Findings of WMT 2025 shared task on Low-resource Indic Languages Translation

Partha Pakray NIT Silchar
Reddi Mohana Krishna NIT Silchar
Santanu Pal Wipro AI Lab45
Advaitha Vetagiri NIT Silchar
Sandeep Kumar Dash NIT Mizoram
Arnab Kumar Maji North-Eastern Hill University
Saralin A. Lyngdoh North-Eastern Hill University
Lenin Laitonjam NIT Mizoram
Anupam Jamatia NIT Agartala
Koj Sambyo NIT Arunachal Pradesh
Ajit Das Bodoland University
Riyanka Manna Amrita Vishwa Vidyapeetham Amaravati

Abstract

This study proposes the results of the low-resource Indic language translation task organized in collaboration with the Tenth Conference on Machine Translation (WMT) 2025. In this workshop, participants were required to build and develop machine translation models for the seven language pairs, which were categorized into two categories. Category 1 is moderate training data available in languages i.e English–Assamese, English–Mizo, English-Khasi, English–Manipuri and English–Nyishi. Category 2 has very limited training data available in languages, i.e English–Bodo and English–Kokborok. This task leverages the enriched IndicNE-corp1.0 dataset, which consists of an extensive collection of parallel and monilingual corpora for north eastern Indic languages. The participant results were evaluated using automatic machine translation metrics, including BLEU, TER, ROUGE-L, ChrF, and METEOR. Along with those metrics, this year's work also includes Cosine similarity for evaluation, which captures the semantic representation of the sentence to measure the performance and accuracy of the models. This work aims to promote innovation and advancements in low-resource Indic languages.

1 Introduction

The Indic MT Shared Task, first organized alongside the Eighth Conference on Machine Translation (WMT) 2023[1] (Pal et al., 2023), demonstrated the critical need for sustained research attention toward low-resourced languages. That inaugural effort not only revealed the untapped potential for advancing machine translation in these linguistic contexts but also provided a robust methodological and collaborative foundation for future work. Motivated by its impact, the task was expanded and refined in the Ninth Conference on Machine Translation (WMT) 2024[2] (Pakray et al., 2024), drawing broader participation and richer system diversity. Building upon these successive advancements, the 2025 edition[3] has emerged as the most successful iteration to date—surpassing previous years in both scale and the quality of contributions—further cementing the task's role as a driving force in low-resource MT research.

India's linguistic landscape is remarkably diverse, encompassing hundreds of languages spoken across its regions. While 22 languages are officially recognized under the Eighth Schedule of the Indian Constitution and receive governmental support in terms of infrastructure, research, and funding, many others—often spoken by indigenous and minority communities—remain excluded from such provisions. These low-resource languages frequently lack standardized orthographies, adequate lexical resources (e.g., corpora), and formal linguistic documentation. Limited institutional support, declining intergenerational transmission, and minimal access to modern technologies further threaten their preservation and vitality.

To address these challenges, our initiative is dedicated to revitalizing and documenting low-resource Indic languages through targeted, technology-driven solutions. Building upon the success of the Indic MT Shared Tasks at WMT 2023 and WMT 2024, which focused on four language pairs (En-

[1] https://www2.statmt.org/wmt23/indic-mt-task.html

[2] https://www2.statmt.org/wmt24/indic-mt-task.html

[3] https://www2.statmt.org/wmt25/indic-mt-task.html

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 532–553
November 8–9, 2025.

glish–Assamese, English–Mizo, English–Khasi, and English–Manipuri) using the enriched IndicNE-Corp1.0 dataset, we have expanded the scope in 2025 to seven language pairs. These are divided into two categories: (1) Languages with moderate amounts of training data: English–Assamese, English–Mizo, English–Khasi, English–Manipuri, and English–Nyishi. (2) Languages with very limited training data: English–Bodo and English–Kokborok.

This year's task, which has already surpassed prior editions in scale and participation, continues to drive innovation in machine translation and NLP, developing solutions specifically adapted to the unique linguistic and resource constraints of low-resource Indic languages. The Indic MT Shared Task initiative is also committed to safeguarding India's rich linguistic diversity and cultural heritage by strengthening the rights and identities of minority language communities. Leveraging state-of-the-art technologies, it seeks to advance the capabilities of low-resource Indic languages, enabling them not only to survive but to flourish within today's increasingly digital and interconnected landscape.

The task is evaluated using a wide range of metrics, integrating automatic lexical evaluation metrics such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and ChrF (Popović, 2015). In addition to the standard evaluation metrics, we used cosine similarity using Sentence-BERT(all-mpnet-base-v2) (Reimers and Gurevych, 2019) to compute the semantic similarity between the predicted and reference sentences in the English language direction. By combining traditional and semantic evaluation methods, this approach provides a thorough assessment of the performance and accuracy of translation systems.

2 Languages

This section is divided into two parts based on the availability of training data: Category 1 includes languages with moderate resources, and Category 2 includes languages with very limited resources. In the WMT 2024 edition of the shared task, the focus was limited to four languages: Assamese, Mizo, Khasi, and Manipuri. Building upon that foundation, the WMT 2025 task expanded the linguistic coverage by including three additional languages: Nyishi, Bodo, and Kokborok. This expansion reflects our ongoing commitment to improving the representation of under-resourced languages in machine translation research and to broadening the scope of technological inclusion for more linguistically marginalised communities.

2.1 Category 1: (Moderate Training Data Available)

2.1.1 The Assamese Language

Assamese, a member of the Indo-Aryan(Wikipedia contributors, 2025a) branch of the Indo-European language family, is primarily spoken in the northeastern Indian state of Assam. It holds official status in the state and functions as a vital lingua franca, bridging communication across the region's diverse ethnic communities. Recognized as one of India's 22 scheduled languages, Assamese occupies a prominent position within the nation's multilingual framework.

With literary roots dating back to the early medieval era, Assamese(Mahanta, 2012) boasts a long-standing and vibrant cultural heritage. The script currently used for the language evolved from the ancient Brahmi script, reflecting centuries of historical development. However, in the contemporary digital age, Assamese confronts a number of challenges, particularly with regard to language technology. The creation and advancement of computational resources for Assamese are essential, not only to support its practical use in modern contexts but also to ensure its continued vitality and preservation in the face of rapid technological change.

2.1.2 The Mizo Language

Mizo, belonging to the Tibeto-Burman (Wikipedia contributors, 2025d) branch of the Sino-Tibetan language family, is primarily spoken in the northeastern Indian state of Mizoram. It serves as the main medium of communication among the Mizo community and is also used by various ethnic groups across neighbouring regions such as Manipur, Tripura, Assam, and even in parts of Myanmar and Bangladesh. Known for its tonal structure and distinct phonological characteristics, Mizo stands out as a linguistically unique language within the Tibeto-Burman group.

The Mizo language (Zothanliana, 2021) is deeply rooted in a vibrant oral tradition that encompasses folklore, songs, and storytelling forms of expression that preserve and reflect the community's cultural identity. The written form of the language began to take shape in the late 19th century, when Christian missionaries

introduced the Roman script, enabling systematic transcription and laying the groundwork for written literature. Since then, Mizo has developed a strong literary presence, with works spanning traditional poetry to contemporary prose. Nonetheless, like many regional languages, Mizo faces significant challenges in the digital era, especially concerning its representation in language technology and digital communication platforms. Addressing these issues is crucial for both the preservation and advancement of the language.

2.1.3 The Khasi Language

Khasi, a language of the Austroasiatic family(Wikipedia contributors, 2025c), is predominantly spoken in the central and eastern regions (Rynjah and Lyngdoh, 2023) of Meghalaya. Prior to 1813, Khasi lacked a writing system. Between 1813 and 1814, the Bengali script was adopted for translating the Bible into Khasi, a decision influenced by the relatively high literacy in Bengali during that period. By 1816, translated excerpts from the Gospel of Matthew had been printed and circulated among Khasi speakers proficient in Bengali. However, a significant shift occurred in 1841 with the arrival of a Welsh missionary, who introduced the Roman script. This led to translations being made in the Sohra (Cherra) dialect, which later became the basis for standardized written Khasi.

Khasi is marked by notable dialectal diversity. Grierson(Grierson, 1903) identified four primary dialects: Standard Khasi, Pnar (or Synteng), Lyngngam, and War. Acharya (Acharya, 1971) reaffirmed this classification, adding that additional sub dialects such as Bhoi, spoken in Meghalaya's northern plains, also exist. Expanding on these observations, Bareh(Bareh, 1977) provides a detailed account of Khasi dialects, classified primarily by their geographical distribution:

- Amwi (southern Jaiñtia hills),
- Shella and Warding (southern Khasi hills),
- Myriaw, Nongkhlaw, Nongspung, Maram, and Mawiang (mid-western Khasi hills),
- Cherra (mid-southern Khasi hills),
- Mylliem, Laitlyngkot, Nongkrem, and Lyniong-Khasi (central Khasi hills),
- Jowai (central Jaintia hills),
- Bhoi (northeastern Khasi hills),
- Manar, Nongwah, and Jirang (northern Khasi hills),
- Khatarblang/Mawpran (mid-southern region),
- Nongstoin and Langrin (western Khasi hills).

Bareh also emphasizes that phonological variation exists within these groups. Among them, Amwi is considered a particularly distinct and conservative dialect. It is often viewed as more agglutinative and less intelligible to speakers of related dialects such as Jowai or Khatarblang. Despite its unique structure, Amwi speakers can generally understand and use neighbouring dialects for communication. Its linguistic features suggest strong retention of Mon-Khmer elements, especially in morphology and phonology.

Bareh (1977) organizes Khasi dialects into three broad branches:

1. **Eastern Dialects**:
   - Jowai (Central Highlands),
   - Amwi and the War dialects (southern region),
   - Bhoi Synteng (northern region).
2. **Central Dialects**:
   - Dialects such as Nongphlang, Cherra, Nongkrem, Mylliem, Nongspung, and others,
   - Northern varieties like Bhoi East (e.g., Mawrong, Bhoi Lymbong) and Bhoi West (e.g., Manar),
   - War Shala and Warding (southern region).
3. **Western Dialects**:
   - Nongstoin, Lyngam, and Langrin.

