Parallel Corpora for Machine Translation in Low-Resource Indic Languages: A Comprehensive Review

Rahul Raja
Carnegie Mellon University
Stanford University
LinkedIn*

Arpita Vats
Boston University
Santa Clara University
LinkedIn*

Abstract

Parallel corpora play an important role in training machine translation (MT) models, particularly for low-resource languages where high-quality bilingual data is scarce. This review provides a comprehensive overview of available parallel corpora for Indic languages, which span diverse linguistic families, scripts, and regional variations. We categorize these corpora into text-to-text, code-switched, and various categories of multimodal datasets, highlighting their significance in the development of robust multilingual MT systems. Beyond resource enumeration, we critically examine the challenges faced in corpus creation, including linguistic diversity, script variation, data scarcity, and the prevalence of informal textual content. We also discuss and evaluate these corpora in various terms such as alignment quality and domain representativeness. Furthermore, we address open challenges such as data imbalance across Indic languages, the trade-off between quality and quantity, and the impact of noisy, informal, and dialectal data on MT performance. Finally, we outline future directions, including leveraging cross-lingual transfer learning, expanding multilingual datasets, and integrating multimodal resources to enhance translation quality. To the best of our knowledge, this paper presents the first comprehensive review of parallel corpora specifically tailored for low-resource Indic languages in the context of machine translation.

1 Introduction

1.1 Importance of parallel corpora

Parallel corpora are collections of texts that contain sentence-aligned translations across two or more languages (Brown et al., 1991). These resources play a fundamental role in machine translation (MT), cross-lingual natural language processing (NLP), and linguistic research. Unlike monolingual corpora, parallel corpora enable direct learning of translation mappings, making them essential for training statistical and neural MT models (Koehn et al., 2020).

Parallel corpora have been crucial in the development of MT models, starting from phrase-based statistical MT (SMT) (Romdhane et al., 2014) to modern neural MT (NMT) approaches (Stahlberg, 2020). In SMT systems, they provided the necessary data for learning phrase alignments and translation probabilities ("Voita and Sennrich). With the rise of transformer-based NMT models (Vaswani et al., 2023), large-scale parallel corpora have become even more critical, as these models rely on extensive aligned data to learn high-quality translation representations.

Beyond MT, parallel corpora are used in cross-lingual NLP tasks such as multilingual word embeddings (Conneau et al., 2020), zero-shot learning (Artetxe and Schwenk, 2019), and multilingual question-answering systems (Hu et al., 2024b). These resources allow models to generalize across languages by leveraging shared semantic representations learned from translation pairs. For low-resource languages, parallel corpora are not just tools for MT but also serve a crucial role in language preservation and revitalization (Hu et al., 2024a). Many Indic languages lack digitized linguistic resources, making them vulnerable to digital extinction. Creating high-quality parallel datasets ensures that these languages remain computationally accessible, enabling future educational tools, digital assistants, and automated translations (Anastasopoulos et al., 2020).

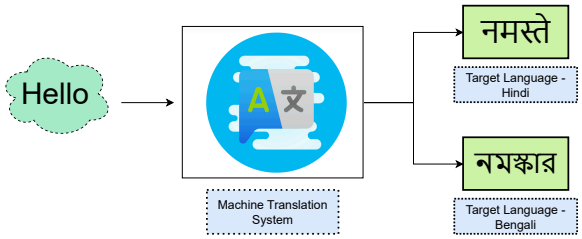

Figure 1: Overview of Machine Translation Model.

1.2 Machine Translation for Indic languages

MT for Indic languages faces numerous challenges due to their linguistic diversity, script variations, and resource constraints (Bala Das et al., 2023a). Unlike high-resource languages, many Indic languages suffer from limited parallel corpora, making it difficult to train robust translation models. The diversity in syntax and phonology across language families further complicates alignment and translation tasks. The presence of mul-

*Work does not relate to position at LinkedIn.

Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025), pages 129–143

tiple scripts and a lack of standardized transliteration mechanisms hinder effective corpus development. Figure 2 shows the overview of challenges in Indic MT

1.2.1 Morphosyntactic Complexity and Linguistic Variability

Indic languages exhibit significant linguistic diversity, primarily categorized into Indo-Aryan and Dravidian language families (Masica, 1993). Indo-Aryan languages, such as Hindi, Bengali, and Marathi, are characterized by inflectional morphology and a relatively flexible subject-object-verb (SOV) (Schouwstra and de Swart, 2014) word order, whereas Dravidian languages, including Tamil, Telugu, and Kannada, employ agglutinative morphology, where words are formed by adding multiple affixes to a root word. These structural differences pose challenges in MT systems, as segmentation strategies that work for one language family may not be effective for another (Rama and Kolachina, 2012). Additionally, phonological distinctions, such as retroflex consonants in Dravidian languages that are absent in many Indo-Aryan languages, complicate speech-to-text and transliteration tasks (Annamalai, 2006).

1.2.2 Multiscript Representation and Orthographic Challenges

Indic languages are written in multiple scripts, which significantly impact corpus creation and text normalization (Manohar et al., 2024), (Hellwig, 2010). For instance, Hindi and Marathi share the Devanagari script, but differences in spelling conventions and phonetic representations require preprocessing before effective alignment (Hellwig, 2010). Bengali and Assamese use the Bengali-Assamese script, while Tamil, Telugu, and Kannada have distinct scripts with unique grapheme-to-phoneme mappings (Gales et al., 2007).We have also created an indic language categorization illustrated in Appendix A. The lack of script standardization introduces inconsistencies in parallel corpora, making text alignment a challenging task. Moreover, the development of optical character recognition (OCR) tools for Indic scripts remains an ongoing challenge, as many scripts have complex ligatures and diacritic variations that reduce OCR accuracy, further limiting the availability of digitized resources for MT (Sengupta et al., 2019).These script-specific challenges underscore the need for robust preprocessing pipelines and script-aware normalization techniques to improve the quality and usability of Indic language corpora.

1.2.3 Data Scarcity and Low-Resource Limitations

The scarcity of high-quality parallel corpora remains a significant obstacle in developing robust MT models for Indic languages (Bala Das et al., 2023b). While languages like Hindi and Bengali have relatively larger corpora, low-resource languages such as Santali, Maithili, and Konkani lack sufficient parallel data, restricting the effectiveness of data-driven MT approaches. The limited availability of bilingual datasets hampers the training of neural MT models, which require vast amounts of parallel text for effective generalization. To mitigate this issue, researchers have explored synthetic data generation techniques such as back-translation and cross-lingual transfer learning. However, these approaches often introduce artifacts that can degrade translation quality, highlighting the need for well-annotated, human-verified corpora to support low-resource Indic language MT (Sengupta et al., 2019).

1.2.4 Register Variability and Linguistic Formality

Most existing parallel corpora for Indic languages are derived from formal sources such as news articles, religious texts, and government documents, which do not capture the informal and conversational aspects of language used in everyday communication (Post et al., 2012). This imbalance affects the performance of MT systems in real-world applications, as they struggle to translate colloquial expressions, dialectal variations, and code-switched text commonly found in social media and user-generated content (Rijhwani et al., 2020). Code-mixing (Khanuja et al., 2020), particularly in Hindi-English and Bengali-English, presents additional challenges, as standard MT models are not optimized for handling intra-sentential language switching (Pratapa et al., 2018). The need for diverse corpora that encompass both formal and informal registers is essential to improve translation accuracy across different linguistic contexts.

