# Bridging Linguistic Gaps: Neural Machine Translation between Hindi and Malayalam

1st Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

2nd Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

3rd Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

4th Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

5th Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

6th Given Name Surname
*dept. name of organization (of Aff.)*
*name of organization (of Aff.)*
City, Country
email address or ORCID

***Abstract*—Neural machine translation has seen significant advances through large-scale multilingual models, which have pushed the boundaries of translation quality across diverse language pairs. However, these models often prioritise high-resource languages and require substantial computational infrastructure, making them less accessible and sustainable for low-resource linguistic scenarios, such as Indian language translation. In particular, bilingual pairs such as Hindi-Malayalam have received limited attention in mainstream multilingual training regimes, despite their linguistic significance.**

**Addressing these limitations, this article proposes a lightweight bilingual machine translation framework tailored for Hindi-Malayalam translation, using a transformer-based architecture. Employing subword-level tokenization and trained on carefully curated parallel corpora, the models prioritise semantic alignment and syntactic generalisation over brute-force scale. The work aims to fill the gap left by large models in underrepresented language pairs and demonstrate the feasibility of compact translation systems in such contexts. The work also includes creating a new test benchmark (MHM dataset) that contains parallel data for Hindi-Malayalam pairs, curated from textbooks to include sentences of varying complexity levels.**

**Extensive evaluations across test datasets show that prototype models deliver competitive performance, particularly on COMET and chrF2++ metrics, where they rival results from significantly larger models. These findings accentuate the potential of efficient bilingual models to deliver high-quality translation in low-resource settings. Machine translation in diverse languages can significantly improve equitable access to educational and informational materials worldwide, mainly for linguistically marginalised communities. The work aligns with the goals of SDG 4, thus contributing towards a responsible and sustainable world.**

***Index Terms*—Natural language processing, Machine translation, Deep learning, Low-resource languages**

## I. Introduction

Machine translation (MT) is defined as systems or models that translate text from one language to another. From a speculative idea once, MT has evolved significantly since its inception, especially with the change from statistical to neural methods, drastically improving translation quality and expanding its use in daily life [1]–[3]. Despite all the advances, the benefits are fruited around high-resource languages with large and substantial data, leaving low-resource languages dry [4]. Low-resource languages, although spread globally, miss the researcher's eyes due to various constraints, including the scarcity of data resources and language-specific tools [4], [5]. Most publicly available Internet materials are in English (65%), but English speakers make up only around 25% of the world's population [6]. Addressing the imbalance in translation models is often hindered by the scarcity and procurement challenges of training data for low-resource languages, which struggle to support modern data-intensive techniques [5], [7].

Neural Machine Translation (NMT) is a deep learning technique that independently translates text between languages employing neural networks that understand the links between words and phrases in large parallel datasets. Multilingual Neural Machine Translation (MNMT) systems have emerged as efficient models that enable translation between multiple language pairs using a single unified architecture [7]. This approach minimises the need for training and maintaining several distinct models, reduces computational overhead, simplifies deployment pipelines, and enhances scalability in multilingual environments. The primary strength of MNMT lies in its ability to share parameters to better generalise in related languages and grasp linguistic transfer, improving the quality of translation for most languages [2], [3], [8].

Indian languages, Hindi and Malayalam, differ considerably in their linguistic structure, typology, and morphological systems. Hindi, an Indo-Aryan language, is written in the Devanagari script and exhibits moderate morphological richness, while Malayalam, a Dravidian language, uses the Malayalam script and is characterised by highly agglutinative and inflectional morphology, particularly in verbs and noun declensions [9].

Despite promising developments, MNMT still faces several limitations, especially when applied to Indian languages with

their rich morphological structures and syntactic diversity [7], [10]. The generalised nature of multilingual models can lead to underfitting or negative transfers, where shared parameters dilute language-specific nuances, resulting in performance degradation for selected languages [5]. Furthermore, the lack of a balanced and representative multilingual corpus exacerbates the issue, as MNMT tends to favour high-resource languages during training while marginalising less-represented ones [11], [12].

### A. Contributions

*1) Hindi-Malayalam Test Benchmark for translation:* Despite progress in multilingual NMT, parallel resources for the Hindi-Malayalam pair remain limited and lack controlled variation in sentence difficulty [12]–[14]. To address this, we introduce the MHM benchmark, a curated test dataset derived from textbooks that covers multiple readability levels, enabling fine-grained evaluation.

