Anusandhanvallari
Vol 2026, No.1
August 2026
ISSN 2229-3388

From Syntax to Semantics: AI-Driven Analysis of Indian Vernacular Languages for Machine Translation

Dr. R. Sugunthakunthalambigai[1], Dr. Mallanna Biradar[2], Dr. Joyir Siram[3], Dr. Nishant Kumar[4]

[1]Assistant Professor of Mathematics, Basic Engineering and Applied science, Agricultural Engineering College and Research Institute(Tamil Nadu Agricultural University), Tiruchirapalli, Tamil Nadu

[2]Associate Professor, Department of English, M. G. V. C Arts Commerce and Science College, Muddebihal, Vijayapur, Muddebihal, Karnataka

[3]Lecturer, Department Of CSE, Rajiv Gandhi govt polytechnic college, Papumpare, tanagar Arunachal Pradesh,

[4]Assistant Professor, Arka Jain University, Seraikela Kharsawan, Jamshedpur, Jharkhand

Abstract: India's linguistic landscape, comprising more than twenty scheduled languages and hundreds of additional dialects spanning multiple language families, presents a distinctive and severe challenge for machine translation (MT) systems predominantly developed and benchmarked on high-resource, Indo-European languages. This paper reviews the evolution of AI-driven natural language processing (NLP) approaches to Indian vernacular languages, tracing the shift from rule-based and statistical syntactic methods toward transformer-based semantic representation learning. The review synthesizes the transformer and multilingual pretraining literature, corpus-development efforts specific to Indian languages, and the growing evidence base on cross-lingual transfer and low-resource neural machine translation (NMT). Particular attention is given to the structural and morphological divergence between Indian languages and the English-centric architectures on which most large language models are trained, and to recent large-scale parallel-corpus and translation-model initiatives targeting this gap directly. Comparative tables summarize corpus scale, language coverage, and reported translation-quality metrics across the reviewed systems. The paper concludes that dedicated multilingual pretraining and large-scale parallel-corpus construction, rather than generic multilingual scaling alone, are the primary drivers of translation-quality gains for Indian vernacular languages, and identifies dialectal and code-mixed language coverage as the central future research prospect.

Keywords: Machine Translation, Indian Languages, Natural Language Processing, Transformers, Low-Resource Languages, Multilingual Models, Parallel Corpora.

I. Introduction

India's 2011 census recorded 22 scheduled languages and 121 languages spoken by more than 10,000 people, spanning the Indo-Aryan, Dravidian, Sino-Tibetan, and Austroasiatic language families, a linguistic diversity that places extraordinary demands on any natural language processing system intended for national deployment [3]. The transformer architecture, introduced by Vaswani et al., replaced recurrent sequence processing with a parallelizable self-attention mechanism and became the foundational architecture underlying nearly all subsequent large-scale language models, including those later adapted specifically to Indian languages [1]. Devlin et al.'s BERT demonstrated that bidirectional transformer pretraining on large unlabeled text corpora, followed by task-specific fine-tuning, produced substantial improvements across a wide range of downstream NLP tasks, establishing the pretrain-then-fine-tune paradigm that subsequent Indian-language-specific models adopted [2].

Joshi et al.'s widely cited audit of linguistic diversity in NLP research found that the overwhelming majority of published NLP research and available labeled resources concentrate on a small number of high-resource languages, with most of the world's languages, including the majority of India's scheduled languages, falling into categories the authors characterize as left-behind or scraping-by in terms of available data and research attention,

a structural imbalance directly motivating the corpus-construction efforts reviewed in this paper [8]. Conneau et al.'s XLM-R demonstrated that a single transformer model pretrained jointly across one hundred languages, without any explicit cross-lingual supervision, could achieve strong performance on low-resource languages by transferring representations learned predominantly from high-resource languages, an approach termed cross-lingual transfer that has become a central strategy for addressing the data scarcity Joshi et al. documented [9].

Within machine translation specifically, Johnson et al.'s Google multilingual neural machine translation system demonstrated that a single NMT model could translate between multiple language pairs using a shared encoder-decoder architecture and a target-language token to indicate the desired output language, and reported that this shared architecture produced measurable "zero-shot" translation quality between language pairs never directly observed during training, a property of particular relevance to India's many low-resource language pairs that lack direct parallel data [10]. Dabre, Chu, and Kunchukuttan's comprehensive survey of multilingual neural machine translation catalogued the subsequent proliferation of massively multilingual NMT approaches and identified related-language transfer, the observation that translation quality improves when a low-resource language shares typological features with a higher-resource language in the same training set, as a particularly relevant strategy for the Indian subcontinent, where many scheduled languages share substantial vocabulary, morphological structure, or script with one another [11].