2 Parallel Corpora for Indic Languages: Modalities and Comparisons

The development of MT systems for Indic languages heavily relies on the availability of high-quality parallel corpora. These corpora serve as the foundation for training neural MT models, aligning linguistic structures across languages, and enabling multilingual applications such as automatic translation, speech recognition, and multimodal understanding. Given the diverse nature of Indic languages and their applications in different contexts, parallel corpora can be classified into various types based on the modality of data they contain. This section provides an overview of different types of parallel corpora available for Indic languages, their characteristics, and their significance in MT research.

2.1 Text-to-Text Parallel Corpora

Text-to-text parallel corpora form the backbone of MT systems, consisting of bilingual or multilingual sentence-aligned datasets that provide direct translations between languages (de Gibert et al., 2024). These corpora are essential for training supervised MT models and are widely used in both statistical and neural machine translation (NMT) frameworks (Raunak et al., 2024).

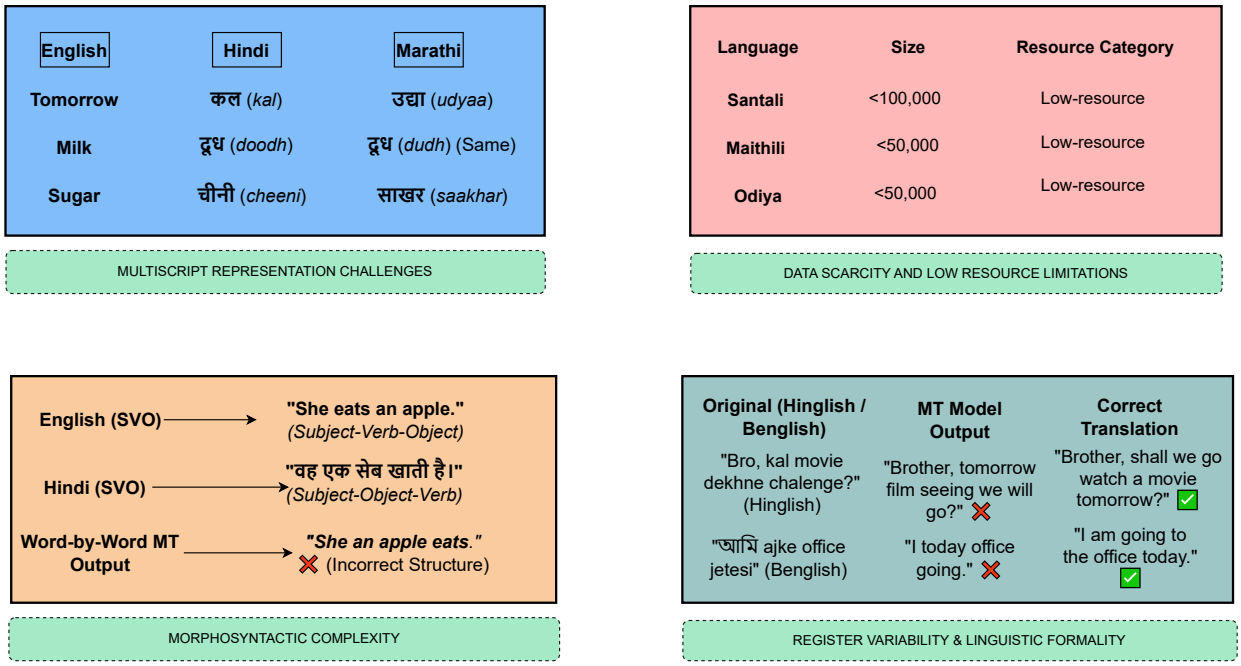

Figure 2: Challenges in Indic Machine Translation: Key issues include morphosyntactic complexity, script variations, low-resource languages, and translation errors in Hinglish and Benglish.

2.1.1 High-Coverage Parallel Corpora

Large-scale parallel corpora play an important role in developing machine translation systems, especially for low-resource languages. We have considered a dataset to be high-coverage or large-scale if it contains more than 10 million sentence pairs, as this volume provides sufficient linguistic diversity and contextual richness for training robust translation models. These corpora serve as the backbone for both statistical and NMT systems, enabling improved generalization, domain adaptation, and cross-lingual transfer learning. BPCC Parallel Corpus (Gala et al., 2023) stands out as the largest, containing 230 million sentence pairs across 22 Indic languages. Its extensive coverage makes it an invaluable resource for multilingual translation tasks, particularly for high-quality English-Indic translations. In comparison, the Samanantar Parallel Corpus (Ramesh et al., 2023), introduced in 2021, includes 46 million sentence pairs between English and 11 Indic languages, along with an additional 82 million sentence pairs between Indic languages. While smaller than BPCC, Samanantar is unique in its extensive Indic-Indic translation pairs, making it highly valuable for intra-Indic translation tasks.

Another notable dataset, CCAligned (El-Kishky et al., 2020), consists of over 100 million document pairs across 137 languages, making it one of the most extensive cross-lingual resources. However, this dataset requires extensive filtering to improve data quality before being used for training MT models. Despite its noisiness, its unparalleled scale and diversity make it useful for large-scale pretraining and domain adaptation. The OPUS corpus (Tiedemann, 2012), widely recognized as a comprehensive multilingual dataset, aggregates multiple parallel corpora across various domains and languages. It covers over 100 million sentence pairs across more than 50 languages, including Hindi, Bengali, Marathi, Tamil, Telugu, and Malayalam, and overlaps significantly with the WAT 2018 Parallel Corpus (Zhang et al., 2020), which contains 10–20 million sentence pairs specifically for Hindi-English and Bengali-English translations. While OPUS provides domain diversity and structured data, it primarily consists of pre-existing datasets, making it less novel compared to BPCC and Samanantar.

Beyond these major corpora, other high-coverage Indic datasets contribute significantly to MT research. The Bhasha Parallel Corpus (Mujadia and Sharma, 2025) includes 44 million sentence pairs across seven Indic languages, supporting cross-lingual and domain adaptation studies. Additionally, the M2M-100 dataset (Fan et al., 2020), developed by Meta AI, contains 12 million sentence pairs spanning over 100 languages, including several Indic languages. This dataset played a key role in the development of the M2M-100 translation model, which enables direct translation between non-English languages.