Within each branch, sub-dialects often display considerable variation, particularly in phonological patterns. Daladier(Daladier, 2002) notes that Khasi is a part of the Mon-Khmer subgroup of the Austroasiatic family retains conservative unwritten forms, especially in the War dialect areas. Pnar and War remain among the most prominent dialects, with War further subdivided into Nongtalang, Amvi, Tremblang, and Shella. The internal

classification of Pnar remains relatively unexplored, War dialects are further sub divided into War-Khasi and War-Jaintia, spoken in the southern regions of the Khasi and Jaiñtia hills.Grierson also provides foundational work on these dialects.

For the purposes of this shared task, we adopt the Sohra (Cherra) dialect as the standardized form of Khasi for translation. Its historical significance, combined with its widespread use in education, religion, and official discourse, makes it a practical and consistent choice. The formal adoption of the Roman script in 1841 further reinforced its position as the standard written variant, ensuring accessibility for both native speakers and learners across the Khasi-speaking population.

2.1.4 The Manipuri Language

Manipuri, also known as Meiteilon, is a Sino-Tibetan language primarily spoken in the north-eastern Indian state of Manipur. It holds official status as one of the 22 scheduled languages of India and functions as a common medium of communication among diverse ethnic communities in the region, thereby playing a key role in facilitating both social interaction and cultural integration.

The language is marked by a rich literary tradition, with historical records indicating the use of written texts since ancient times. Manipuri employs both the indigenous Meitei Mayek script and, more recently, the Bengali script for written communication. Despite its cultural and historical significance, the language faces notable challenges related to preservation and modernization, particularly within the context of technological development and digital communication. The advancement of computational tools and linguistic resources is critical for ensuring the sustained vitality and broader accessibility of Manipuri in the digital age. In recent years, there has been an increasing academic and technological interest in developing natural language processing (NLP) tools tailored to low-resource languages, including Manipuri (Allen, 2003). However, several persistent issues continue to hinder progress (Gyanendro Singh et al., 2016). Chief among these is the scarcity of annotated corpora and high-quality linguistic datasets, which are crucial for training effective machine learning models. This lack of data significantly constrains the development of key NLP applications such as machine translation (Pal et al., 2023), sentiment analysis (Singh and Singh, 2017), and automatic speech recognition (Gyanendro Singh et al., 2016). Another major obstacle lies in the linguistic complexity of Manipuri itself. The language features a highly inflectional morphological system, posing difficulties for standard NLP models, which are typically optimized for morphologically simpler and better-resourced languages like English. Additionally, issues of script representation and lack of digital standardization complicate text processing, as existing tools often struggle with script conversion, normalization, and consistency across platforms.

Ongoing research is working to mitigate these challenges by building foundational linguistic resources, designing language-specific processing algorithms, and modifying existing NLP architectures to better accommodate the structural characteristics of Manipuri. Despite these promising developments, there remains a substantial gap between Manipuri and more digitally privileged languages a gap that must be addressed through sustained linguistic, technological, and policy-driven efforts.

2.1.5 The Nyishi Language

Nyishi(Kakum et al., 2023; Wikipedia contributors, 2025e), also known as Nyising, Leil, Aya, Nisi, Bangni-Bangru, or Akang, is a Tani language within the Sino-Tibetan family, spoken across eight districts in Arunachal Pradesh. It exhibits distinct typological features, including tonal phonology and context-dependent semantics.

The language uses a modified Roman script comprising seven vowels, eighteen consonants, consonant clusters, a glottal component, and a semivowel. Nyishi is tonal, employing rising, falling, and level tones to distinguish meaning an essential feature in its monosyllabic lexical system.

Syntactically, Nyishi follows a default Subject–Object–Verb (SOV) word order, with occasional Subject–Verb–Object (SVO) patterns. It blends isolating and agglutinative morphological traits. Words often carry multiple meanings based on context for instance, *taxy* may mean “squirrel”, “ginger”, or “animal lice”.

Gender is marked via suffixes rather than noun inflection, using forms like *kibu* (male dog) and *kine* (female dog). There are no plural markers or verb agreement for person or number. Negation is expressed uniformly through the particle *ma*.

Despite its cultural and linguistic value, Nyishi remains underrepresented in technological and computational domains. Expanding NLP efforts

and developing language resources are crucial for its digital preservation and broader accessibility.

2.2 Category 2: (Very Limited Training Data)

2.2.1 The Bodo Language

Bodo, also referred to as Boro (Wikipedia contributors, 2025b), is a member of the Bodo-Garo subgroup within the Tibeto-Burman branch of the Sino-Tibetan language family. It is primarily spoken by the Bodo people in the Bodoland Territorial Region (BTR) of Assam, India, with additional speaker communities in neighboring states such as Arunachal Pradesh, Meghalaya, and Nagaland, as well as in parts of Nepal and Bangladesh. According to the 2011 Census of India, the language is spoken by approximately 1.4 million people.

Bodo holds official recognition as one of the 22 scheduled languages of India(Bhattacharya, 1977), having been included in the Eighth Schedule of the Constitution in 2003. It also enjoys the status of an associate official language in Assam and is used as a medium of instruction in educational institutions within the BTR. The language was historically written in the Latin and Assamese scripts, but since 1975, the Devanagari script has been officially adopted.

Linguistically, Bodo is characterized as a tonal and agglutinative language. Its tonal system assigns semantic distinctions based on pitch, while its morphology supports the formation of complex words through the use of multiple affixes. The syntactic structure typically follows a Subject-Object-Verb (SOV) (Pathak et al., 2025) order, aligning with patterns common to Tibeto-Burman languages.

The Bodo literary tradition has developed (Boro, 2021)significantly in the modern period, particularly following the establishment of the Bodo Sahitya Sabha in 1952, which has been central to efforts in language standardization, publication, and literary development. Today, Bodo literature encompasses diverse genres, including poetry, fiction, and drama, reflecting the sociocultural life of its speakers.

Despite its official status and growing corpus, Bodo remains relatively under-resourced in the digital and computational linguistic domains. Continued initiatives in documentation, corpus building, and NLP development are essential to ensure its sustained vitality and technological integration.

2.2.2 The Kokborok Language

Kokborok, also known as Tripuri, is a Tibeto-Burman language belonging to the Bodo-Garo subgroup, primarily spoken in the Indian state of Tripura and parts of southern Assam, Mizoram, and the Chittagong Hill Tracts of Bangladesh (Debbarma et al., 2012; Nagaraja, 2015). It serves as a lingua franca among indigenous communities, including the Tripuri, Reang, Jamatia, Debbarma and other Borok tribes. According to the 2011 Census of India (Census of India, 2011), Kokborok has approximately 800,000 speakers in India, with additional speakers in Bangladesh, though exact figures for the latter are less documented. Though Kokborok is one of the official language of Tripura, but in urban areas like Agartala, there is a noticeable shift toward Bengali due to its dominance in administration, education, and media.

Linguistically, Kokborok has a rich phonological system, featuring six monophthong vowels and a consonant inventory that includes stops, nasals, fricatives, and approximants. Historically tonal, with pitch distinctions marking lexical meaning, the language is undergoing phonological simplification, particularly in urban settings, due to prolonged contact with non-tonal languages like Bengali and Hindi. Morphologically, Kokborok is agglutinative, employing affixation and compounding to form complex words(Hoque, 2014). Its syntax follows a Subject-Object-Verb (SOV) order, with a robust case system and verbal inflections for tense, aspect, mood, and person, reflecting its Tibeto-Burman roots.

Kokborok is written in both Roman and Bengali scripts, creating challenges for standardization and literacy efforts. Since 1979, Kokborok has been recognized as an official language in Tripura and is integrated into primary and secondary education curricula, with efforts to develop textbooks and teaching materials (Roy et al., 2022). The language remains a cornerstone of Borok cultural identity, expressed through oral traditions, folklore, songs, and ritual practices, such as those tied to festivals like Garia, Kharchi, Ker, and Hojagiri(Jacquesson, 2003).

Despite its cultural significance, UNESCO's Atlas of the World's Languages in danger classifies Kokborok as "vulnerable", reflecting threats from language shift and limited intergenerational transmission in urban areas(UNESCO, 2010). Revitalization efforts are ongoing, including curricu-

lum development, cultural festivals, and media programming like radio broadcasts and local television. Computational innovations, such as Linear Predictive Cepstral Coefficients (LPCC) for vowel recognition, support language documentation and preservation(Debbarma, 2012). Advocacy for Kokborok's inclusion in the Eighth Schedule of the Indian Constitution continues, emphasizing its linguistic and political significance for the Borok people.

3 Low-Resource Indic Language Translation 2025 Shared Task

3.1 Overview and Task Description

Following the success of the "Shared Task: Low-Resource Indic Language Translation" at WMT 2024, which attracted widespread international participation, the initiative will continue as part of the Tenth Conference on Machine Translation (WMT 2025). While recent advancements in machine translation (MT), particularly through multilingual modelling and transfer learning, have led to significant performance gains, developing effective MT systems for low-resource languages remains a critical challenge. This difficulty primarily stems from the limited availability of high-quality parallel corpora, which are essential for training robust and accurate translation models. The shared task aims to address this gap by fostering research and collaboration in low-resource Indic MT and promoting the creation and evaluation of translation systems for linguistically diverse and underrepresented languages.

The WMT 2025 Indic Machine Translation Shared Task aims to tackle the persistent challenges of low-resource translation by focusing on a diverse set of Indic languages drawn from multiple language families. This year, the task is organized around two categories based on the availability of training data.

- **Category 1** includes language pairs with moderate amounts of parallel data: English ⇔ Assamese, English ⇔ Mizo, English ⇔ Khasi, English ⇔ Manipuri, and English ⇔ Nyishi.
- **Category 2** consists of language pairs with extremely limited training resources: English ⇔ Bodo and English ⇔ Kokborok. By highlighting both moderately and severely resource-constrained languages, the task encourages the development of adaptable and data-efficient machine translation approaches capable of addressing the varying degrees of resource scarcity.