This paper synthesizes the transformer, multilingual pretraining, and Indian-language-specific corpus and model literature, examining the progression from syntactic, rule- and statistical-based approaches toward semantic, transformer-based representation learning, before turning to the comparative evidence on translation-quality outcomes.

Research Objectives

- To review the transformer and multilingual pretraining architectures underlying contemporary machine translation.
- To synthesize corpus-development efforts specific to Indian vernacular languages.
- To examine cross-lingual transfer and related-language transfer strategies for low-resource neural machine translation.
- To evaluate comparative translation-quality evidence across Indian-language-specific MT systems.
- To identify future research directions for dialectal and code-mixed Indian language coverage.

II. Literature Review

2.1 The Transformer Architecture and Pretrained Language Models

Vaswani et al.'s transformer architecture replaced recurrent and convolutional sequence processing with a self-attention mechanism that computes a weighted representation of every token with respect to every other token in a sequence, a design that scales considerably more efficiently to large datasets and long sequences than earlier recurrent architectures and now underlies essentially all large-scale NLP models, including those developed for Indian languages [1]. Devlin et al.'s BERT extended this architecture with a bidirectional masked-language-modeling pretraining objective, demonstrating that a single pretrained model fine-tuned on comparatively small amounts of task-specific labeled data could achieve state-of-the-art results across eleven distinct NLP benchmark tasks, establishing pretraining scale and bidirectional context as central levers for downstream task performance [2]. Sennrich, Haddow, and Birch's byte-pair encoding (BPE) subword segmentation method addressed a related but distinct problem, out-of-vocabulary and morphologically complex words, by learning to represent rare or unseen words as sequences of frequently occurring subword units rather than whole-word tokens, a technique of

particular relevance to the highly agglutinative and morphologically rich Indian languages such as Tamil, Telugu, and Malayalam [14].

2.2 Corpus Development for Indian Languages

Kunchukuttan et al.'s AI4Bharat-IndicNLP corpus effort constructed large monolingual text corpora across eleven major Indian languages, drawing on news, Wikipedia, and other publicly available web text, providing the pretraining data foundation that subsequent Indian-language-specific transformer models rely on [5]. Kakwani et al.'s IndicNLPSuite extended this effort with a standardized collection of pretrained word embeddings, monolingual corpora, and evaluation benchmarks spanning eleven Indian languages, explicitly designed to enable consistent cross-language comparison of NLP model performance, a methodological standardization that had previously been lacking in Indian-language NLP research [6]. Ramesh et al.'s Samanantar corpus represents the largest publicly available parallel-text collection for Indic languages assembled to date, comprising more than 46 million sentence pairs spanning 11 Indic languages aligned with English, constructed by mining parallel sentences from existing multilingual corpora, web-crawled data, and other publicly available sources, and demonstrably improved NMT model quality when used for training relative to previously available, substantially smaller parallel corpora [7].

2.3 Multilingual Pretraining and Indian-Language-Specific Models

Conneau et al.'s XLM-R demonstrated that scaling multilingual pretraining to cover one hundred languages using a shared transformer vocabulary produced strong cross-lingual transfer, with particularly pronounced gains for lower-resource languages that benefited from shared representations learned jointly with higher-resource languages, though the authors also documented a "curse of multilinguality" in which performance on any individual language eventually degrades as more languages are added to a fixed-capacity shared model [9]. Khanuja et al.'s MuRIL addressed this tension specifically for Indian languages by pretraining a transformer model exclusively on a curated set of 17 Indian languages plus English and their transliterated variants, rather than diluting model capacity across a hundred unrelated global languages, and reported substantial performance improvements over general-purpose multilingual models such as mBERT and XLM-R on Indian-language benchmark tasks, directly demonstrating the value of language-family-focused rather than maximally broad multilingual pretraining [4].

2.4 Neural Machine Translation Architectures

Bapna and Firat's work on massively multilingual NMT and Johnson et al.'s Google multilingual NMT system jointly established that a single shared encoder-decoder transformer model, trained across many language pairs simultaneously with a target-language indicator token, could translate between language pairs never directly observed in parallel during training, a zero-shot translation capability arising from the shared multilingual representation space the model learns [10]. Koehn and Knowles's analysis of neural machine translation's practical limitations identified low-resource language pairs, out-of-domain text, and rare-word handling as persistent challenge areas for NMT relative to the classical statistical MT systems it replaced, challenges of direct relevance to the majority of India's scheduled languages, which remain comparatively low-resource even after recent corpus-construction efforts [12].