Several other corpora derived from Wikipedia and Common Crawl have also contributed to large-scale MT training. The WikiMatrix Corpus (Schwenk et al., 2021), developed by Meta, consists of parallel sentences extracted from Wikipedia using LASER-based sentence alignment (Artetxe and Schwenk, 2019). It offers a vast number of sentence pairs across numerous languages, it includes around 3.5 million Hindi-English sentence pairs and around 7 millions of other indic language pairs, making it a valuable resource for training MT

models (Niehues and Waibel, 2011). However, its reliance on Wikipedia content means that the domain of the text is primarily encyclopedic, limiting its applicability for more conversational or domain-specific translations. Similarly, Wikititles (Liu et al., 2017), another Wikipedia-based corpus, extracts bilingual and multilingual article titles from Wikipedia, For Hindi-English, the dataset contains approximately 1.3 million parallel titles. Compared to WikiMatrix, Wikititles provides smaller, well-aligned phrase pairs rather than full sentences, making it particularly useful for training models that focus on short-form content, such as entity names, search queries, or phrase-based MT systems. Due to its structure, Wikititles is less prone to misalignment errors than WikiMatrix but lacks sentence-level parallelism. In contrast, CCMatrix (Schwenk et al., 2021), developed by Meta, is a much larger dataset mined from the CommonCrawl web corpus (Panchenko et al., 2018). It contains billions of parallel sentences across multiple languages, surpassing both WikiMatrix and Wikititles in sheer volume, making it a powerful resource for large-scale NMT training. However, its primary drawback is the noisiness of web-mined content, which often includes misaligned or irrelevant text pairs that require extensive filtering. Among these datasets, WikiMatrix offers a balanced trade-off between scale and accuracy, providing a large yet relatively clean dataset for training MT models, whereas Wikititles ensures precise alignment quality but is limited in scope due to its focus on article titles rather than full sentences. CCMatrix, on the other hand, offers the most extensive collection of parallel sentences but requires aggressive filtering to ensure usability. While BPCC holds the advantage in terms of sheer size and multilingual support, Samanantar's inclusion of Indic-Indic translation pairs makes it particularly valuable for intra-Indic translation tasks. Researchers must carefully analyze these corpora to determine the most suitable dataset for their translation models, ensuring an optimal balance between high-quality curated translations and large-scale mined data. The choice of corpus ultimately depends on the specific needs of an MT system—whether prioritizing size, quality, or domain coverage.

2.2 Low-Coverage Parallel Corpora

While large-scale parallel corpora provide extensive training data for MT, many Indic languages remain low-resource, lacking sufficient parallel data for robust model development. For this study, we define low-resource parallel corpora as datasets containing fewer than 10 million sentence pairs. These datasets are crucial for developing MT models for underrepresented Indic languages, particularly those that lack substantial digital text resources. Despite their smaller size, these corpora serve as valuable benchmarks for fine-tuning, domain adaptation, and zero-shot learning approaches in machine translation.

The IIT Bombay Parallel Corpus (Kunchukuttan et al., 2018) is one of the most widely used low-resource datasets, containing 1.5 million sentence pairs for English-Hindi translation. The dataset is derived from news and government documents, making it well-suited for formal text translation but less effective for conversational and domain-specific tasks. Due to its clean alignment and high-quality translations, it is often used for benchmarking and fine-tuning NMT models for Hindi-English translation. For Bangla-English translation, the `BUETEnglishBanglaCorpus` (Islam et al., 2021) offers 2.7 million sentence pairs, primarily sourced from news articles, books, and religious texts. While smaller than large-scale corpora like BPCC or OPUS, BUET (Islam et al., 2021) is an important resource for Bangla machine translation, particularly for formal and literary domains. The dataset provides high-quality bilingual sentence alignments, making it a valuable resource for training MT models that need precise and domain-specific translations.

For historical and classical language translation, the Itihasa Parallel Corpus (Krishna et al., 2020) is one of the few available datasets, offering 93,000 sentence pairs for English-Sanskrit translation. Given Sanskrit's morphologically rich structure and complex syntax, this dataset provides a rare opportunity for training translation models in ancient and scholarly texts. Due to its small size, MT models trained on Itihasa rely on data augmentation techniques, such as back-translation and transfer learning, to improve performance. A major initiative for low-resource machine translation across multiple Indic languages is the TICO-19 dataset (Anastasopoulos et al., 2020). TICO-19 provides parallel sentence pairs across multiple underrepresented Indic languages, including Maithili, Manipuri, and Sindhi. Unlike many general-purpose corpora, TICO-19 is specifically designed for medical and technical translations, making it highly valuable for domain-specific machine translation models. Given the importance of healthcare communication in multilingual settings, this dataset plays a critical role in enabling low-resource language translation for public health applications.

Another key dataset is NLLB (No Language Left Behind) (Team et al., 2022b), which provides small-scale training data for multiple Indic languages, including Kashmiri, Maithili, and Bhojpuri. The NLLB project is part of Meta AI's initiative to support low-resource language translation, aiming to improve direct translation between non-English language pairs. While individual language pairs in NLLB have limited sentence pairs, the dataset's wide coverage across many underrepresented Indic languages makes it highly useful for zero-shot and few-shot learning applications in MT.

In contrast to large-scale corpora such as BPCC and Samanantar, these low-resource parallel datasets serve as critical benchmarks for low-resource Indic languages, enabling research in domain adaptation, transfer learning, and cross-lingual generalization. Given the scarcity of annotated parallel corpora for many Indic languages, data augmentation, back-translation, and synthetic data generation play a crucial role in improving translation

quality for underrepresented languages.

2.3 Multimodal Corpora

Multimodal corpora extend beyond traditional text-based datasets by incorporating multiple data types, such as text, speech, and visual information, into a unified dataset (Baltrušaitis et al., 2019). These corpora are instrumental in building more comprehensive MT models capable of handling real-world scenarios involving multiple modalities (Li et al., 2020). IndicMultiModal (Kothapalli et al., 2021), for instance, provides text, speech, and image datasets aligned across multiple Indic languages, supporting research in multimodal translation, speech synthesis, and cross-lingual retrieval. Such corpora are particularly beneficial for applications in digital accessibility (Sun et al., 2021), interactive AI assistants.

2.3.1 Speech-to-Text corpora

Speech-to-text align spoken language with its textual translation, making them essential for speech translation and automatic speech recognition (ASR) systems (Jouvet et al., 2019). These datasets are particularly valuable for creating voice-enabled translation models and developing ASR systems for Indic languages (Sitaram et al., 2020).

One of the important corpora in this category is CVIT-IIITH Mann ki Baat Corpus (Philip et al., 2021), mined from Indian Prime Minister Narendra Modi's Mann ki Baat speeches. Since these speeches are carefully prepared and delivered in formal Hindi with occasional English phrases, the dataset is well-suited for studying political speech translation and handling Hindi-English code-switching. However, given its highly structured nature, it may not fully capture the variability of spontaneous speech, which is often a challenge for ASR models. Compared to other datasets, this corpus is more domain-specific, focusing on political communication. A more extensive alternative is the PMIndia corpus, which, while not explicitly speech data, consists of transcriptions of spoken news content into text (Haddow and Kirefu, 2020). It extends beyond Mann ki Baat by providing spoken speech transcriptions with translations across multiple Indian languages. While both PMIndia and Mann ki Baat focus on government-related content, PMIndia includes a broader set of formal speeches, policy discussions, and governance-related material. This makes it valuable for multilingual speech translation systems, though, like Mann ki Baat, its primary limitation is that government discourse follows a standardized linguistic structure, lacking the variation seen in informal conversations or spontaneous speech.

Another dataset designed for a specific domain is the QED Corpus (Lamm et al., 2021). It is a text data which is derived from the video transcripts. It focuses on educational video transcripts in English and Hindi. Which cover a wide range of topics, QED is optimized for academic discourse, including lecture-style content. This makes it particularly beneficial for ASR and translation models targeting online education platforms, academic lectures, and instructional content. QED ensures high-quality transcriptions and translations tailored for educational use cases. QED Corpus fills the gap in academic and instructional content, making it an essential resource for educational applications. The choice of corpus depends on the desired application—whether for structured government communication, spontaneous public discourse, or domain-specific speech processing.