3.2 Categories

This year's task features two main categories based on the availability of training data:

3.2.1 Category 1: Moderate Training Data

- English ⇔ Assamese (en⇔as)
- English ⇔ Mizo (en⇔lus)
- English ⇔ Khasi (en⇔kha)
- English ⇔ Manipuri (en⇔mni)
- English ⇔ Nyishi (en⇔njz)

3.2.2 Category 2: Very Limited Training Data

- English ⇔ Kokborok (en⇔trp)
- English ⇔ Bodo (en⇔bodo)

3.3 Goal

This shared task goal is to build machine translation systems that can generate accurate translations regardless of the limitations of limited data availability. Participants are motivated to explore different innovative approaches, including:

- **Leveraging Monolingual Resources**: Utilizing monolingual corpora to improve the performance of translation systems.
- **Multilingual Strategies**: Exploring cross-lingual transfer techniques to support translation for under-resourced language pairs.
- **Cross-lingual Transfer Learning**: Employing models pretrained on high-resource languages and adapting them to low-resource scenarios.
- **Novel Methodologies**: Applying cutting-edge or customized approaches designed specifically for data-scarce environments.

3.4 Data

3.4.1 Training

This WMT 2025 Indic Machine Translation Shared Task leverages a dataset that consists of both parallel and monolingual corpora for Assamese, Khasi, Mizo, Manipuri, Nyishi, Bodo and Kokborok taken from the IndicNE-corp1.0 dataset.

3.4.2 Testing

For the testing section, we have created 2000 language pair sentences for each of the following language pairs:

- English ⇔ Assamese (en⇔as)
- English ⇔ Mizo (en⇔lus)
- English ⇔ Khasi (en⇔kha)
- English ⇔ Manipuri (en⇔mni)
- English ⇔ Nyishi (en⇔njz)
- English ⇔ Kokborok (en⇔trp)
- English ⇔ Bodo (en⇔bodo)

Out of these 2000 sentences, the first 1000 are provided in English, the participant needs to translate them into the specific target(Indic) language, and the remaining 1000 are given in the target language and are to be translated to English.

3.5 Evaluation

All the machine translation systems that are submitted by the participants are evaluated using automatic assessments to ensure balanced analysis of the translation systems. Automatic evaluation is being carried out by the following metrics such as BLEU, TER, ROUGE-L, ChrF and METEOR. Along with those metrics, this year's work also includes Cosine similarity using sentence transformer (all-mpnet-base) model based embeddings for evaluation, which captures the semantic representation of the sentences in the English language direction to measure the performance and accuracy of the models.

4 Dataset

4.1 Training

The dataset for the WMT 2024 Shared Task on Low-Resource Indic Language Translation is primarily based on the IndicNE-Corp1.0 dataset[4]. This corpus was built by aggregating datasets from previous research, including significant contributions from (Laskar et al., 2020) (Laskar et al., 2022), (Khenglawt et al., 2022), and (Laitonjam and Ranbir Singh, 2021). The compiled datasets encompass both parallel and monolingual corpora across four languages: Assamese, Mizo, Khasi, and Manipuri.

In earlier studies, we focused on developing parallel and monolingual corpora for English ⇔ Assamese (en⇔as) (Laskar et al., 2020, 2022), English ⇔ Mizo (en⇔lus) (Khenglawt et al., 2022), English ⇔ Khasi (en⇔kha) (Laskar et al., 2021), and English ⇔ Manipuri (en⇔mni) (Laitonjam and Ranbir Singh, 2021). The data was sourced from a variety of online platforms, including the Bible, multilingual dictionaries (such as Xobdo and Glosbe), multilingual question papers, PMIndia (Haddow and Kirefu, 2020), web pages, blogs, and online newspapers.

Table 1 shows the detailed statistics of the parallel datasets used for training and validation for each language pair.

Type	Sentences	Tokens (eng)	Tokens (target)
Assamese	54,000	1,033,580	878,466
Mizo	50,000	981,513	1,044,077
Khasi	26,000	778,689	948,853
Manipuri	23,687	422,522	357,524
Nyishi	60,000	337,887	323,876
Bodo	15,215	228,219	204,926
Kokborok	2,269	55,634	51,268

Table 1: Parallel data statistics for train and validation.

4.2 Testing

The testing dataset for the 2024 shared task was meticulously curated to present a substantial challenge beyond previous years' datasets. It comprised 1000 samples for each language pair, spanning four distinct and diverse domains: News, Travel, Sports, Entertainment, and Business. This domain-specific distribution aimed to comprehensively evaluate models' performance across varied and complex linguistic contexts, reflecting real-world translation demands. A collaborative approach was employed to create these testing samples, involving four specialized teams, each dedicated to one domain. These teams were provided 1000 English sentences, which they translated into their assigned target languages. The translation teams were instructed to maintain high fidelity to the source material while ensuring the translations were idiomatic and contextually appropriate for each domain.

The test set release process was intentionally staged to introduce additional complexity and rigour. In the first phase, 500 English sentences were released, requiring participants to translate these into the target languages. This forward trans-

[4] https://data.statmt.org/wmt23/indic-mt/

Language Pair	Entertainment	Sports	Healthcare	Travel	Political
English → Assamese	400	400	400	400	400
English → Mizo	400	400	400	400	400
English → Khasi	400	400	400	400	400
English → Manipuri	400	400	400	400	400
English → Nyishi	400	400	400	400	400
English → Bodo	400	400	400	400	400
English → Kokborok	400	400	400	400	400

Table 2: Domain-wise distribution of the 2025 test dataset across all language pairs. Each pair contains 2000 sentences, distributed evenly over five domains.

lation task required participants to demonstrate their models' proficiency in capturing nuances and domain-specific terminology in the target languages. In the second phase, 500 sentences in the target languages were provided, requiring translation back into English. This reverse translation task assessed the models' ability to accurately render the meaning, tone, and subtleties of the original sentences in English, thus testing bidirectional translation capability. The combined forward and reverse tasks aimed to evaluate the accuracy, fluency, and idiomatic correctness of the translations. The careful selection of diverse domains and the structured release of the test set were intended to challenge the generalization capabilities of the participating models. The goal was to ensure that only the most robust models, capable of handling a wide range of real-world scenarios, would excel.

This approach ensures a rigorous and multifaceted evaluation, capturing the subtleties of each language pair's translation performance across different domains.

5 Participants and System Descriptions

Language Pair	Submissions
English - Assamese	17 (primary), 18 (contrastive)
English - Mizo	5 (primary), 9 (contrastive)
English - Khasi	6 (primary), 19 (contrastive)
English - Manipuri	11 (primary), 7 (contrastive)
English - Nyishi	6 (primary), 12 (contrastive)
English - Bodo	6 (primary), 7 (contrastive)
English - Kokborok	3 (primary), 7 (contrastive)

Table 3: Number of participants in the low-resource Indic language translations

In this WMT 2025 Indic MT Shared Task, a total of 17 teams, as illustrated in the Table 4, registered and contributed for this year which is a huge improvement over the last year. We gathered the outputs produced by participant systems, including both primary and contrastive submissions.

A3-108: This team (Yadav and Shrivastava, 2025) system focused on translation for low-resource language pairs, combining a phrase-based SMT framework with subword segmentation through multiple BPE merge operations (500–3000 merges). Their approach involved concatenating and deduplicating segmented bitext to enhance vocabulary coverage and reduce sparsity, supported by KenLM-trained target-side language models. They submitted results for four English–X pairs: Nyishi, Manipuri, Khasi, and Assamese.

AkibaNLP-TUT: The AkibaNLP-TUT (Hamada et al., 2025) team tackled the WMT25 IndicMT task with Transformer-based models implemented in Fairseq. Their approach combined official parallel datasets with additional Bengali–English and Assamese monolingual corpora. A key technique was language-specific word-level noise injection to enhance robustness in low-resource settings, complemented by back-translation to augment English→X training data.

ANVITA: This team (Sivabhavani et al., 2025)submitted systems for three low-resource Indic languages Nyishi, Khasi, and Kokborok covering both primary and contrastive tracks. Their models were built using transfer learning, fine-tuning public pre-trained architectures such as byt5-base and nllb-200-distilled-600M with selective vocabulary expansion and targeted post-editing. The primary submissions relied on organizer-provided datasets, while the contrastive runs incorporated data augmentation through back-translation, sentence concatenation, and proprietary crawled resources. Language-specific strategies included leveraging Bodo data for Kokborok and tailoring vocabulary for Khasi.

BibaoMT: This team submission explored a

Team Name	Name of University/Lab/Industry/Group
A3-108	IIIT Hyderabad
AkibaNLP-TUT	Toyohashi University of Technology / NLP Lab.
ANVITA	CAIR
BilbaoMT	University of the Basque Country
BVSLP	Banasthali Vidyapith
CITK_MT	Central Institute of Technology Kokrajhar
DELAB-IIITM	Indian Institute of Information Technology, Senapati, Manipur
DoDS-IITPKD	Indian Institute of Technology, Palakkad, Kerala
DPKM	Dynamic Partial Knowledge Model Group
Hope for best	University of Delhi
JU-NLP	Jadavpur University
MT@HLT-BLR_Amrita	Amrita School of Computing, Bengaluru, Amrita Vishwa Vidyapeetham, India
NLPTng-NITAP	National Institute of Technology Arunachal Pradesh
RBG-AI	Resilience Business Grids LLP
SRIB-NMT	Samsung Research Institute Bangalore
Trasnformers	Centre for Development of Advanced Computing (C-DAC), Pune
TranssionMT	Transsion Translation

Table 4: The following table provides an overview of the teams registered for the low-resource Indic language translation task at WMT 2025.

lightweight neural MT model with just 22.4M parameters (280 MB) to tackle low-resource English-to-Indic translation. Their approach used a two-stage training pipeline: multilingual pretraining on seven task languages plus three high-resource languages, followed by fine-tuning on target languages. Training data combined official shared-task corpora with NLLB-mined bitexts, Samanantar, HPLT Bengali–English and Hindi–English, and OpenSubtitles Spanish–English datasets, enabling efficient translation despite limited resources.