*2) Systematic Evaluation of Bilingual Translation Quality of MNMT Models:* Although multilingual models achieve strong performance across many languages, their behaviour on typologically distant pairs such as Hindi-Malayalam is not well understood, particularly across varying difficulty levels [12], [13]. We offer a comprehensive evaluation of major MNMT models on this pair using BLEU, chrF2++, and COMET across the proposed benchmarks.

*3) Lightweight Direct Translation Models:* Large-scale MNMT models often struggle with semantic fidelity and syntactic alignment for low-resource, morphologically divergent language pairs. To overcome these limitations, we develop a compact transformer-based bilingual model specifically tuned for Hindi-Malayalam, demonstrating competitive performance with significantly lower computational cost.

The remainder of the section is as follows. Section II reviews the multilingual models that are considered for the comparison of translation accuracy for the Hindi-Malayalam pair. Section III describes the steps involved in developing the translation models, along with the details of training and evaluation. The results are discussed in Section IV followed by the conclusion in Section V.

## II. Related Works

In this section, prominent multilingual translation models are briefly reviewed, focussing on their architectural features and training strategies. MNMT models based on transformers with dedicated Indian language support came to picture with BART [15].

mBART-50 [16] is a sequence-to-sequence auto-encoder model that underwent two stages of training: supervised fine-tuning for many-to-many translation across 50 languages and multilingual denoising on sizable unlabelled corpora. The use of shared positional embeddings, language-specific prefix tags, and a SentencePiece vocabulary enables strong zero-shot translation.

M2M100 [8], with 7.5 billion phrases in 100 languages and balanced sampling, provided direct many-to-many translation over 2,200 language pairs without the need for language tokens or English as a pivot.

TABLE I
Language models description

| Name of the Model | Parameters | Embedding Size | Max Seq Length | Vocabulary Size |
|---|---|---|---|---|
| NLLB-200 [2] | 1.3B | 1024 | 512 | 256K |
| MBART-50 [16] | 610M | 1024 | 1024 | 250K |
| m2m-100 [8] | 418M | 1024 | 1024 | 128K |
| IndicTrans2 [3] | 1.1B | 1024 | 256 | 128K |

NLLB-200 [2] greatly increased multilingual capabilities by training in more than 3.3 billion sentence pairs from 200 languages [17] using filtered CommonCrawl and curated datasets. Using adaptive sampling and knowledge distillation, it has a compact model that includes language-specific prefixes and a smaller version of the 1.3B parameter for easier use.

The most recent of all of these models, IndicTrans2 [3], uses a 6-layer transformer trained on bilingual and back-translated corpora from 22 Indic languages with an emphasis on Indian languages and English. Task-specific SentencePiece tokenisers and domain-adaptive fine-tuning enable effective translation in both general and low-resource domains. It is especially well-suited to the Indian language environment due to its support for numerous domains and mixed-script inputs.

With their increased language coverage and training effectiveness, each of these models represents a step forward in multilingual neural machine translation. Table I gives the details of the embedding and vocabulary size, the maximum sequence length, and the total number of parameters of the translation models described above.

The model specifications indicate that they are indeed "large" in terms of their specifications and require significant time and resources for training. These models are tested for Hindi-Malayalam language pair translation, providing a detailed analysis whose results are presented in Section IV.

## III. Modelling and Evaluation Framework

The multilingual neural machine translation conceptualises as a sequence-to-sequence task built on the transformer architecture [18], where an encoder processes an input sequence in the source language and a decoder produces the output sequence in the designated target language. Given a source sentence $\mathbf{x}$, the corresponding target sentence $\mathbf{y}$, and the specified source-target language pair, the model is trained to maximise the likelihood of generating the correct target sequence. This objective is expressed as:

$$\theta^* = \arg\max_{\theta} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \sum_{t=1}^{T} \log P_\theta(y_t \mid y_{<t}, \mathbf{x}) \qquad (1)$$

where $\mathcal{D}$ denotes the parallel training corpus, $\theta$ represents the model parameters, $T$ is the length of the target sequence, and $y_{<t}$ is the partial target sequence generated before time step $t$.