2.4 Text-to-Speech Corpora

Text-to-Speech (TTS) technology plays an important role in enhancing accessibility and language inclusivity by converting textual information into natural-sounding speech. The development of high-quality TTS systems for Indic languages has gained momentum with the availability of large-scale corpora and open-source models.

Two notable contributions in this area are AI4Bharat Indic-TTS (Kumar and S, 2023) and BhasaAnuvaad (Jain et al., 2024), both of which provide extensive linguistic resources to improve speech synthesis and ASR systems. AI4Bharat Indic-TTS is an open-source TTS model that supports 13 Indic languages, including Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Sanskrit, Tamil, Telugu, and Urdu. It is designed to produce high-quality synthetic speech with various speaker styles, making it suitable for applications such as language learning, assistive technologies, and media content creation. The model provides fine-grained control over speech parameters, including pitch, speed, and voice modulation, ensuring natural and expressive speech synthesis. Complementing this, BhasaAnuvaad serves as a comprehensive dataset for speech translation, encompassing over 44,400 hours of speech and 17 million text segments across 13 Indic languages and English. By incorporating both mined and high-quality curated parallel speech data, BhasaAnuvaad is a valuable resource for developing ASR and TTS systems, as it enables sentence-to-audio alignment, facilitating the training of robust speech synthesis models. These initiatives significantly contribute to advancing TTS technology for Indic languages, fostering inclusivity and accessibility for a diverse user base.

2.4.1 Image-to-Text Corpora

Image-to-text corpora helps in advancing multimodal learning, enabling models to understand and generate text based on visual input (Guo et al., 2024). In the context of Indic languages, several high-quality datasets have been developed to support multimodal research, particularly in image captioning, visual question answering (VQA), and image-grounded translation (Özdemir and Akagündüz, 2024). Among the most significant contributions are the Hindi Visual Genome, Bengali Visual Genome, and Malayalam Visual Genome datasets. The Hindi Visual Genome dataset (Parida et al., 2019) contains 31K multimodal pairs aligned in Hindi-English,

providing a rich resource for training models in tasks such as image captioning and cross-lingual understanding. Similarly, the Bengali Visual Genome (Sen et al., 2021) offers 29K multimodal pairs in Bengali-English, while the Malayalam Visual Genome includes 29K multimodal pairs in Malayalam-English. These datasets are designed to bridge the gap in low-resource Indic languages for multimodal AI applications. They are particularly useful in training (Xue et al., 2024) and evaluating MT models, cross-lingual retrieval systems, and vision-language models (VLMs) (Bordes et al., 2024) for Indic languages. These corpora contribute significantly to the development of multimodal AI for Indic languages, facilitating better captioning, improved VQA systems, and enhanced multilingual vision-language applications. By providing a strong benchmark for multimodal learning, they enable robust model training for real-world applications such as automated image description generation and visual assistive technologies in Indian languages.

2.5 Code-Switched Corpora

Code-switching, the practice of mixing two or more languages within a single conversation or sentence, is common in multilingual communities, including those using Indic languages (Garg et al., 2021). Code-switched corpora are crucial for developing translation models that can handle real-world conversations where users frequently switch between languages such as Hindi-English, Bengali-English, and Tamil-English (Bali et al., 2014).

GLUECoS (Khanuja et al., 2020) is a well known code-switched corpus, which contains Hindi-English and Bengali-English code-switched data and also datasets from social media platforms like Twitter and WhatsApp, where code-mixing is prevalent (Sitaram et al., 2019). Training MT systems on such corpora improves their ability to handle informal and conversational text (Winata et al., 2021a). The PHINC (Parallel Hinglish Social Media Code-Mixed Corpus) (Srivastava and Singh, 2020) consists of 13.7k Hinglish (Hindi-English) sentences, making it one of the most comprehensive resources for studying mixed-language usage in digital communication. It is particularly valuable for social media NLP tasks, where speakers often switch between Hindi and English within a single sentence.

Similarly, the IIIT-H en-hi-codemixed-corpus (Dhar et al., 2018) is a code-mixed dataset consisting of 6k English-Hindi sentences. Compared to PHINC, this corpus has a smaller dataset size but higher-quality annotation, ensuring accurate training data for models dealing with Hinglish content. Its focus on token-level annotations makes it especially useful for tasks such as word-level language identification and code-mixed text normalization. The CALCS 2021 Eng-Hinglish dataset (Appicharla et al., 2021) provides 10k parallel sentence pairs, focusing on formal and informal contexts of Hinglish usage. Compared to PHINC and IIIT-H, CALCS is particularly valuable for machine translation between English and Hinglish, helping models bridge the gap between standard English and code-mixed vernacular speech.

3 Evaluation of Parallel Corpora

3.1 Evaluation Metrics

Evaluating parallel corpora for Indic languages requires a combination of automatic evaluation metrics that compare machine-generated translations with human reference translations. Common metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), COMET (Rei et al., 2020), and Translation Edit Rate (TER) (Snover et al., 2006) are frequently used, each offering different perspectives on translation quality. BLEU, which focuses on n-gram precision (Callison-Burch et al., 2006), is a widely used metric but can struggle with Indic languages due to their lexical complexity and flexible word order. For instance, languages like Hindi and Tamil exhibit significant syntactic differences from English, which BLEU may fail to capture adequately. METEOR improves upon BLEU by incorporating recall, synonym matching, and stemming, which makes it more effective for languages with rich morphology and varied word forms, such as Hindi, Bengali, and Telugu (Li et al., 2024). However, like BLEU, METEOR still struggles with capturing semantic meaning and context, which are crucial for languages with high syntactic divergence from English.

To address these limitations, COMET, a newer metric, utilizes neural embeddings and contextualized models to evaluate translations based on semantic similarity and contextual understanding (Sun and Wang, 2024). This makes COMET particularly valuable for Indic languages, where it is essential to capture contextual meaning and semantic equivalence rather than just surface-level n-gram matches (Rei et al., 2020). Despite its advantages, COMET requires pretrained models and significant computational resources, which may not always be feasible for resource-constrained settings (Larionov et al., 2024). Additionally, TER measures the edit distance between the machine-generated translation and the reference (Stanchev et al., 2019). By counting the minimum number of insertions, deletions, or substitutions needed to transform one translation into the other, TER is particularly useful for identifying structural mismatches between languages, especially those with flexible sentence structures, such as Hindi and Tamil (Snover et al., 2006). However, TER focuses on structural alignment rather than semantic accuracy, making it a complementary metric rather than a standalone tool.

3.2 Human-Translated Evaluation Datasets

High-quality, human-translated evaluation datasets are essential for benchmarking machine translation models, ensuring that they are assessed on accurate and reliable translations (Yan et al., 2024). Unlike automatically mined corpora, these datasets are manually curated

by professional translators, making them gold-standard resources for evaluating translation quality across different language pairs. Human-translated corpora are particularly crucial for low-resource languages, where the availability of clean, parallel data is often limited (Haddow et al., 2022).