BVSLP: The BV-SLP(Joshi et al., 2025) team developed MT systems for five language pairs: English⇔Assamese, English⇔Manipuri, and English→Bodo. Their pipeline integrates a rule-based named entity recognition and translation module prior to NMT training, handling organisation and location names via translation or transliteration from a knowledge base. After preprocessing, byte pair encoding (BPE) was applied to prepare data for Transformer-based NMT training, enabling improved handling of named entities in low-resource scenarios.

CITK-MT: The CITK-MT(Wary et al., 2025) team proposed an end-to-end NMT system targeting English→Bodo translation, leveraging a Seq2Seq model with GRU-based encoder–decoder layers and Bahdanau attention to enhance contextual alignment. Their pipeline included extensive data preprocessing, careful hyperparameter tuning (embedding size, hidden units, dropout), and training on Google Colab with NVIDIA A100 GPUs. The system was optimized using early stopping and evaluated via BLEU scores, demonstrating a focused approach for low-resource language translation.

DELAB-IIITM: The DELAB-IIITM(Oinam and Saharia, 2025) team addressed low-resource Indic translation for English⇔Assamese and English⇔Manipuri by fine-tuning the NLLB-200 multilingual model with synthetic parallel data augmentation. They generated synthetic corpora by leveraging bilingual pairs (Manipuri–English, Assamese–English) to create additional data for target languages, yielding 77K sentences after strict data cleaning. Fine-tuning employed the Seq2SeqTrainer framework with the Adafactor optimizer (2e-5 learning rate) and two training epochs, alongside careful train–test splits to mitigate overfitting. Evaluation showed notable BLEU score gains over baseline NLLB models across most directions (e.g., mni-as 0.45 vs. 0.10 baseline).

DoDS-IITPKD: The DoDS IIT Palakkad (Khongthaw et al., 2025) team tackled low-resource Indic language translation by participating with four languages: Khasi, Mizo, Assamese (Category-1) and Bodo (Category-2). Their primary system fine-tuned facebook/nllb-200-distilled-600M for English⇔Khasi and English⇔Mizo, while IndicTrans2 was used for Assamese and Bodo. For the contrastive system, training data was expanded with external corpora

such as PMINDIA and Google SMOL, enabling broader coverage. Both systems applied Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning within the Hugging Face Transformers + PEFT framework, along with language-specific tagging for preprocessing. This modular design balanced translation quality with computational efficiency.

DPKM: The DPKM(Kumar et al., 2025) team presents a low-resource NMT approach for English–Kokborok and English–Bodo translation, leveraging the LLaMA2-8B model with LoRA-based parameter-efficient fine-tuning. Their pipeline involves pretraining on monolingual Kokborok and English corpora (70k Kokborok / 30k English for Kokborok, and 350k Kokborok / 125k English for Bodo) prior to instruction tuning using WMT25 datasets converted into Alpaca format to suit instruction-following objectives. The fine-tuning method integrates LoRA adapters to minimize training overhead on large models.

Hope for best: This team deployed pre-trained IndicTrans2 transformer models for English–Assamese translation without additional fine-tuning, prioritizing fast deployment under CPU-only constraints. They formatted inputs with language tags and applied minimal preprocessing, achieving balanced quality with beam search and batch inference. Their system highlights how off-the-shelf multilingual models can still perform competitively in low-resource shared tasks.

JU-NLP: The JUNLP(Acharya et al., 2025) team addressed English ⇔ Assamese, Mizo, Manipuri, and Bodo translation by fine-tuning multilingual NLLB and IndicTrans2 models using parameter-efficient methods like LoRA and DORA. Their pipeline featured rigorous preprocessing, including deduplication, script harmonization, and alignment filtering to improve data quality. Evaluation on WMT datasets showed competitive BLEU/ChrF scores despite low-resource constraints.

MTHLT-BLR_Amrita: This team(Sheshadri and Gupta, 2025) tackled Assamese and Bodo to English translation using IndicTrans2 enhanced with a novel Representation Fine-tuning (ReFT) method, inserting lightweight modules into encoder layers for targeted adaptation. They optimized ReFT hyperparameters via Bayesian search with Optuna and fine-tuned only 0.5M parameters to avoid overfitting. Experiments on WMT data demonstrated competitive BLEU scores under strict low-resource constraints.

NLPTng-NITAP: The team from NIT Arunachal Pradesh addressed English ⇔ Nyishi translation using the mBART50 multilingual model with Task-Adaptive Fine-Tuning (TAFT). They introduced a custom Nyishi token (<nyi_IN>) and performed full model fine-tuning on WMT25 parallel corpora, leveraging language prefixing for direction control. Their method demonstrates effective transfer learning for new Indic languages under severe low-resource constraints.

RBG-AI: The RBG-AI(H and Ptaszynski, 2025) team developed a multilingual translation pipeline using the MADLAD-400 T5-based model, optimized for both high- and low-resource languages. Their approach employed 4-bit quantization to reduce memory usage and speed up inference on RTX3090 hardware without compromising translation quality. The system incorporated language-specific tags and beam search decoding to improve fluency and directionality. This design balanced translation accuracy with deployment efficiency, suitable for edge or resource-limited environments.

SRIB-NMT: SRIB-NMT participated in the WMT-25 Low-Resource Indic MT challenge with contrastive submissions for four language pairs: English–Assamese, English–Mizo, English–Khasi, and English–Manipuri. Their system used pre-trained NLLB models combined with LoRA fine-tuning to efficiently adapt to low-resource settings. By leveraging cross-lingual transfer techniques, they achieved notable gains in SacreBLEU on blind test sets. The submission highlights parameter-efficient adaptation strategies for multilingual translation tasks.

Transformers: This team(Gupta et al., 2025) developed NMT models for English to Assamese, Bodo, and Manipuri using the OpenNMT framework with a Transformer-based encoder–decoder architecture. Their approach included extensive pre-processing tokenisation, BPE segmentation, and vocabulary generation along with transfer learning from Samanantar v2 to boost low-resource performance. The models were fine-tuned on WMT25 Indic datasets and evaluated with BLEU and perplexity metrics. Deployment was optimised through CTranslate2 for efficient runtime translation on GPUs.

TranssionMT: This team employed a dual-model strategy using IndicTrans2_1B and NLLB_3.3B for low-resource Indic translation. Their system applied cross-iterative back-translation of monolingual data to create high-quality pseudo-parallel corpora and semantic filtering (all-mpnet-base-v2) to enhance domain similarity. Rigorous data cleaning removed noise like URLs and untranslated segments, ensuring improved training quality. The final translations combined outputs from both models to achieve optimal results.

6 Results and Discussion

The results of the WMT 2025 Indic Machine Translation (MT) Shared Task are illustrated in the tables below. For clarity, results are reported separately for each language pair and direction: Assamese–English (as–en) in Table 5 and English–Assamese (en–as) in Table 6. Similarly, results for English–Manipuri (en–mni) and Manipuri–English (mni–en) are presented in Tables 7 and 8, respectively. The English–Khasi (en–kha) and Khasi–English (kha–en) directions are summarized in Tables 9 and 10. For English–Mizo (en–lus) and Mizo–English (lus–en), results are provided in Tables 11 and 12. The English–Nyishi (en–njz) and Nyishi–English (njz–en) results are shown in Tables 14 and 13. Similarly, results for and English–Bodo (en–bodo) and Bodo–English (bodo–en) are reported in Tables 15 and 16. Finally, results for English–Kokborok (en–trp) and Kokborok–English (trp–en) are detailed in Tables 17 and 18. Each table lists the participating systems in descending order of performance, along with their respective evaluation scores. This section presents the evaluation scores of the participants and their submitted system outputs and corresponding papers. Although participants submitted results for both primary and contrastive systems, only the primary system results are highlighted in the corresponding tables.

An evaluation of the quantitative results was performed using metrics like BLEU, METEOR, ROUGE-L, ChrF, and TER. BLEU measures the precision of n-grams in candidate translations relative to reference translations. TER quantifies the number of edits required to transform the candidate translation into the reference. ROUGE-L evaluates the longest common subsequence between the candidate and reference, emphasizing recall-oriented aspects of translation quality. ChrF computes the character n-gram F-score, providing sensitivity to morphological variations. METEOR combines precision, recall, and synonym matching to capture translation adequacy and fluency. In addition to the traditional statistical metrics this year's evaluation also incorporates semantic similarity based on cosine similarity between sentence embeddings generated by the all-mpnet-base model for the Indic language to English direction.

Discussion

For the Assamese language, the team TranssionMT achieved a higher BLEU score of 23.20 in primary mode and 22.41 in contrastive mode for the as-en direction with cosine similarity of 0.92 in primary mode of evaluation. For the en-as direction, this team also achieved a higher score in both primary and contrastive with BLEU scores of 20.97 and 67.50, respectively. This team employed a dual-model strategy using IndicTrans2_1B and NLLB_3.3B for low-resource Indic translation. Their system applied cross-iterative back-translation of monolingual data to create high-quality pseudo-parallel corpora and semantic filtering (all-mpnet-base-v2) to enhance domain similarity.

For the Manipuri language, the team BVSLP achieved a higher BLEU score of 4.15 in the primary system and also achieved a higher cosine similarity of 89.60. In contrastive system submission team TranssionMT achieved BLEU score of 4.49 for the en-mni direction. Team BVSLP pipeline integrates a rule-based named entity recognition and translation module prior to NMT training, handling organisation and location names via translation or transliteration from a knowledge base. After preprocessing, byte pair encoding (BPE) was applied to prepare data for Transformer-based NMT training. For the mni-en direction team TranssionMT achieved higher BLEU scores of 13.37,14.86, and cosine similarities of 0.859, 0.860 in both primary and contrastive systems using their dual model strategy.

