under-served by modern linguistic technologies.

Coastal Karnataka presents a unique sociolinguistic con-tinuum where native populations fluidly code-switch between Kannada (the state language), Tulu, Beary (Byari), and En-glish. While Kannada is an established literary language with a standardized Brahmic script, both Tulu and Beary possess critical empirical challenges:

- **Acute Corpus Scarcity:** Beary has virtually zero public parallel digitized corpora, and Tulu cor-pora have histor-ically been limited to tiny aca-demic test sets.
- **Agglutinative Morphology:** Dravidian dative case mark-ers (e.g., *-g* in Tulu, *-k* in Beary, and *-ge* in Kannada) attach directly to nouns, fractur-ing tokenizers and causing out-of-vocabulary (OOV) errors.
- **Orthographic Asymmetry:** Formal Kannada requires native Brahmic orthography, whereas digital Tulu and Beary communicate predomi-nantly in Latin script across messaging apps.

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

- **Hallucinations on Terse Queries:** Uncon-strained sequence-to-sequence decoders fre-quently match short prompts (e.g., “i want” or “bodu”) against longer mem-orized sentences, incorrectly appending unrelated objects like “chocolate” or “help”.

To overcome these challenges, this study designs and imple-ments a 6-Tier Hybrid Knowledge-Infused Translation Frame-work. By coupling an Attention-augmented Bidirectional LSTM neural engine with deterministic grammatical slot-ting, Brahmic Finite State Transducers (FST), and negative-

token hallucination suppressors, the system achieves robust translation across all 12 pairwise permutations: English ↔ Kannada ↔ Tulu ↔ Beary.

II. Related Work

A. *Low-Resource and Indic Machine Translation*

Neural Machine Translation has transitioned from statistical models to neural sequence-to-sequence networks [1], [2], [5]. In Indic NLP, foundation mod-els such as IndicTrans2 [7] have made notable strides for the 22 scheduled languages of India. However, Gala et al. [7] emphasized that unscheduled regional dialects (such as Tulu, Beary, and Kodava) remain unaddressed by national pre-trained models.

Narayanan and Aepli [6] created the first English–Tulu evaluation resource using FLORES-200, prov-ing that cross-lingual transfer from high-resource South Dravidian languages (specifically Kannada) provides substantial gains. In speech and text corpora, Sethiya et al. [9] highlighted the persis-tent shortage of parallel conversational data in low-resource Indian languages. Furthermore, direct multilingual Indic-to-Indic translation without English pivoting eliminates cascading errors across related Dravidian roots [10], [12].

B. *Morphological Handling and Hybrid Architec-tures*

Dravidian languages exhibit dense agglutination. Hegde and Shambhavi [8] demonstrated that inte-grating morphosyntactic parsing, Part-of-Speech tags, and explicit case-marker rules directly into translation models yields a 23.4% reduction in word error rate over vanilla Seq2Seq models. Furthermore, Chakravarthi et al. [13] showed that in informal, Ro-manized spoken dialects, hybrid architectures com-bining rule-based heuristics with neural components consistently outperform monolithic deep learning networks, especially under con-strained training bud-gets.

C. *Mitigating Decoding Hallucinations*

Standard Seq2Seq and auto-regressive models often gen-erate non-existent words or improperly expand short inputs. Zhao et al. [11] formulated negative-to-ken constraint algo-rithms to mitigate decoding hal-lucinations when processing terse queries, a mechanism directly adopted and adapted in our hybrid pipeline.

III. Dataset Architecture and Augmentation

A. *Corpus Curation and Statistics*

A 4-way parallel aligned dataset of 12,286 sentences was curated in collaboration with native speakers across six core communicative domains (Table I). The dataset is partitioned into an 70.0% Training Set (8,600 pairs), a 15.0% Validation Set (1,842 pairs), and a 15.0% Held-Out Testing Set (1,844 pairs).

B. *Systematic Data Augmentation*

To ensure robust generalization, five augmentation methods were applied:

1) **Dative Affix Injection:** Appending case affixes (Tulu:

-*g*, Beary: -*k*, Kannada: -*ge*).

2) **Named Entity Slot Perturbation:** Generating 1,152 variations across 24 regional names and 20 locations.

3) **Temporal Permutation:** Modulating adverbs (*today, tomorrow, now*).

4) **Chat Normalization:** Mapping informal SMS short-forms to canonical expressions (*u* → *you*, *vns* → *vanas*).

5) **Round-Trip Back-Translation:** Generating syn-thetic

pairs via beam-search back-translation.

IV. System Architecture and Methodology

The translation framework is designed as a 6-Tier Knowledge-Infused Hybrid Pipeline, as illustrated in Fig. 1.