Several notable human-annotated evaluation datasets have been developed to facilitate rigorous benchmarking of multilingual machine translation systems. Among these, the FLORES-101 dataset (Guzmán et al., 2019), developed by Meta AI, is one of the most comprehensive evaluation resources. It provides human-translated test sets for 101 languages, including 14 Indic languages, making it a critical benchmark for assessing translation models in diverse linguistic settings. Following its success, Meta AI expanded the dataset to FLORES-200 (Guzmán et al., 2019), covering 200 languages, including 24 Indic languages. FLORES-200 represents one of the largest human-annotated evaluation datasets for multilingual translation, allowing researchers to systematically analyze the performance of models across a wide range of linguistic families. Both FLORES-101 and FLORES-200 use n-way parallel translation, meaning each sentence is consistently translated across all supported languages, enabling direct multilingual comparisons. These datasets are highly useful for benchmarking, but due to their relatively small sentence count, they are not suited for large-scale training.

Meta AI had also introduced No Language Left Behind (NLLB) (Team et al., 2022a) a benchmarks for large-scale translation efforts, NLLB-Seed (Team et al., 2022b), a small but valuable human-translated dataset specifically designed for evaluating very low-resource languages. This dataset focuses on five Indian languages—Kashmiri, Manipuri, Maithili, Bhojpuri, and Chhattisgarhi—where high-quality parallel data is scarce. While FLORES-200 provides extensive language coverage, it does not always include languages with very limited training data. NLLB-Seed (Team, 2022) fills this gap by prioritizing data quality and focusing on extremely low-resource languages, ensuring that translation models trained on scarce data sources can still be evaluated effectively. A complementary dataset, NLLB-MD (Multi-Domain) (Team et al., 2022a), extends this effort by providing human-annotated parallel translations across three key domains: news, unscripted informal speech, and health. Unlike FLORES datasets, which contain mostly general-purpose text, NLLB-MD allows for more fine-grained evaluation of translation models across different content types, addressing challenges such as domain adaptation and stylistic variation in machine translation. This makes it a valuable resource for improving translation models that operate in specific fields such as journalism, healthcare, or conversational AI.

3.3 Domain Adaptation and Bias

The usability of parallel corpora for MT is contingent on their domain coverage and representativeness (Labaka et al., 2016). Most of the Indic corpora like samantar, PMIndia exhibit a strong bias toward formal and government-regulated domains, including legislative proceedings, religious scriptures, and legal texts (Khanuja et al., 2020). While these datasets facilitate structured translation tasks, they lack the coverage necessary for informal and domain-specific language variations essential in social media, e-commerce, and medical translations. For example, Specialized corpora, such as TICO-19 (Anastasopoulos et al., 2020), have been curated to enhance healthcare-related translation, significantly improving domain-specific MT performance. Expanding parallel corpora to include informal, code-switched datasets from platforms like Twitter, WhatsApp, and online forums is crucial for improving translation quality in conversational and low-resource settings. Beyond domain generalization, data size and language representation remain pivotal. While large-scale corpora such as Bhasha (Jain et al., 2024) provide substantial bilingual sentence pairs, their distribution skews heavily toward high-resource languages like Hindi and Bengali, leaving low-resource languages such as Santali, Konkani, and Maithili significantly underrepresented (Resnik, 1999). Addressing this imbalance requires techniques such as cross-lingual transfer learning, wherein models pretrained on high-resource Indic languages are fine-tuned on their low-resource counterparts (Lample and Conneau, 2019). Gender bias, particularly in languages like bengali with grammatical gender agreement, often results in incorrect translations of gender-neutral references (Stanovsky et al., 2019). Furthermore, political biases inherent in government-curated corpora such as PMIndia and Tico-19 may introduce ideological skew, influencing translation fidelity. Addressing such biases necessitates adversarial debiasing strategies, including counterfactual translation augmentation, reinforcement learning-based neutralization, and bias-aware adversarial training (Sun et al., 2019). Ensuring fair and inclusive translations mandates continuous evaluation and mitigation of systemic biases, particularly for Indic languages with diverse sociopolitical and linguistic landscapes.

4 Future Directions

The development of parallel corpora for Indic languages has advanced significantly, yet challenges related to data scarcity, domain diversity, and alignment quality persist. Future research must focus on expanding low-resource language coverage, improving domain adaptation, leveraging multimodal data, and enhancing automatic data generation techniques to build more robust MT and NLP systems for Indic languages.

4.1 Expanding Coverage for Low-Resource and Dialectal Variants

Many Indic languages, such as Santali, Bodo, Manipuri, and Konkani, remain underrepresented in existing parallel corpora. Most datasets focus on high-resource

languages like Hindi, Bengali, and Tamil, creating an imbalance that hinders the development of MT models for low-resource languages (Lupascu et al., 2025). Future efforts should prioritize the collection of bilingual and multilingual parallel data from vernacular media, government archives, social media, and oral histories (Guzmán et al., 2019). Crowdsourcing initiatives and community-driven data curation can further help improve linguistic diversity and increase the representation of marginalized languages in NLP applications.

4.2 Enhancing Domain Adaptation and Contextual Alignment

Most existing Indic parallel corpora are domain-specific, with a strong bias toward news, religious texts, and government documents. This limits their applicability in scientific, medical, legal, and conversational domains (Hu et al., 2024b). To improve cross-domain generalization, future work should focus on constructing multi-domain parallel corpora and training domain-adaptive MT models (Dong et al., 2025). Furthermore, context-aware alignment techniques, such as document-level parallel corpora and sentence embedding-based alignment, can enhance semantic consistency and translation fluency across diverse textual genres.

4.3 Leveraging Multimodal and Code-Switched Parallel Data

Multimodal MT, which involves image-text and speech-text parallel corpora, is becoming increasingly relevant for Indic languages. Initial datasets like Hindi Visual Genome and Bengali Visual Genome demonstrate the potential of multimodal learning, but larger, more diverse datasets are needed to improve multimodal translation systems (Sen et al., 2021). Similarly, code-switching is prevalent in Hindi-English, Bengali-English, and Tamil-English interactions, yet parallel code-switched corpora remain scarce. Expanding multimodal and code-switched datasets will enhance MT models' performance in real-world multilingual communication and improve their ability to handle informal language (Winata et al., 2021b).

4.4 Automatic Data Generation

Given the scarcity of human-annotated parallel corpora, automatic data generation techniques such as back-translation, synthetic data augmentation, and parallel data mining have gained prominence (Shu et al., 2024). Back-translation, where monolingual target-language data is translated into the source language using pre-trained models, has been widely used to augment data for low-resource Indic languages (Sennrich et al., 2016). Similarly, parallel sentence mining techniques, such as LASER have been applied to extract sentence-aligned parallel data from large-scale web corpora. Additionally, zero-shot learning and self-supervised approaches can further help bootstrap translation models for languages with minimal parallel data. Future work should focus on refining these techniques to improve alignment accuracy and reduce noise in automatically generated corpora.

5 Conclusion

This paper has provided a comprehensive overview of parallel corpora for Indic languages, emphasizing their role in improving MT performance. While large-scale datasets like BPCC and Samanantar exist, many languages remain underrepresented, necessitating more diverse and high-quality resources. Challenges such as lexical diversity, script variations, and data scarcity require innovative approaches like crowdsourcing, domain-specific text collection, and multimodal resources. Code-switching in digital communication also demands corpora that capture informal and mixed-language text. Automatic data generation techniques like back-translation and parallel sentence mining help augment corpora, but ensuring data quality is critical. Future research should address domain bias, improve evaluation metrics, and expand multimodal and low-resource language coverage. Collaborative efforts between researchers and linguistic communities will be essential in enhancing accessibility and translation accuracy for Indic languages.