For the Khasi language, team DoDS-IITPKD achieved a higher BLEU score of 14.20 in primary system and TranssionMT achieved higher BLEU score of 82.56 in contrastive system for the en-kha direction. This team also achieved higher BLEU score of 4.31 in kha-en direction with their primary system submission, while Trans-

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
TranssionMT	**primary**	**23.20**	**0.703**	**0.699**	**67.71**	**55.82**	**0.920**
TranssionMT	contrastive1	22.41	0.699	0.705	67.08	55.37	0.918
DoDS-IITPKD	contrastive	21.75	0.690	0.703	65.77	53.77	0.913
DoDS-IITPKD	primary	21.40	0.695	0.701	66.14	54.90	0.918
TranssionMT	contrastive2	20.09	0.709	0.694	67.42	60.76	0.929
RBG-AI	contrastive	15.27	0.626	0.632	60.36	68.50	0.882
DELAB-IIITM	primary	15.02	0.604	0.605	59.37	75.25	0.869
BVSLP	primary	14.91	0.615	0.613	60.29	71.33	0.893
SRIB-NMT	contrastive	12.68	0.004	0.601	57.69	88.43	0.849
AkibaNLP-TUT	primary	12.28	0.539	0.557	55.61	78.24	0.826
MT@HLT-BLR_Amrtita	primary	10.58	0.609	0.539	58.06	148.91	0.875
Transformers	primary	7.63	0.437	0.472	47.36	102.55	0.743
JU-NLP	primary	0.37	0.013	0.022	14.26	116.71	0.039
A3-108	constraint	0.33	0.023	0.021	17.50	286.28	0.077
A3-108	contrastive1	0.33	0.023	0.021	17.50	286.32	0.077
A3-108	contrastive2	0.33	0.023	0.021	17.50	286.21	0.077
A3-108	primary	0.33	0.023	0.021	17.50	286.21	0.077

Table 5: Evaluation results for Assamese → English (as→en) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
TranssionMT	contrastive2	67.50	0.772	0.0174	82.46	28.12
TranssionMT	**primary**	**20.97**	**0.470**	**0.0023**	**61.57**	**62.18**
TranssionMT	contrastive1	19.29	0.451	0.0062	60.39	63.86
DoDS-IITPKD	contrastive	17.64	0.422	0.0074	57.71	74.81
DoDS-IITPKD	primary	17.54	0.422	0.0074	57.75	71.17
JU-NLP	primary	16.72	0.412	0.0039	57.22	70.69
DELAB-IIITM	primary	16.11	0.406	0.0030	55.70	68.32
AkibaNLP-TUT	primary	14.03	0.376	0.0132	53.76	74.08
BilbaoMT	contrastive	10.23	0.284	0.0084	43.99	77.84
RBG-AI	contrastive	9.09	0.281	0.0060	45.48	81.73
Transformers	primary	6.92	0.234	0.0010	41.92	89.99
A3-108	constraint	3.03	0.115	0	31.63	108.91
A3-108	contrastive1	3.03	0.114	0	31.30	107.33
A3-108	primary	2.97	0.113	0	31.46	107.35
A3-108	contrastive2	2.93	0.109	0	30.49	104.26
BVSLP	primary	1.81	0.058	0.0030	27.45	98.66
HopeForBest	contrastive	0.00	0.000	0.0000	1.43	152.04

Table 6: Evaluation results for English → Assamese (en→as) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
TranssionMT	contrastive1	4.49	0.164	0.0125	44.11	85.96
BVSLP	**primary**	**4.15**	**0.146**	**0.0099**	**41.43**	**89.60**
JU-NLP	primary	4.12	0.155	0.0113	43.87	93.16
RBG-AI	contrastive	4.11	0.144	0.0037	40.64	90.04
TranssionMT	primary	3.66	0.135	0.0100	38.64	92.76
SRIB-NMT	contrastive	3.26	0.000	0.0093	39.52	104.80
DELAB-IIITM	primary	3.15	0.113	0.0087	37.51	132.05
Transformers	primary	2.79	0.099	0.0040	33.41	98.76
BilbaoMT	contrastive	2.75	0.091	0.0096	31.69	90.46

Table 7: Evaluation results for English → Manipuri (en→mni) translation direction

sionMT system achieved higher BLEU score of 24.17 with their contrastive system submission. Team DoDS-IITPKD's primary system fine-tuned NLLB-200-distilled-600M for English–Khasi and English–Mizo, and used IndicTrans2 for Assamese and Bodo. The contrastive system incorporated additional corpora (e.g., PMINDIA, Google SMOL) to expand training data. Both systems employed

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
TranssionMT	contrastive1	14.86	0.555	0.576	58.08	76.10	0.8601
TranssionMT	**primary**	**13.37**	**0.554**	**0.565**	**56.71**	**79.31**	**0.8599**
JU-NLP	primary	8.10	0.480	0.495	49.60	100.29	0.7974
RBG-AI	contrastive	7.85	0.431	0.468	48.08	94.56	0.7608
DELAB-IIITM	primary	7.35	0.464	0.479	48.78	103.20	0.7645
AkibaNLP-TUT	primary	5.74	0.328	0.370	41.28	109.95	0.6179
Transformers	primary	4.27	0.291	0.327	36.97	125.99	0.5548
BVSLP	primary	3.06	0.221	0.251	35.61	139.03	0.5671
SRIB-NMT	contrastive	0.34	0.026	0.034	12.64	261.49	0.0685

Table 8: Evaluation results for Manipuri → English (mni→en) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
TranssionMT	contrastive2	82.56	0.906	0.915	90.17	11.32
DoDS-IITPKD	contrastive	20.08	0.452	0.534	47.36	59.98
ANVITA	contrastive2	19.43	0.457	0.549	45.93	54.41
ANVITA	contrastive	18.83	0.451	0.543	45.48	55.75
DoDS-IITPKD	**primary**	**14.20**	**0.370**	**0.431**	**39.95**	**87.50**
RBG-AI	contrastive	10.31	0.265	0.344	32.10	77.01
BilbaoMT	contrastive	8.03	0.253	0.352	30.32	73.72
ANVITA	primary	7.34	0.248	0.343	28.34	75.77
A3-108	primary	4.26	0.192	0.255	26.80	96.24
A3-108	contrastive2	4.24	0.188	0.254	26.55	94.71
A3-108	contrastive1	4.23	0.193	0.255	26.76	97.94
SRIB-NMT	contrastive	4.19	0.000	0.227	25.63	113.97
A3-108	constraint	4.10	0.194	0.252	26.90	100.62

Table 9: Evaluation results for English → Khasi(en→kha) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
TranssionMT	contrastive2	24.17	0.635	0.686	63.04	52.81	0.879
ANVITA	contrastive	7.44	0.376	0.416	41.85	102.87	0.738
RBG-AI	contrastive	5.64	0.273	0.323	36.36	122.12	0.624
DoDS-IITPKD	contrastive	5.52	0.289	0.349	34.85	113.30	0.644
ANVITA	contrastive2	4.39	0.220	0.284	30.65	123.25	0.551
DoDS-IITPKD	**primary**	**4.31**	**0.239**	**0.293**	**31.33**	**131.86**	**0.579**
ANVITA	primary	1.99	0.106	0.137	20.88	223.26	0.297
A3-108	contrastive2	1.09	0.081	0.114	19.26	171.43	0.243
A3-108	contrastive1	1.06	0.080	0.111	19.46	176.13	0.246
A3-108	primary	1.05	0.079	0.111	19.47	177.43	0.246
A3-108	constraint	1.05	0.081	0.111	19.57	179.16	0.247
SRIB-NMT	contrastive	0.34	0.026	0.034	12.64	261.49	0.069

Table 10: Evaluation results for Khasi → English (kha→en) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
TranssionMT	contrastive2	37.81	0.660	0.704	69.93	41.80
JU-NLP	**primary**	**15.83**	**0.419**	**0.548**	**52.00**	**69.01**
DoDS-IITPKD	contrastive	14.72	0.407	0.506	48.55	69.49
DoDS-IITPKD	primary	14.26	0.415	0.515	48.51	72.22
SRIB-NMT	contrastive	12.45	0.368	0.509	47.53	78.69
RBG-AI	contrastive	12.44	0.359	0.476	46.83	76.47
BilbaoMT	contrastive	11.06	0.325	0.453	40.83	69.20
DoDS-IITPKD	primary (dup)	10.38	0.537	0.576	55.09	86.84

Table 11: Evaluation results for English → Mizo (en→lus) translation direction

LoRA-based parameter-efficient fine-tuning within the Hugging Face Transformers + PEFT frame-

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
TranssionMT	contrastive2	18.45	0.669	0.684	63.13	61.60	0.915
JU-NLP	**primary**	**12.30**	**0.578**	**0.620**	**58.14**	**78.81**	**0.889**
RBG-AI	contrastive	11.92	0.557	0.588	55.76	79.96	0.871
DoDS-IITPKD	contrastive	11.81	0.544	0.581	55.17	74.39	0.865
DoDS-IITPKD	primary	10.38	0.537	0.576	55.09	86.84	0.874
SRIB-NMT	contrastive	0.007	0.001	0.002	6.12	160.06	0.065

Table 12: Evaluation results for Mizo → English (lus→en) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
ANVITA	**primary**	**11.59**	**0.414**	**0.511**	**49.85**	74.09	0.786
ANVITA	contrastive2	11.25	0.404	0.513	49.36	**73.79**	0.783
ANVITA	contrastive	11.13	0.416	0.506	48.92	74.17	**0.798**
RBG-AI	contrastive	9.62	0.369	0.387	49.43	93.23	0.624
NLPTng-NITAP	primary	5.42	0.307	0.371	41.37	105.61	0.687
A3-108	primary	1.27	0.086	0.121	23.44	138.23	0.211
A3-108	contrastive_2	1.26	0.083	0.119	23.29	139.45	0.203
A3-108	contrastive_1	1.19	0.081	0.116	22.98	145.27	0.205
A3-108	constraint	1.19	0.081	0.113	23.35	147.92	0.201

Table 13: Evaluation results for Nyishi → English (njz→en) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
ANVITA	**primary**	**6.21**	**0.210**	**0.283**	**34.01**	**81.53**
ANVITA	contrastive	5.92	0.203	0.274	34.08	82.83
BilbaoMT	contrastive	3.92	0.132	0.190	29.38	87.77
NLPTng-NITAP	primary	3.40	0.105	0.180	24.58	92.87
RBG-AI	contrastive	2.45	0.080	0.160	12.57	97.19
A3-108	contrastive_2	1.23	0.049	0.078	20.21	120.46
A3-108	primary	1.19	0.049	0.078	20.37	123.93
A3-108	contrastive_1	1.18	0.050	0.077	20.43	124.40
A3-108	constraint	1.17	0.050	0.077	20.65	127.13

Table 14: Evaluation results for English → Nyishi (en→njz) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
DoDS-IITPKD	contrastive	24.97	0.519	0.169	67.81	51.50
DoDS-IITPKD	**primary**	**24.45**	**0.513**	0.168	**67.71**	**51.84**
JU-NLP	primary	19.71	0.455	**0.169**	62.47	64.97
Transformers	contrastive	19.30	0.452	0.168	67.29	72.92
BilbaoMT	contrastive	10.18	0.283	0.160	46.87	71.09
DPKM	primary	4.38	0.132	0.009	35.50	92.56
BVSLP	primary	1.35	0.040	0.168	17.05	106.11
CITK_MT	primary	0.31	0.019	0.003	7.24	808.91
RBG-AI	contrastive	0.20	0.006	0.027	0.81	131.96

Table 15: Evaluation results for English → Bodo (en→bodo) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
DoDS-IITPKD	contrastive	22.11	0.629	0.688	63.55	52.84	0.897
DoDS-IITPKD	**primary**	**21.68**	**0.627**	**0.679**	**62.95**	**54.29**	**0.888**
Transformers	contrastive	11.83	0.526	0.559	54.38	85.73	0.831
RBG-AI	contrastive	1.40	0.071	0.101	19.45	206.05	0.231

Table 16: Evaluation results for Bodo → English (bodo→en) translation direction

work and applied language-specific tagging during preprocessing, achieving a balance between translation quality and computational efficiency.