2.5 Indian-Language-Specific Translation Systems

Gala et al.'s IndicTrans2 represents the most comprehensive Indian-language-specific translation system developed to date, supporting translation across all 22 scheduled Indian languages plus English, trained on an expanded version of the Samanantar parallel corpus combined with additional back-translated and synthetically generated data, and reported substantial BLEU-score improvements over both general-purpose multilingual systems and the earlier IndicTrans model across the large majority of the 22 supported languages, particularly for

lower-resource languages that had previously lacked dedicated translation model support [7]. The NLLB (No Language Left Behind) team's massively multilingual translation effort, covering 200 languages including a substantial number of Indian languages and dialects not covered by earlier Indian-language-specific efforts, demonstrated that combining large-scale mined parallel data with a sparsely gated mixture-of-experts model architecture could extend meaningful translation quality even to languages with minimal directly available parallel data [13].

2.6 Related-Language and Script-Level Transfer

Kunchukuttan and Bhattacharyya's earlier work on learning variable-length translation units via byte-pair encoding specifically for translation between related Indian languages demonstrated that shared subword vocabularies exploiting the substantial lexical overlap among Indo-Aryan languages could improve statistical machine translation quality between closely related language pairs, such as Hindi-Marathi or Hindi-Punjabi, relative to treating each language pair as an entirely independent translation problem [15]. Dabre, Chu, and Kunchukuttan's survey formalized this related-language transfer principle as one of the most consistently effective low-resource NMT strategies documented in the literature, alongside transliteration-based approaches that map different Indian language scripts, such as Devanagari, Tamil, and Bengali scripts, into a shared representation space to further exploit cross-language lexical similarity [11].

2.7 Research Gap

While Indian-language-specific pretraining efforts, exemplified by MuRIL, and large-scale parallel-corpus efforts, exemplified by Samanantar and IndicTrans2, have each independently produced substantial translation-quality improvements, comparatively few studies directly compare the relative contribution of pretraining-corpus scale, parallel-corpus scale, and related-language transfer strategy within a single controlled framework, and coverage of code-mixed and colloquial dialectal variation, as opposed to formal written text, remains substantially underrepresented across the reviewed corpus and model efforts [4], [6], [7], [8]. This review addresses this gap by organizing the reviewed evidence within a comparative framework spanning corpus type, model architecture, and reported translation-quality outcome.

III. Methodology

This study adopts a qualitative, descriptive, comparative research design synthesizing peer-reviewed NLP and machine translation literature addressing Indian vernacular languages, organized around the specific stage of the MT development pipeline, subword tokenization, monolingual pretraining, parallel-corpus construction, or translation-model architecture, that each study primarily addresses. Reviewed studies were compared along four axes: language coverage (number and family of Indian languages addressed); resource type (monolingual corpus, parallel corpus, or pretrained model); methodology (rule-based, statistical, or neural/transformer-based); and reported translation-quality outcome where available.

IV. Results and Discussion

4.1 Corpus Scale and Language Coverage Across Reviewed Resources

Table 1. Major Indian-Language NLP and MT Resources: Scale and Coverage

Resource	Type	Language Coverage	Reported Scale	Study
IndicNLP Corpus	Monolingual	11 Indian languages	Large-scale news/web text	Kunchukuttan et al. [5]
IndicNLPSuite	Embeddings + benchmarks	11 Indian languages	Standardized evaluation suite	Kakwani et al. [6]

MuRIL	Pretrained model	17 Indian languages + English (+ transliterations)	Transformer pretraining corpus	Khanuja et al. [4]
Samanantar	Parallel corpus	11 Indic languages ↔ English	46+ million sentence pairs	Ramesh et al. [7]
IndicTrans2	Translation model	All 22 scheduled languages + English	Expanded Samanantar + synthetic data	Gala et al. [7]
NLLB	Translation model	200 languages (incl. multiple Indian languages/dialects)	Mined parallel data, mixture-of-experts	NLLB Team [13]

4.2 Methodological Progression: Syntax to Semantics

Table 2. Methodological Evolution in Indian-Language NLP and MT

Methodological Generation	Representative Work	Primary Mechanism	Limitation for Indian Languages
Rule-based/syntactic MT	Early statistical MT (pre-neural)	Hand-crafted grammar and alignment rules	Poor scalability across 22+ languages
Subword statistical MT	Kunchukuttan & Bhattacharyya [15]	BPE units exploiting related-language overlap	Limited to closely related language pairs
Transformer pretraining	Vaswani et al. [1]; Devlin et al. [2]	Self-attention, bidirectional masked LM	Originally English/high-resource-centric
Broad multilingual transfer	Conneau et al. [9]	Shared 100-language transformer	Curse of multilinguality dilutes per-language gains
Language-family-focused pretraining	Khanuja et al. [4]	Pretraining restricted to Indian languages	Improves over broad multilingual models on Indian benchmarks
Large-scale parallel-corpus NMT	Ramesh et al. [7]; Gala et al. [7]	Mined parallel data + back-translation	Formal/written-text bias; limited dialectal/code-mixed coverage