6 Limitations

Despite providing a comprehensive review of parallel corpora for low-resource Indic languages in machine translation, several limitations must be acknowledged. The scope of the review is constrained by the availability of datasets, and while we have made an effort to cover a wide range of resources, many low-resource languages remain underrepresented. Additionally, the quality of the corpora varies significantly, with some datasets suffering from issues like inconsistent translations, noisy data, and domain-specific biases, which could limit their applicability in building robust machine translation systems. Moreover, this review focuses primarily on the datasets themselves and does not delve deeply into the models or evaluation metrics employed, which are crucial factors in the effectiveness of any MT system. Finally, access to some corpora may be restricted due to licensing issues, and in some cases, dataset metadata may not be fully available, limiting the depth of evaluation that can be performed. These limitations highlight the need for ongoing research and continuous updates in this evolving field.

References

Antonios Anastasopoulos, Noah Constant, Amir Feder, Dan Garrette, Richard Hatcher, John Hewitt, Zhong Zhou Wang, and Yiming Xu. 2020. Tico-19: The translation initiative for covid-19. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4760–4772.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks

and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

E. Annamalai. 2006. *Language in South Asia*. Cambridge University Press.

Ramakrishna Appicharla, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2021. IITP-MT at CALCS2021: English to Hinglish neural machine translation using unsupervised synthetic code-mixed parallel corpus. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, Online.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Sudhansu Bala Das, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kr. Patra. 2023a. Improving multilingual neural machine translation system for indic languages. 22(6).

Sudhansu Bala Das, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kr. Patra. 2023b. Improving multilingual neural machine translation system for indic languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6).

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. Code-mixing: A challenge for language identification in the indian context. In *Proceedings of the First Workshop on Computational Approaches to Code Switching (EMNLP)*, pages 13–23.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. 2024. An introduction to vision-language modeling. *Preprint*, arXiv:2405.17247.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. ACL '91, page 169–176, USA. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. *Preprint*, arXiv:2403.14009.

Mrinal Dhar, Vaibhav Kumar, and Manish Shrivastava. 2018. Enabling code-mixed translation: Parallel corpus creation and MT augmentation approach. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 131–140, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hao Dong, Moru Liu, Kaiyang Zhou, Eleni Chatzi, Juho Kannala, Cyrill Stachniss, and Olga Fink. 2025. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *Preprint*, arXiv:2501.18592.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint arXiv:2010.11125*.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Mark J. F. Gales, Kate M. Knill, Anton Ragni, and Shakti P. Rath. 2007. Application of grapheme-to-phoneme mappings in speech technology. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4453–4456. IEEE.

Ayush Garg, Anoop Kunchukuttan, and Monojit Choudhury. 2021. Code-switching in indian languages: Linguistic aspects and computational challenges. *Computational Linguistics*, 47(2):285–319.

Ruifeng Guo, Jingxuan Wei, Linzhuang Sun, Bihui Yu, Guiyong Chang, Dawei Liu, Sibo Zhang, Zhengbing Yao, Mingjun Xu, and Liping Bu. 2024. A survey on advancements in image-text multimodal models: From general techniques to biomedical implementations. *Comput. Biol. Medicine*, 178:108709.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, , et al. 2019. The flores evaluation datasets for low-resource machine translation: Benchmarking progress in many-to-many translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2246–2251.

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.

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india. *Preprint*, arXiv:2001.09907.

Oliver Hellwig. 2010. The interaction of scripts and languages in south asia. *Written Language & Literacy*, 13(1):62–85.

Jia Cheng Hu, Roberto Cavicchioli, Giulia Berardinelli, and Alessandro Capotondi. 2024a. Learning from wrong predictions in low-resource neural machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10263–10273, Torino, Italia. ELRA and ICCL.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Dong Zhang, Zhehuai Chen, and Eng Siong Chng. 2024b. Gentranslate: Large language models are generative multilingual speech and machine translators. *Preprint*, arXiv:2402.06894.

Md. Islam et al. 2021. Buet english-bangla parallel corpus for neural machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(4):1–15.

Sparsh Jain, Ashwin Sankar, Devilal Choudhary, Dhairya Suman, Nikhil Narasimhan, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, and Raj Dabre. 2024. Bhasaanuvaad: A speech translation dataset for 14 indian languages. *arXiv preprint arXiv: 2411.04699*.

Denis Jouvet, Martine Adda-Decker, and Laurent Besacier. 2019. Asr for under-resourced languages: A survey. In *Proceedings of Interspeech*, pages 160–164.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, and Sunayana Sitaram. 2020. Gluecos: An evaluation benchmark for code-switched nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3575–3585.

Philipp Koehn et al. 2020. Mining parallel data for low-resource machine translation. In *Proceedings of EMNLP 2020*.

Ravi Kothapalli, Anurag Sharma, and Sandeep Subramaniam. 2021. Indicmultimodal: A multimodal dataset for indic language processing. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 3121–3130.

G. Krishna et al. 2020. Itihasa parallel corpus: A large-scale english-sanskrit parallel dataset. *arXiv preprint arXiv:2006.04585*.

Gokul Karthik Kumar and Praveen S. 2023. Towards building text-to-speech systems for the next billion users. *Preprint*, arXiv:2211.09536.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gorka Labaka, Iñaki Alegria, and Kepa Sarasola. 2016. Domain adaptation in MT using titles in Wikipedia as a parallel corpus: Resources and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2209–2213, Portorož, Slovenia. European Language Resources Association (ELRA).

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2021. Qed: A framework and dataset for explanations in question answering. *Transactions of the Association for Computational Linguistics*, 9:790–806.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 7057–7067.

Daniil Larionov, Mikhail Seleznyov, Vasiliy Viskov, Alexander Panchenko, and Steffen Eger. 2024. xcomet-lite: Bridging the gap between efficiency and quality in learned mt evaluation metrics. *Preprint*, arXiv:2406.14553.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. Culturepark: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Xutai Li, Zhe Yao, Yihan Zhang, et al. 2020. Vivo: Visual vocabulary pre-training for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3478–3485.

Frederick Liu, Han Lu, Chieh Lo, and Graham Neubig. 2017. Learning character-level compositionality with visual features. *Preprint*, arXiv:1704.04859.

Marian Lupascu, Ana-Cristina Rogoz, Mihai Sorin Stupariu, and Radu Tudor Ionescu. 2025. Large multimodal models for low-resource languages: A survey. *Preprint*, arXiv:2502.05568.

Kavya Manohar, Leena G Pillai, and Elizabeth Sherly. 2024. What is lost in normalization? exploring pitfalls in multilingual asr model evaluations. *Preprint*, arXiv:2409.02449.

Colin P. Masica. 1993. *The Indo-Aryan Languages*. Cambridge University Press, Cambridge, UK.

Vandan Mujadia and Dipti Misra Sharma. 2025. Bhashaverse : Translation ecosystem for indian subcontinent languages. *Preprint*, arXiv:2412.04351.

Jan Niehues and Alex Waibel. 2011. Using Wikipedia to translate domain-specific terms in SMT. In *Proceedings of the 8th International Workshop on Spoken Language Translation: Papers*, pages 230–237, San Francisco, California.