For the Mizo language, team JU-NLP achieved

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
ANVITA	contrastive	6.997	0.300	0.367	38.08	76.26
RBG-AI	contrastive	2.220	0.134	0.204	22.85	103.51
ANVITA	**primary**	**1.756**	**0.107**	**0.168**	**18.58**	**104.04**
BilbaoMT	contrastive	1.417	0.076	0.134	20.08	91.93
ANVITA	contrastive2	0.553	0.041	0.054	13.38	335.55
DPKM	primary	0.179	0.006	0.015	5.60	105.49

Table 17: Evaluation results for English → Kokborok (en→trp) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
ANVITA	contrastive	2.99	0.163	0.218	25.52	117.73	0.487
ANVITA	**primary**	**2.41**	**0.108**	**0.175**	**23.55**	**129.15**	**0.359**
RBG-AI	contrastive	1.59	0.086	0.125	20.00	147.75	0.302
ANVITA	contrastive2	0.79	0.051	0.081	16.46	170.61	0.201

Table 18: Evaluation results for Kokborok → English (trp→en) translation direction

higher BLUE score of 15.83 in primary system submission and TranssionMT achieved BLEU score of 37.81 in contrastive system submissions in en-lus direction. JUNLP also achieved higher BLEU score of 12.30 with cosine similarity of 0.889 in primary mode and TranssionMT achieved higher BLUE score of 18.45 with cosinie similarity of 0.915 contrastive submissions in lus-en direction.The JUNLP team addressed English–Assamese, Mizo, Manipuri, and Bodo translation by fine-tuning NLLB and IndicTrans2 using parameter-efficient methods (LoRA and DORA). Their approach emphasized extensive preprocessing including deduplication, script harmonization, and alignment filtering to enhance data quality.

For the Nyishi language, team ANVITA achieved a higher BLUE scores of 11.59 with cosine similarity of 0.786 in primary submission out of all submissions in njz-en direction and achieved higher BLUE score of 6.21 in en-njz direction. Their models employed transfer learning by fine-tuning public pre-trained architectures such as ByT5-base and NLLB-200-distilled-600M, incorporating selective vocabulary expansion and targeted post-editing. The primary submissions utilized organizer-provided datasets, while the contrastive runs applied data augmentation through back-translation, sentence concatenation, and proprietary crawled resources.

For the Bodo language, team DoDS-IITPKD achieved higher BLEU score of 24.45,24.97 in primary and contrastive modes respectively in en-bodo direction. This team also achieved higher BLEU scores of 21.68, 22.11 in primary and contrastive submissions respectively in bodo-en direction.Team DoDS-IITPKD's primary system fine-tuned NLLB-200-distilled-600M for English–Khasi and English–Mizo, and used IndicTrans2 for Assamese and Bodo. The contrastive system incorporated additional corpora (e.g., PMINDIA, Google SMOL) to expand training data. Both systems employed LoRA-based parameter-efficient fine-tuning techniques.

For the Kokborok language, team ANVITA achieved higher BLEU scores of 2.41 in trp-en and 1.756 in en-trp directions with their primary submissions. This team also presented their higher scores in contrastive submissions also. Their models leveraged transfer learning by fine-tuning public pre-trained architectures such as ByT5-base and NLLB-200-distilled-600M, combined with selective vocabulary expansion and targeted post-editing. Primary submissions used organizer-provided datasets, while contrastive runs employed data augmentation via back-translation, sentence concatenation, and proprietary crawled resources. Additionally, language-specific strategies included leveraging Bodo data for Kokborok and tailoring vocabulary for Khasi.

7 Analysis

The evaluation results across multiple translation directions reveal a competitive landscape with significant variations in performance. A key finding is the direct correlation between the size of the parallel training data and the translation quality, although some notable exceptions exist. The figures (Figure 1, 2, and 3) illustrate these findings, providing a visual context for the observations.

Figure 1 shows the best **primary** BLEU score for

each language direction (the highest BLEU among primary submissions for that direction). Figure 2 visualizes BLEU vs. ChrF for the set of all primary submissions.

Observations

- **Correlation with Data Size:** There is a strong general trend that language pairs with larger parallel datasets, such as Assamese and Mizo, have higher translation scores. Conversely, languages with very limited data, like Kokborok, show the lowest performance. This confirms that data scarcity remains a significant bottleneck for low-resource languages.
- **Outlier Performance in Bodo:** A particularly noteworthy finding is the performance of the Bodo language pair. Despite having a relatively small dataset of only 15,215 sentences, it achieved the highest overall BLEU score of 24.45 for the $en \rightarrow bodo$ translation. This suggests that the quality of the Bodo data, or the highly effective model and training strategies employed by teams like DoDS-IITPKD, compensated for the limited size. This performance highlights that data quality and model optimization can sometimes outweigh the sheer quantity of data.
- **Dominance of Key Teams:** Teams such as TranssionMT and DoDS-IITPKD consistently delivered high-performing models, frequently securing the top spot in the language pairs they participated in.
- **Asymmetry in Translation Direction:** A consistent pattern emerged where the translation quality for one direction of a language pair was notably different from the other. This could be due to differences in data quality for each direction or inherent linguistic challenges in translating into a specific language.
- **Correlation of Metrics:** The Figure 2 scatter plot illustrates a clear positive correlation between BLEU and ChrF scores. This indicates that models that perform well on one metric of translation quality generally also perform well on the other, reinforcing the validity of these metrics as indicators of good performance.

Language-wise Analysis

The key findings for each language pair are given below:

- $Assamese \leftrightarrow English$: With one of the largest datasets (54,000 sentences), this pair yielded strong results. TranssionMT was the top performer in both directions, with BLEU scores of 23.20 for $as \rightarrow en$ and 20.97 for $en \rightarrow as$. The performance here aligns with the substantial training data available.
- $Mizo \leftrightarrow English$: This pair also had a large dataset (50,000 sentences), and the results reflect this. JU-NLP consistently outperformed DoDS-IITPKD, achieving the highest BLEU scores for both $en \rightarrow lus$ (15.83) and $lus \rightarrow en$ (12.30).
- $Khasi \leftrightarrow English$: With 26,000 sentences, the performance was moderate. DoDS-IITPKD excelled in this pair, securing the highest BLEU scores in both $en \rightarrow kha$ (14.20) and $kha \rightarrow en$ (4.31). The significant performance gap between the two directions is a point of interest.
- $Manipuri \leftrightarrow English$: Despite a dataset of 23,687 sentences, the $en \rightarrow mni$ direction proved exceptionally challenging, with the top BLEU score being only 4.15. In the reverse direction ($mni \rightarrow en$), TranssionMT led with a much higher BLEU of 13.37. This disparity is a key finding, suggesting that the complexity of translating English into a tonal, agglutinative language like Manipuri is a significant hurdle.
- $Nyishi \leftrightarrow English$: This language pair had the highest sentence count (60,000), but the lowest token count among the larger datasets, suggesting shorter sentences. ANVITA's performance (11.59 and 6.21) was moderate, indicating that sentence quantity alone is not the sole determinant of success.
- $Bodo \leftrightarrow English$: This language pair is the most striking example of data quality and model effectiveness. Despite a small dataset of only 15,215 sentences, DoDS-IITPKD achieved the highest BLEU score (24.45), demonstrating that high-quality data and strong modeling can overcome the limitations of data size.
- $Kokborok \leftrightarrow English$: With the least amount of data (2,269 sentences), this pair

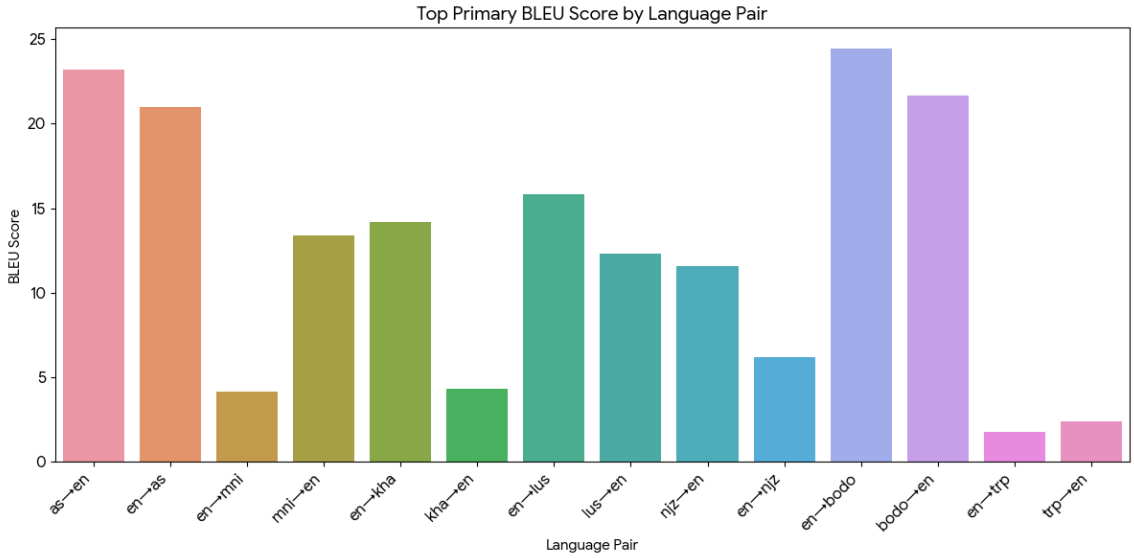

Figure 1: Top BLEU among **primary** submissions per language pair.