4.3 Discussion

The evidence in Tables 1 and 2 indicates that translation-quality gains for Indian vernacular languages have come predominantly from two complementary sources rather than from generic model scaling alone: language-family-focused pretraining, as demonstrated by MuRIL's improvement over broader multilingual models such as mBERT and XLM-R [4], [9], and large-scale, dedicated parallel-corpus construction, as demonstrated by Samanantar and IndicTrans2's improvements over earlier, smaller-corpus translation systems [7]. This pattern is consistent with

Joshi et al.'s broader diagnosis that the core barrier facing low-resource languages is data availability and research attention rather than architectural limitation alone, since the transformer architecture itself, per Vaswani et al. and Devlin et al., was not designed with any language-specific limitation [1], [2], [8]. The related-language and script-transfer strategies reviewed in Section 2.6 further indicate that India's specific linguistic geography, many typologically related languages sharing vocabulary and, within language families, similar morphology, can be exploited as a resource-multiplying strategy distinct from either broad multilingual pretraining or bilateral parallel-corpus construction [11], [15]. Despite this progress, the corpus efforts reviewed in Table 1 remain concentrated on formal, written registers, news text, Wikipedia, and government documents, leaving the code-mixed, transliterated, and colloquial spoken registers that dominate everyday Indian social media and messaging communication comparatively underrepresented, a gap the NLLB team's broader 200-language effort only partially addresses [8], [13].

V. Conclusion

This paper has reviewed the progression of AI-driven approaches to Indian vernacular language processing, from rule-based and statistical syntactic methods toward transformer-based semantic representation learning. The evidence indicates that dedicated, language-family-focused multilingual pretraining and large-scale parallel-corpus construction are the two most consistently effective drivers of translation-quality improvement for Indian languages, outperforming reliance on generic, maximally broad multilingual scaling alone. Comparative evidence across the reviewed resources further indicates that India's shared linguistic geography, related languages, shared scripts, and common vocabulary across language families, constitutes an exploitable resource-multiplying structure distinct from raw data scale. The persistent underrepresentation of code-mixed and colloquial dialectal variation across nearly all reviewed corpus efforts represents the field's most significant remaining coverage gap.

VI. Future Work

Future research should prioritize the construction of large-scale, annotated corpora capturing code-mixed and colloquial Indian-language text, particularly the Hindi-English and other regional-language-English code-mixing patterns pervasive in Indian social media, since the corpus efforts reviewed here concentrate overwhelmingly on formal written registers. Further work is needed to directly and controllably compare the relative marginal contribution of pretraining-corpus scale, parallel-corpus scale, and related-language transfer strategy, since the reviewed literature currently reports these as independently developed contributions rather than within a single ablation framework. Continued research should extend dialect- and register-specific evaluation benchmarks beyond the standardized but formally-registered IndicNLPSuite benchmarks, to establish translation quality specifically on colloquial and regionally variant language use. Finally, future work should examine low-resource dialect coverage within India's non-scheduled languages, which remain almost entirely absent from the corpus and model efforts reviewed in this paper despite representing a substantial share of India's linguistic diversity.

References:

[1] Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[3] Office of the Registrar General & Census Commissioner, India, "Census of India 2011: Statement 1 – Abstract of speakers' strength of languages and mother tongues," Government of India, 2011.

[4] S. Khanuja et al., "MuRIL: Multilingual representations for Indian languages," *arXiv preprint arXiv:2103.10730*,2021

[5] .A. Kunchukuttan, D. Kakwani, S. Golla, et al., "AI4Bharat-IndicNLP corpus: Monolingual corpora and word embeddings for Indic languages," *arXiv preprint arXiv:2005.00085*, 2020.

[6] D. Kakwani et al., "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in *Findings of EMNLP*, 2020, pp. 4948–4961.

[7] G. Ramesh et al., "Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 145–162, 2022. (IndicTrans2 extension: J. Gala et al., "IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages," *arXiv preprint arXiv:2305.16307*, 2023.)

[8] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," in *Proc. ACL*, 2020, pp. 6282–6293.

[9] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proc. ACL*, 2020, pp. 8440–8451.

[10] M. Johnson et al., "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[11] R. Dabre, C. Chu, and A. Kunchukuttan, "A survey of multilingual neural machine translation," *ACM Computing Surveys*, vol. 53, no. 5, art. 99, 2020.

[12] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proc. First Workshop on Neural Machine Translation*, 2017, pp. 28–39.

[13] NLLB Team, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[14] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. ACL*, 2016, pp. 1715–1725.

[15] A. Kunchukuttan and P. Bhattacharyya, "Learning variable length units for SMT between related languages via byte pair encoding," in *Proc. First Workshop on Subword and Character Level Models in NLP*, 2016, pp. 14–24.