Övgü Özdemir and Erdem Akagündüz. 2024. Enhancing visual question answering through question-driven image captions as prompts. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1562–1571.

Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone Paolo Ponzetto, and Chris Biemann. 2018. Building a web-scale dependency-parsed corpus from commoncrawl. *Preprint*, arXiv:1710.01779.

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.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multimodal english-to-hindi machine translation. *Preprint*, arXiv:1907.08948.

Jerin Philip, Shashank Siripragada, Vinay P. Namboodiri, and C. V. Jawahar. 2021. Revisiting low resource status of indian languages in machine translation. CODS COMAD 2021, page 178–187. ACM.

Matt Post, Chris Callison-Burch, and Sanjeev Khudanpur. 2012. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT)*, pages 401–409.

Adithya Pratapa, Monojit Choudhury, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory-based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1543–1553.

Taraka Rama and Sudheer Kolachina. 2012. Morphological complexity of dravidian languages, as measured by prefix, suffix, and infix counts. *Linguistic Typology*, 16(3):453–482.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2023. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Preprint*, arXiv:2104.05596.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Vikas Raunak, Roman Grundkiewicz, and Marcin Junczys-Dowmunt. 2024. On instruction-finetuning neural machine translation models. *Preprint*, arXiv:2410.05553.

Ricardo Rei, José G. C. Teixeira, Tiago Coelho, Telmo Pires, and André F. T. Martins. 2020. COMET: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702.

Philip Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534.

Shruti Rijhwani, Kalika Bali, and Monojit Choudhury. 2020. Handling multilinguality in low-resource mt: The case of indian languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1532–1543.

Achraf Ben Romdhane, Salma Jamoussi, Abdelmajid Ben Hamadou, and Kamel Smaïli. 2014. Phrase-based language modelling for statistical machine translation. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 96–99, Lake Tahoe, California.

Marieke Schouwstra and Henriëtte de Swart. 2014. The semantic origins of word order. *Cognition*, 131(3):431–436.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondrej Bojar, and Satya Ranjan Dash. 2021. Bengali visual genome: A multimodal dataset for machine translation and image captioning. In *International Conference on Frontiers in Intelligent Computing: Theory and Applications*.

Anirban Sengupta, Sudip Ghosh, and Abhishek Basu. 2019. Challenges in ocr for indic scripts: A survey and case study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1807–1819.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96.

Peng Shu, Junhao Chen, Zhengliang Liu, Hui Wang, Zihao Wu, Tianyang Zhong, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Yifan Zhou, Constance Owl, Xiaoming Zhai, Ninghao Liu, Claudio Saunt, and Tianming Liu. 2024. Transcending language boundaries: Harnessing llms for low-resource language translation. *Preprint*, arXiv:2411.11295.

Sunayana Sitaram, Kalika Bali, Monojit Choudhury, and Alan W Black. 2019. A survey of code-switched speech and language processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 73–94.

Sunayana Sitaram, Sandeep Kar, and Saurabh Taneja. 2020. Asr challenges for indian languages: A survey. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 5432–5438.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231.

Vivek Srivastava and Mayank Singh. 2020. PHINC: A parallel Hinglish social media code-mixed corpus for machine translation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 41–49, Online. Association for Computational Linguistics.

Felix Stahlberg. 2020. Neural machine translation: A review and survey. *Preprint*, arXiv:1912.02047.

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. EED: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 514–520, Florence, Italy. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684.

Kun Sun and Rong Wang. 2024. Textual similarity as a key metric in machine translation quality estimation. *Preprint*, arXiv:2406.07440.

Tony Sun, Alex Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640.

Xuan Sun, Wei Zhang, and Jialiang Liu. 2021. Enhancing digital accessibility with multimodal translation models. *ACM Transactions on Accessible Computing (TACCESS)*, 14(3):1–18.

NLLB Team. 2022. Nllb-seed: Human-translated parallel corpora for low-resource languages. `https://github.com/facebookresearch/flores/tree/main/nllb_seed`. Accessed: 2025-02-10.

NLLB Team, Marta R. Costa-jussà, and James Cross. 2022a. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

NLLB Team, Marta R. Costa-jussà, James Cross, and Onur Çelebi. 2022b. No language left behind: Scaling human-centered machine translation.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Elena "Voita and Sennrich. Language modeling, lexical translation, reordering: The training process of NMT through the lens of classical SMT. In *"Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing"*, pages "8478–8491".

Genta Indra Winata, Zihan Lin, Jamin Shin, and Pascale Fung. 2021a. Multilingual code-switching for zero-shot cross-lingual intent prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2920–2930.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2021b. Multimodal code-switching language modeling with visual grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1893–1905. Association for Computational Linguistics.

Dizhan Xue, Shengsheng Qian, and Changsheng Xu. 2024. Few-shot multimodal explanation for visual question answering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24,

page 1875–1884, New York, NY, USA. Association for Computing Machinery.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Jing Li, Xianchao Zhu, and Yue Zhang. 2024. Benchmarking gpt-4 against human translators: A comprehensive evaluation across languages, domains, and expertise levels. *Preprint*, arXiv:2411.13775.

Boliang Zhang, Ajay Nagesh, and Kevin Knight. 2020. Parallel corpus filtering via pre-trained language models. *Preprint*, arXiv:2005.06166.

A Appendix

Corpus Name	Corpus Type	Sentence Pairs	Languages	Link
BPCC	Text-to-Text	230M	English-22 Indic languages	BPCC
Samanantar	Text-to-Text	46M En-IL, 82M IL-IL	11 Indic languages	Samanantar
IIT Bombay	Text-to-Text	1.5M	English-Hindi	IIT Bombay
CVIT-IIITH PIB	Text-to-Text	N/A	Several Indic languages	CVIT-IIITH PIB
OPUS	Text-to-Text	100M+	50+ languages, several Indic	OPUS
WAT 2018	Text-to-Text	10-20M+	Hindi-English, Bengali-English	WAT 2018
CCAligned	Text-to-Text	100M+	137 languages	CCAligned
Bhasha	Text-to-Text	44M+ rows	7 Indic languages	Bhasha
M2M-100	Text-to-Text	12M+	100+ languages	M2M-100
Itihasa	Text-to-Text	93K	English-Sanskrit	Itihasa Corpus
BUET Eng-Bn Corpus	Text-to-Text	2.7M	English-Bangla	BUET Corpus
Sanskrit-Hindi-MT	Text-to-Text	N/A	Sanskrit-English, Sanskrit-Hindi	Sanskrit-Hindi MT
Kangri Corpus	Text-to-Text	27,362	Hindi-Kangri	Kangri Corpus
MTEnglish2Odia	Text-to-Text	42K	English-Odia	MTEnglish2Odia
IndoWordNet	Text-to-Text	6.3M	18 Indic languages	IndoWordNet Corpus
NLLB Seed	Text-to-Text	N/A	Kashmiri, Maithili, Bhojpuri	NLLB Seed
PHINC	Text-to-Text	13,738	Hindi-English Code-mixed	PHINC
NLLB MD	Text-to-Text	9000+	Bhojpuri	NLLB-MD
PMIndia	Text-to-Text	N/A	Hindi-English	PMIndia
QED	Text-to-Text	43K	English-Hindi	QED Corpus
CoPara	Text-to-Text	2.5K passage pairs	4 Dravidian languages	CoPara
Uka Tarsadia	Text-to-Text	65K	English-Gujarati	Uka Tarsadia
TICO 19	Text-to-Text	N/A	Multiple indic languages	TICO 19
BhasaAnuvaad	Speech<->Text	44,400+ hrs	13 Indic languages	BhasaAnuvaad
Mann ki Baat	Speech<->Text	N/A	Hindi	Mann ki Baat
IndicTTS	Speech<->Text	100+ hrs/language	7 Indic languages	IndicTTS
GLUECoS	Code-Switched	8K-22K	Hindi-English Code-mixed	GLUECoS
PHINC	Code-Switched	13,738	Hindi-English Code-mixed	PHINC
IIIT-H en-hi-codemixed	Code-Switched	6K	English-Hindi	N/A
CALCS 2021	Code-Switched	10K	English-Hinglish	CALCS 2021
Hi Visual Genome	Multimodal	31K	Hindi-English	Hindi Visual Genome
Bn Visual Genome	Multimodal	29K	Bengali-English	Bengali Visual Genome
Ml Visual Genome	Multimodal	29K	Malayalam-English	Malayalam Visual Genome