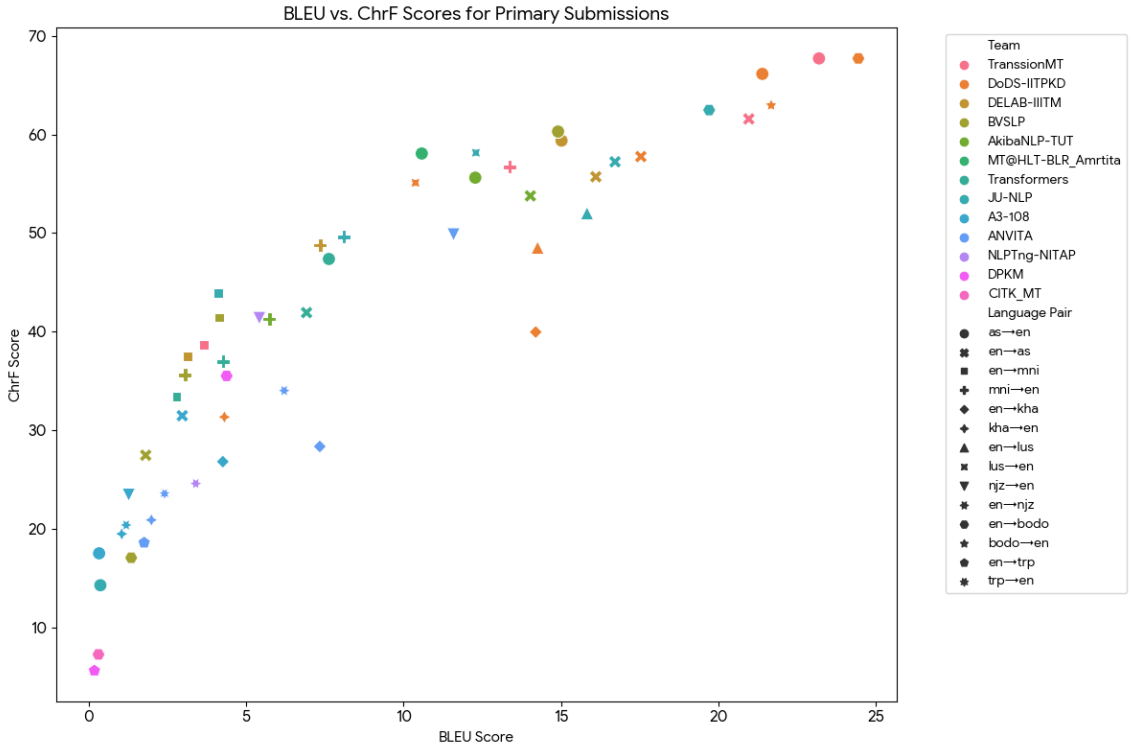

Figure 2: BLEU vs. ChrF for **primary** submissions. Each point is labelled with "team:language".

exhibited the lowest BLEU scores for both directions (1.76 and 2.41), confirming that data scarcity is the primary limiting factor for this language.

Team-wise Analysis

- TranssionMT: This team demonstrated exceptional performance in the Assamese-English and Manipuri-English pairs, consistently ranking at the top. Their models achieved the highest BLEU scores for $as \rightarrow en$ (23.20), $en \rightarrow as$ (20.97), and $mni \rightarrow en$ (13.37), highlighting their strength in these specific languages.

- DoDS-IITPKD: With the highest overall BLEU score of 24.45 for en→bodo, DoDS-IITPKD proved to be a dominant force, espe-

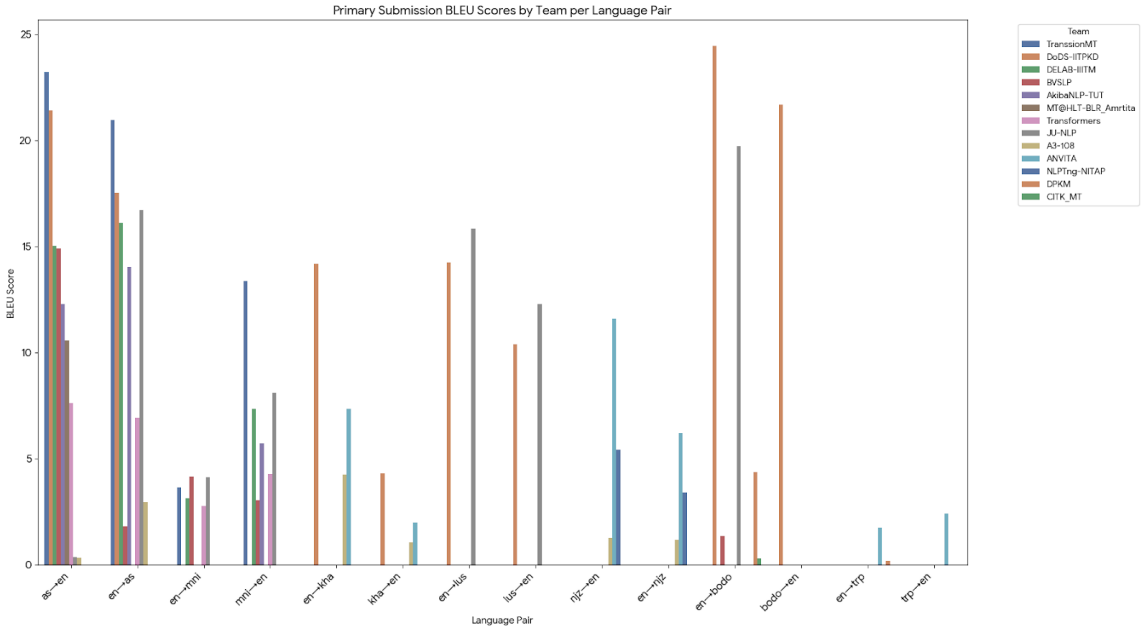

Figure 3: BLEU by team per language pair for **primary** submissions.

cially for the Bodo and Khasi language pairs, where they led in both translation directions.

- JU-NLP: This team's primary submissions were most effective for the Mizo language pairs, where they achieved the highest BLEU scores in both en→lus and lus→en.
- ANVITA: While ANVITA's performance varied, they were the clear leaders in the Nyishi and Kokborok language pairs. Their models, despite the challenges, achieved the best scores for all four directions involving these languages.
- BVSLP: BVSLP's top performance was for the en→mni translation direction, with a BLEU score of 4.15.
- Other Teams: Teams such as DELAB-IIITM, AkibaNLP-TUT, Transformers, and others participated in a number of language pairs, generally achieving lower, though still competitive, scores compared to the top performers.

Categorization of Approaches

Based on the system descriptions, the submitted approaches can be classified into the following methodological categories described in Table19. Note that some teams appear in more than one category when their systems span multiple techniques.

Language-wise Impact and Approach Trends

We further analyze the distribution of techniques across languages for primary submissions in both translation directions (Table 20).

- **>50k pairs** – Teams favored standard Transformer fine-tuning with minor augmentation. Substantial gains were observed from clean fine-tuning alone.
- **20k–30k pairs** – Back-translation and cross-lingual transfer were the most common strategies.
- **<20k pairs** – Parameter-efficient methods and transfer from related languages dominated. Data synthesis played a critical role in achieving competitive performance.

Conclusion

The outcomes of the participating teams in the WMT 2025 translation task have been comprehensively evaluated using a combination of automated metrics and semantic similarity measures. This year's shared task on low-resource Indic language translation utilized the updated **IndicNE-Corp2.0** dataset, which introduced broader domain coverage and incorporated three additional languages Nyishi, Bodo, and Kokborok extending the scope beyond the four language pairs evaluated in 2024. A newly curated test set with higher linguistic and

Category	Teams	Key Characteristics
Phrase-based SMT & Statistical Methods	A3-108	Traditional SMT with BPE, KenLM, deduplication, and focus on vocabulary coverage.
Transformer-based NMT (Standard)	AkibaNLP-TUT, BVSLP, JU-NLP, SRIB-NMT, Transformers	Transformer encoder–decoder architectures (Fairseq, OpenNMT, IndicTrans2, NLLB) with various preprocessing and fine-tuning strategies.
Parameter-efficient Fine-tuning (LoRA, DORA, ReFT)	DoDS-IITPKD, DPKM, JU-NLP, MT@HLT-BLR_Amrita, SRIB-NMT	Efficient adaptation of large multilingual models with minimal computational overhead.
Pretrained Multilingual Models (Zero/Few-shot)	Hope for Best, NLPTng-NITAP, RBG-AI	Direct use of pretrained models (IndicTrans2, mBART50, MADLAD-400) with minimal or targeted adaptation.
Back-translation / Synthetic Data Augmentation	AkibaNLP-TUT, ANVITA, DELAB-IIITM, TranssionMT	Creation of pseudo-parallel data from monolingual corpora to improve low-resource performance.
Transfer Learning & Multilingual Pretraining	ANVITA, BibaoMT, DoDS-IITPKD, Transformers	Leveraging high-resource languages or multilingual corpora to improve target language performance.
Custom Architectures / Specialized Modules	BVSLP (NER module), CITK-MT (GRU + Bahdanau), MT@HLT-BLR_Amrita (ReFT)	Architectures or modules tailored for specific challenges such as named entity handling or fine-grained adaptation.

Table 19: Categorization of submitted systems by methodological approach.