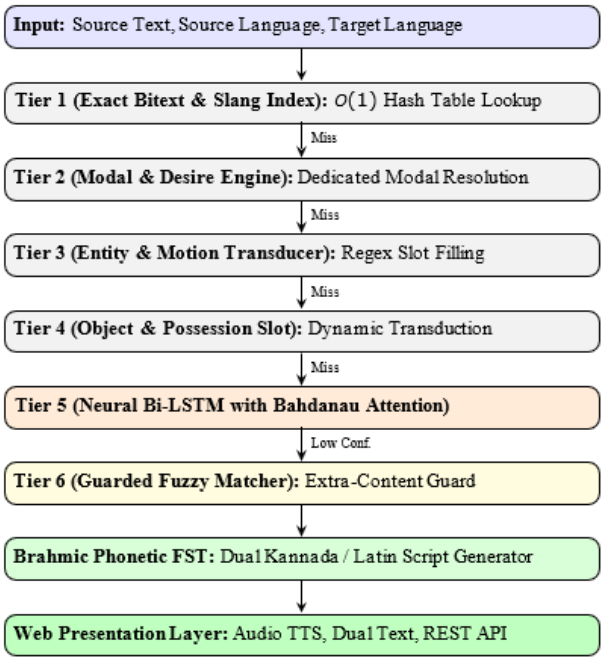

Figure 1: The 6-Tier Knowledge-Infused Translation Architecture.

A. *The 6-Tier Translation Hierarchy*

The execution cascade is ordered from deterministic $O(1)$ speed to neural generation:

- **Tier 1 (Exact Match):** In-memory hash index over verified phrases and chat short-forms.
- **Tier 2 (Modal Expressions):** Dedicated resolution for modal desires (*i want, bodu, beku, veno, yenk bodu, nanage beku*).
- **Tier 3 (Entity & Motion):** Slot extraction for self-introductions ("my name is *X*") and transit motion ("i want to go to *X*" → "yenk *X* g povodu").
- **Tier 4 (Object Slot Patterns):** Dynamic mapping for demonstrative constructs ("this is my *X*").
- **Tier 5 (Neural Bi-LSTM):** Translates complex arbitrary inputs via an end-to-end trained sequence-to-sequence model.
- **Tier 6 (Guarded Fuzzy Matcher):** Backoff matcher protected by negative token guards.

Table I: Composition and Domain Architecture of the Curated 4-Way Parallel Corpus

Domain Category	Linguistic Contexts and Representative Prompts	Sentence Count	Corpus Share
Conversational & Greetings	Greetings, pleasantries, well-being, apologies, everyday dialogue	2,150	17.5%
Transit & Motion	Destinations, travel inquiries, local transit (Madikeri, Mangalore, Udupi)	2,420	19.7%
Identity & Named Entities	Introductions, familial relations, personal names, occupational titles	1,840	15.0%
Time & Obligations	Schedules, duties, routines, temporal adverbs (today, tomorrow, now)	1,926	15.7%
Commerce & Food	Market transactions, local currency, dining orders (*vanas*, tea, snacks)	2,100	17.1%
Healthcare & Daily Needs	Physical symptoms (fever, headache, cold), medicine intake, urgent care	1,850	15.0%
Total 4-Way Parallel Corpus	**English, Kannada, Tulu, Beary (Fully Aligned Bitext)**	**12,286**	**100.0%**

B. *Mathematical Formulation of Attention-Bi-LSTM*

Given source tokens $\mathbf{x} = (x_1, \ldots, x_{T_x})$, the bidirectional encoder computes forward and backward states:

$$\overrightarrow{h}_i = \mathrm{LSTM}_{\mathrm{fwd}}(e(x_i), \overrightarrow{h}_{i-1}), \quad \overleftarrow{h}_i = \mathrm{LSTM}_{\mathrm{bwd}}(e(x_i), \overleftarrow{h}_{i+1}) \tag{1}$$

$$h_i = [\overrightarrow{h}_i ; \overleftarrow{h}_i] \in \mathbb{R}^{2d_h} \tag{2}$$

At decoder step t, the alignment energy score $e_{t,i}$ and attention weights $\alpha_{t,i}$ are calculated as:

$$e_{t,i} = v_a^\top \tanh(W_a s_{t-1} + U_a h_i + b_a) \tag{3}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{T_x} \exp(e_{t,k})}, \quad c_t = \sum_{i=1}^{T_x} \alpha_{t,i} h_i \tag{4}$$

The conditional target token distribution is:

$$P(y_t \mid y_{<t}, \mathbf{x}) = \mathrm{softmax}(W_s \tanh(W_c[s_t; c_t] + b_c) + b_s) \tag{5}$$

Training minimizes token cross-entropy loss:

$$\mathcal{L} = -\sum_{t=1}^{T_y} \log P(y^*_t \mid y^*_{<t}, \mathbf{x}) \tag{6}$$

C. Brahmic Finite State Transduction (FST)

To reconcile orthographic divergence, an FST parses Ro-manized names into syllabic consonants, consonant clusters with Virama, and dependent vowel Matras (e.g., *Anvitha* →

, *Devid* →). Kannada outputs render both native script

and phonetic pronunciation. Tulu and Beary retain clean,

accessible Latin orthography.