Table 1: Comprehensive overview of major parallel corpora available for Indic languages, spanning various modalities including text-to-text, speech-to-text, code-switched, and multimodal datasets. The table highlights the corpus type, number of sentence pairs or duration (where applicable), supported language pairs (with a focus on English-Indic and intra-Indic combinations), and links to official sources for access. These resources play a vital role in enabling research in machine translation, multilingual NLP, and low-resource language processing across the Indic language spectrum.

Script	Languages	Region	Script Family
Devanagari	Hindi, Marathi, Sanskrit, Nepali, Konkani, Maithili, Bhojpuri, Sindhi	North, Central India, Nepal	Brahmic
Bengali	Bengali, Assamese, Sylheti, Bodo	Eastern India, Bangladesh	Brahmic
Sharada	Kashmiri (historical script)	Kashmir (historical)	Brahmic
Gurmukhi	Punjabi	Punjab (India and Pakistan)	Brahmic
Gujarati	Gujarati	Gujarat, Daman and Diu	Brahmic
Odia	Odia	Odisha	Brahmic
Grantha	Tamil (Sanskrit texts), Kannada	Tamil Nadu, Karnataka (historical)	Brahmic
Tamil	Tamil	Tamil Nadu, Sri Lanka, Singapore	Brahmic
Telugu	Telugu	Andhra Pradesh, Telangana	Brahmic
Kannada	Kannada	Karnataka	Brahmic
Malayalam	Malayalam	Kerala	Brahmic
Urdu (Arabic script)	Urdu, Kashmiri, Dakhini	North India, Pakistan, Kashmir	Arabic
Arabic	Arabic, Sindhi	North India, Pakistan, Jammu	Kashmir
Tibetan	Tibetan (spoken in Ladakh, Sikkim)	Ladakh, Sikkim, Tibet	Tibetic
Meitei Mayek	Manipuri	Manipur	Brahmic
Brahmi	Ancient Indian texts, Prakrits, early Sanskrit	Pan-Indian (historical)	Brahmic
Sinhala	Sinhala	Sri Lanka (but used by Tamil diaspora in India)	Brahmic
Lepcha	Lepcha	Sikkim, Darjeeling	Brahmic
Limbu	Limbu	Sikkim, Darjeeling, eastern Nepal	Brahmic
Tirhuta	Maithili	Bihar, Nepal	Brahmic
Kaithi	Hindi (historical script)	Bihar, Uttar Pradesh	Brahmic
Sylheti Nagari	Sylheti	Bangladesh, India (Assam)	Brahmic
Chakma	Chakma	Chittagong Hill Tracts (Bangladesh)	Brahmic
Burmese	Burmese	Myanmar, parts of India (Mizoram)	Burmese
Thai	Thai	Thailand (historical influence in India)	Thai
Khmer	Khmer (used historically in Southeast India)	Cambodia, some historical presence in India	Khmer

Table 2: Comprehensive overview of the diverse scripts used across Indic languages, categorized by associated languages, geographic regions, and their respective script families. This table, referenced in Section 1.2.2, highlights both modern and historical scripts, including Brahmic-derived scripts (e.g., Devanagari, Tamil, Bengali), Perso-Arabic adaptations (e.g., Urdu, Kashmiri), and lesser-known indigenous scripts (e.g., Meitei Mayek, Lepcha, Chakma). The representation illustrates the linguistic diversity and orthographic complexity of the Indian subcontinent—factors that critically affect text normalization, OCR development, and

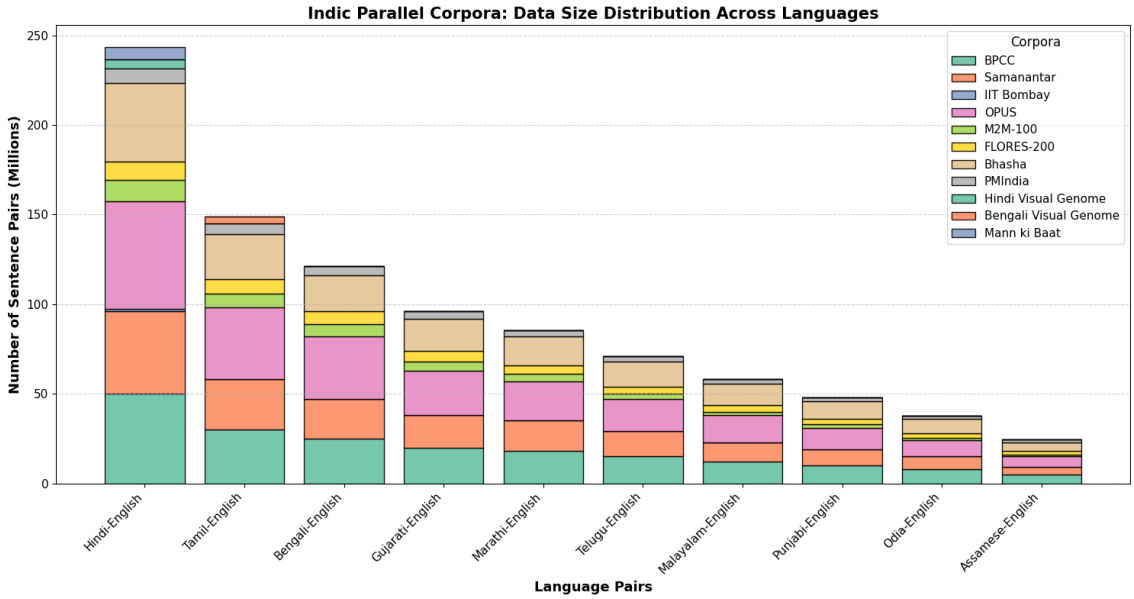

Figure 3: Indic Parallel Corpora: Data Size Distribution Across Languages. A stacked bar chart showing the number of sentence pairs (in millions) for various Indic-English language pairs across major corpora. Languages like Hindi, Tamil, and Bengali are well-resourced, while others such as Assamese and Odia have significantly less data. This highlights the data imbalance in Indic NLP and the need for better resource coverage.