Language (Parallel Data Size)	Common Approaches Seen	Findings
Assamese (54k)	Transformer fine-tuning, direct pretrained model usage, some SMT	Largest resource size in the set; multiple teams reported strong BLEU gains with LoRA fine-tuning.
Mizo (50k)	Transformer fine-tuning + LoRA, back-translation	AkibaNLP-TUT and DoDS-IITPKD achieved consistent gains with monolingual augmentation.
Khasi (26k)	Transfer learning (ByT5, NLLB), BPE-based SMT	SMT still competitive for specific pairs; some systems leveraged Bodo data for transfer.
Manipuri (23.6k)	NLLB fine-tuning, Transformer training, back-translation	Popular among teams due to moderate resource availability.
Nyishi (60k)	SMT + mBART50 fine-tuning	Larger corpus size but fewer participating teams; most relied on transfer learning with prefix tokens.
Bodo (15.2k)	LoRA fine-tuning, ReFT, custom GRU Seq2Seq	Very low-resource; teams adopted parameter-efficient tuning or synthetic data generation.
Kokborok (2.3k)	Transfer learning from related languages, instruction-tuned LLaMA2	Extremely low-resource; innovative data sourcing and vocabulary sharing strategies applied.

Table 20: Language-wise trends in approach adoption for primary submissions in both directions.

structural complexity was also introduced, providing a more rigorous benchmark for system performance. These enhancements are aimed at capturing finer-grained differences in translation quality and reflecting more realistic application scenarios for low-resource Indic languages.

Acknowledgements

We acknowledge the use of linguistic resources and prior descriptive works on Khasi and related languages, which provided valuable background for this study. In particular, we refer to foundational contributions such as (Bareh, 1977), (Grierson, 1928), (Grierson, 1903), (Gurdon, 1904), (Gurdon, 1907), (Nagaraja, 1985), and (Pyrse, 1855).

References

Priyobroto Acharya, Haranath Mondal, Dipanjan Saha, Dipankar Das, and Sivaji Bandyopadhyay. 2025. JU-NLP: Improving low-resource indic translation system with efficient lora-based adaptation. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

S. K. Acharya. 1971. Languages of khasis. *Mainstream*, May 22:19–26.

James F. Allen. 2003. *Natural language processing*, page 1218–1222. John Wiley and Sons Ltd., GBR.

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

Hamlet Bareh. 1977. *The Language and Literature of Meghalaya*. Indian Institute of Advanced Study, Shimla.

Pramod Chandra Bhattacharya. 1977. *A Descriptive Analysis of the Boro Language*. Department of Publication, Gauhati University, Gauhati, Assam. Originally presented as thesis, Univ. of Gauhati, 1965.

Krishna Boro. 2021. Focus enclitics in bodo. *Linguistics of the Tibeto-Burman Area*, 44(1):75–112.

Census of India. 2011. Language data: Tripura, 2011 census.

Anne Daladier. 2002. Definiteness in amwi: grammaticalization and syntax. *Recherches linguistiques de Vincennes*, 31:61–78. Online edition: mis en ligne le 06 juin 2005; accessed 2025-08-15.

Abhijit Debbarma. 2012. Isolated kokborok vowels recognition. In *Global Trends in Information Systems and Software Applications*, pages 489–493, Berlin, Heidelberg. Springer Berlin Heidelberg.

Khumbar Debbarma, Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2012. Morphological analyzer for kokborok. In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, pages 41–52, Mumbai, India. The COLING 2012 Organizing Committee.

George Abraham Grierson. 1903. *Linguistic Survey of India, Vol II, Part 1*. Banarsidass, Delhi.

George Abraham Grierson. 1928. *Linguistic Survey of India, Vol 11: Mon-Khmer and Siamese-Chinese Family (including Khasi and Tai)*. Motilal Banarsidass, Delhi.

Neha Gupta, Saurabh Salunkhe, Bhagyashree Wagh, and Harish Bapat. 2025. Transformers : Leveraging opennmt and transfer learning for low-resource indian language translation. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

P. R. T. Gurdon. 1904. *English Khasi Dictionary*. Mittal Publication, New Delhi.

P. R. T. Gurdon. 1907. *The Khasis*. Macmillan & Co, London.

Loitongbam Gyanendro Singh, Lenin Laitonjam, and Sanasam Ranbir Singh. 2016. Automatic syllabification for Manipuri language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 349–357, Osaka, Japan. The COLING 2016 Organizing Committee.

Barathi Ganesh H and Michal Ptaszynski. 2025. RBG-AI: Benefits of multilingual language models for low-resource languages. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india.

Shoki Hamada, Tomoyoshi Akiba, and Hajime Tsukada. 2025. AkibaNLP-TUT: Injecting language-specific word-level noise for low-resource language translation. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, November 15-16,2025 Suzhou, China.

F. Hoque. 2014. Kokborok: A major tribal language of tripura. *IOSR Journal of Humanities and Social Science*.

François Jacquesson. 2003. Kokborok, a short analysis. In *Hukumu, 10th anniversary volume*, pages 109–122. Kokborok Tei Hukumu Mission.

Nisheeth Joshi, Palak Arora, Anju Krishnia, Riya Lonchenpa, and Mahsilenuo Vizo. 2025. BVSLP: Machine translation using linguistic embellishments for indicmt shared task 2025. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, November 15-16,2025 Suzhou, China.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources english-nyishi pair. *Sādhanā*, 48(4):237.

Vanlalmuansangi Khenglawt, Sahinur Rahman Laskar, Santanu Pal, Partha Pakray, and Ajoy Kumar Khan. 2022. Language resource building and English-to-mizo neural machine translation encountering tonal words. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 48–54, Marseille, France. European Language Resources Association.

Ontiwell Khongthaw, G. L. John Salvin, Budde Shrikant Tryambak, Abigairl Nyasha Chigwededza, Dhruvadeep Malkar, and Swapnil Hingmire. 2025. DoDS-IITPKD : Submissions to the wmt25 low-resource indic language translation task. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

Deepak Kumar, Laishram Thoibisana Devi, and Asif Ekbal. 2025. DPKM : Tackling low-resource nmt with instruction-tuned llama2: A study on kokborok and bodo. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

Lenin Laitonjam and Sanasam Ranbir Singh. 2021. Manipuri-English machine translation using comparable corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 78–88, Virtual. Association for Machine Translation in the Americas.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji Darsh Kaushik, Partha Pakray, and Sivaji Bandyopadhyay. 2021. EnKhCorp1.0: An English–Khasi corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 89–95, Virtual. Association for Machine Translation in the Americas.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. EnAsCorp1.0: English-Assamese corpus. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 62–68, Suzhou, China. Association for Computational Linguistics.

Sahinur Rahman Laskar, Riyanka Manna, Partha Pakray, and Sivaji Bandyopadhyay. 2022. A domain specific parallel corpus and enhanced english-assamese neural machine translation. *Computación y Sistemas*, 26(4):1669–1687.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shakuntala Mahanta. 2012. Assamese. *Journal of the International Phonetic Association*, 42(2):217–224.

K. S. Nagaraja. 1985. *Khasi: A Descriptive Analysis*. Deccan College, Poona.

K.S. Nagaraja. 2015. Kokborok grammar (an old and rare book). *Exoticindiaart.com*. Published: 2015-05-07.

Dingku Singh Oinam and Navanath Saharia. 2025. DELAB-IIITM : Enhancing low-resource machine translation for manipuri and assamese. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dhrubajyoti Pathak, Sanjib Narzary, Sukumar Nandi, and Bidisha Som. 2025. Part-of-speech tagger for bodo language using deep learning approach. *Natural Language Processing*, 31(2):215–229.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

W. Pyrse. 1855. *Introduction to the Khasi Language*. School Book Society Press, Calcutta.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gargi Roy, Rajesh Kumar, and Kārumūri V. Subbārāo. 2022. Control structures in kokborok: A case of syntactic convergence. *Lingua Posnaniensis*, 63(1):21–52.

Rumphang K. Rynjah and Saralin A. Lyngdoh. 2023. Cross-linguistic comparisons of noun phrase constructions in khasi varieties. *Indian Journal of Language and Linguistics*, 4(2):42–53.

Shaillashree K Sheshadri and Deepa Gupta. 2025. MT@HLT-BLR_Amrita: Bayesian optimization of representation-finetuned adapters for low-resource indic multilingual neural machine translation. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

Loitongbam Gyanendro Singh and Sanasam Ranbir Singh. 2017. Word polarity detection using syllable features for manipuri language. In *2017 International Conference on Asian Language Processing (IALP)*, pages 206–209.

J Sivabhavani, Daneshwari Kankawadi, Abhinav Mishra, and Biswajit Paul. 2025. ANVITA : A multi-pronged approach for enhancing machine translation of extremely low-resource indian languages. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, November 15-16,2025 Suzhou, China.

Matthew Snover, Bonnie Dorr, Rich 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: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

UNESCO. 2010. Atlas of the world's languages in danger.

Subhash Kumar Wary, Birhang Borgoyary, Akher Uddin Ahmed, Mohanji Prasad Sah, and Apurbalal Senapati. 2025. CITK_MT: An attention-based neural translation system for english to bodo. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

Wikipedia contributors. 2025a. Assamese language. https://en.wikipedia.org/wiki/Assamese_language. Accessed: 2025-08-15.

Wikipedia contributors. 2025b. Boro language (india). https://en.wikipedia.org/wiki/Boro_language_(India). Accessed: 2025-08-15.

Wikipedia contributors. 2025c. Khasi language. https://en.wikipedia.org/wiki/Khasi_language. Accessed: 2025-08-15.

Wikipedia contributors. 2025d. Mizo language. https://en.wikipedia.org/wiki/Mizo_language. Accessed: 2025-08-15.

Wikipedia contributors. 2025e. Nishi language. https://en.wikipedia.org/wiki/Nishi_language. Accessed: 2025-08-15.

Saumitra Yadav and Manish Shrivastava. 2025. A3-108 : A preliminary exploration of phrase-based smt and multi-bpe segmentations for low-resource indian languages. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

R. Zothanliana. 2021. *A Study of the Development of Mizo Language in Relation to Word Formation*. Ph.d. thesis, Mizoram University, Aizawl, India.