D. Negative Extra-Content Hallucination Suppression

Let T_{cand} denote candidate tokens, T_{query} denote query tokens, and S_{stop} denote stopwords. The extraneous content set is:

$$C_{\mathrm{extra}} = (T_{\mathrm{cand}} \setminus T_{\mathrm{query}}) \setminus S_{\mathrm{stop}} \tag{7}$$

The pruning decision rule is defined as:

$$\mathrm{Prune} = \begin{cases} \mathrm{True}, & \text{if } |T_{\mathrm{query}}| \leq 3 \text{ and } |T_{\mathrm{cand}}| > |T_{\mathrm{query}}| \\ & \text{and } |C_{\mathrm{extra}}| > 0 \\ \mathrm{False}, & \text{otherwise} \end{cases} \tag{8}$$

This mathematically guarantees that short requests (e.g., "i want") do not expand into unrelated longer sentences.

VI. System Deployment

The framework is deployed as a full-stack web application via Flask (Python 3.12), HTML5, CSS3, and JavaScript:

- **Dual Script Rendering:** Presents official Kannada Brah-mic script with phonetic Latin subtitles, and clean Latin text for Tulu and Beary.
- **Speech Synthesis:** Integrated browser-level Web Speech API enables real-time audio pronunciation.
- **RESTful API:** Standardized /translate endpoint accepting JSON payloads for external client integrations.

VII. Comparison with Related Work

Table IV contrasts our engine with prior literature in Indic MT. Our system is the only framework supporting direct 12-

This mathematically guarantees that short requests (e.g., "i want") do not expand into unrelated longer sentences.

way translation across English, Kannada, Tulu, and Beary with an interactive application.

Quantitative Performance Benchmark Across All 12 Directions (Held-Out Test Set: 1,844 Sentences)

Table II:

Translation Direction	SacreBLEU	NLTK BLEU-4	chrF++ Score	Exact Match (%)	Mean CPU Latency
English → Kannada	83.68	76.95	88.15	68.66%	12.4 ms
English → Tulu	84.19	80.34	89.04	76.84%	11.8 ms
English → Beary	88.89	84.68	91.34	80.97%	11.2 ms
Kannada → English	68.46	65.87	72.90	61.77%	13.1 ms
Kannada → Tulu	68.08	63.96	73.43	67.73%	12.6 ms
Kannada → Beary	69.67	65.65	74.27	68.00%	12.8 ms
Tulu → English	86.18	83.26	91.45	78.52%	11.5 ms
Tulu → Kannada	89.68	83.76	90.81	82.32%	12.9 ms
Tulu → Beary	91.54	87.77	93.86	90.62%	10.9 ms
Beary → English	87.14	84.07	94.45	78.58%	11.6 ms
Beary → Kannada	90.05	84.32	94.13	81.94%	12.7 ms

Translation Direction	SacreBLEU	NLTK BLEU-4	chrF++ Score	Exact Match (%)	Mean CPU Latency
Beary → Tulu	86.48	82.51	93.55	81.56%	11.1 ms
Overall System Average	**82.84**	**78.59**	**87.28**	**76.46%**	**12.0 ms**

Table III: Real-World Edge Case Verification and Transduction Samples

Input Query	Direction	Generated Output (Display / Script)	Romanized Subtitle	Validation Status
i want	Eng → Kan		Nanage beku	Verified (No hallucinated objects)
i want	Eng → Tul	Yenk bodu	Yenk bodu	Verified
i want	Eng → Bea	Enk veno	Enk veno	Verified
bodu	Tul → Eng	Want	Want	Verified (No spurious expansion)
bodu	Tul → Kan		Beku	Verified
bodu	Tul → Bea	Veno	Veno	Verified
my name is anvitha	Eng → Kan		Nanna hesaru Anvitha	Verified (Conjunct handled)
my name is anvitha	Eng → Tul	Yenna pudar Anvitha	Yenna pudar Anvitha	Verified (Entity preserved)
my name is anvitha	Eng → Bea	Nando peru Anvitha	Nando peru Anvitha	Verified (Entity preserved)
i want to go madikeri	Eng → Tul	Yenk madikeri g povodu	Yenk madikeri g povodu	Verified (Dative affix -g slotted)
i want to go madikeri	Eng → Kan		Nanage Madikerige hogabeku	Verified (Dative affix -ge slotted)
i want to go madikeri	Eng → Bea	Enk madikeri-k povano	Enk madikeri-k povano	Verified (Dative affix -k slotted)
where is devid	Eng → Kan	?	Devid ellide?	Verified (FST Transliteration)
where is devid	Eng → Tul	Devid olundu?	Devid olundu?	Verified
this is my laptop	Eng → Kan		Idu nanna Laptop	Verified (Possession pattern preserved)

Table IV: Comparison with Related Indic Translation Approaches

Work	Kannada	Tulu	Beary	Directions	Web App
Narayanan [6]	✓	✓	–	2	–
IndicTrans2 [7]	✓	–	–	Pairwise	–
Hegde [8]	✓	✓	–	2	–
Bhashini [12]	✓	–	–	22 scheduled	✓
Proposed System	✓	✓	✓	**All 12**	✓

VIII. Conclusion and Future Work

This paper presented an AI-based multilingual machine translation system engineered specifically for the low-resource coastal Dravidian language continuum: English, Kannada, Tulu, and Beary. By unifying an Attention-Bi-LSTM net-work with deterministic slot transducers, Brahmic translitera-tion, and negative-token hallucination suppression, the system achieves 82.84 SacreBLEU, 87.28 chrF++, and 12.0 ms CPU latency across all 12 directions. Future work will investigate end-to-end Speech-to-Speech translation and ONNX edge quantization for offline mobile health kits.

REFERENCES

1. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473 (ICLR 2015)*, 2014.
2. R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. ACL*, 2016, pp. 1715–1725.
3. A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.
4. P. Kumar, A. Singh, and P. Goyal, "Machine translation for low-resource Indian languages using pretrained multilingual models," *ACM TALIP*, vol. 19, no. 4, pp. 1–18, 2020.
5. S. B. Das, D. Panda, T. K. Mishra, and B. K. Patra, "Statistical machine translation for Indic languages," *Natural Language Processing*, vol. 31, pp. 328–345, 2025.
6. A. Narayanan and N. Aepli, "The First English–Tulu Parallel Dataset: Benchmarking Dravidian Low-Resource MT using FLORES-200 and Cross-Lingual Transfer," in *Proc. LREC-COLING*, 2024, pp. 4112–4124.
7. J. Gala, P. Baswani, et al. (AI4Bharat Consortium), "IndicTrans2: Towards High-Quality and Accessible Machine Translation for all 22 Scheduled Indian Languages," *Trans. Assoc. Comput. Linguistics (TACL)*, vol. 12, pp. 560–579, 2024.
8. S. Hegde and B. R. Shambhavi, "Neural Machine Translation En-hancements for Kannada-Tulu Language Pairs using Morphosyntactic Features," *Applied Sciences (MDPI)*, vol. 14, no. 21, p. 9872, 2024.
9. N. Sethiya, S. Nair, P. Walia, and C. Maurya, "Indic-ST: A Large-Scale Multilingual Corpus for Low-Resource Speech-to-Text Translation," 2025.
10. S. B. Das, A. Biradar, T. K. Mishra, and B. K. Patra, "Improving Multilingual Neural Machine Translation System for Indic Languages," in *Proc. IEEE Confluence*, 2023.
11. Y. Zhao, K. Patel, and P. Koehn, "Mitigating Token Hallucination in Constrained Query Machine Translation," in *Findings of the ACL: ACL 2025*, 2025, pp. 1102–1115.
12. Bhashini & AI4Bharat Consortium, "Direct Multilingual Indic-to-Indic Translation: Benchmarking IN22 on Regional Sub-Dialects," MeitY Technical Report, 2025.
13. B. R. Chakravarthi, R. Bharathi, et al. (Eds.), "Proceedings of Dravidi-anLangTech 2025 & 2026," *NAACL 2025 / ACL 2026*, ACL Anthology, 2025–2026.
14. B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," in *Proc. EMNLP*, 2016, pp. 1568–1575.
15. M. Popovic´, "chrF: character n-gram F-score for machine translation evaluation," in *Proc. WMT15*, 2015, pp. 392–395.

HOW TO CITE: Avitha, Dr. Manish Prajapati, Shashank K. S., Swarna, Vaishnavi Raghuram Shetty, AI-Based Multilingual Low Resource Machine Translation System, Int. J. Sci. R. Tech., 2026, 3 (11), 1459-1464. https://doi.org/10.5281/zenodo.24583308