This Section discusses the data involved in model training, followed by the preprocessing and tokenization techniques

TABLE II
TRAINING DATASET DETAILS

| Dataset | Domain | #Sentences |
|---|---|---|
| NLLB [2], [17] | Generic | 1.8M |
| Subtitles [19] | Conversational | 0.3M |
| Shiksha [20] | Academic | 67K |

involved in processing the sentences. It then presents the model's training and architectural details, concluding with a description of the evaluation benchmarks and metrics.

### A. Training data

The training data for the prototype model consist of three separate sources: the NLLB dataset [2], [17], the movie subtitle dataset [19] and the Shiksha dataset [20]. Table II provides details about the datasets.

The NLLB dataset [2], [17] is an extensive and high-quality corpus of aligned text pairings over 200 languages, specifically designed to facilitate translation for low-resource languages. It offers comprehensive coverage for Indian languages, including Hindi and Malayalam.

The dataset of multilingual movie subtitles [19] includes pairs extracted from publicly accessible films and television shows, along with extensive contextual and conversational data. This source documents colloquial idioms, cultural subtleties, and informal speech patterns, which are generally underrepresented in formal corpora.

The Shiksha dataset [20] is a domain-specific translation dataset focused on the educational and technical sectors, designed to support Indian languages. The dataset includes aligned texts in Hindi, Malayalam, and other Indian languages, drawn from academic materials, instructional content, and scientific documentation. Together, these datasets ensure that the model learns both structured and natural languages, making it stronger and more adaptable in different areas.

### B. Preprocessing and Tokenization

The training data undergoes standard preprocessing, which includes removing redundant white spaces, duplicate sentences, and special characters. The tokenization is performed using SentencePiece [21] with Byte Pair Encoding (BPE) [22], which breaks words into subword tokens. This technique helps deal with morphologically rich languages, where a large number of words exist outside the vocabulary [23].

The embedding size is 512, with a vocabulary of 32,000 words, developed from the training data. The maximum sequence length is set to 350 tokens, enabling the model to handle relatively long sentences. Special tokens such as [PAD] (padding), [SOS] (start of sequence), and [EOS] (end of sequence) are also included to manage input and output sequences effectively.

### C. Architecture and Training

The base translation model is a transformer-based sequence-to-sequence architecture optimised for low-resource translation between Hindi and Malayalam. The model integrates standard Transformer components: multi-head self-attention, position encoding, and deep feedforward layers, augmented with tokenization and training strategies tailored for morphologically rich languages [18]. Table III describes the configuration of the base model.

TABLE III
BASE TRANSFORMER MODEL: ARCHITECTURE AND TRAINING CONFIGURATION

| Component | Details |
|---|---|
| Architecture | Transformer (Encoder-Decoder) |
| Encoder-Decoder Layers | 6/6 |
| Attention Heads per Layer | 8 |
| Hidden Dimension | 512 |
| Feedforward Layer Size | 2048 |
| Embedding Size | 512 |
| Vocabulary Size | 32,000 |
| Tokenization Technique | SentencePiece (Byte Pair Encoding) |
| Special Tokens | [PAD], [SOS], [EOS] |
| Positional Encoding | Sinusoidal |
| Max Sequence Length | 350 tokens |
| Optimizer | Adam |
| Learning Rate Scheduler | Inverse Square Root Decay |
| Loss Function | Cross-Entropy with Label Smoothing |
| Dropout | Applied to Encoder and Decoder |
| Layer Normalization | Applied Throughout |
| Training-Validation Split | 90% - 10% |
| Total Parameters | 75M approx. |

TABLE IV
DETAILS OF EVALUATION BENCHMARKS

| Dataset | Category of text | #Sentences |
|---|---|---|
| IN22 [3] | Wikipedia, Web, Conversation data | 2527 |
| FLORES-200 [2] | Generic text | 1000 |
| MHM | Academic textbook contents | 120 |

Two prototype models, MHM-P1 and MHM-P2, were developed first to assess translation quality and serve as baseline models. MHM-P1 is trained using the movie subtitle dataset, which consists of 0.3 million sentences. MHM-P2 uses larger generic data. The model is trained on the NLLB and Shiksha datasets, which comprise 1.9M sentences in size.

### D. Evaluation Benchmarks

The models are evaluated on the following publicly available benchmarks: IN22 [3] and FLORES-200 [2]. IN22 [3] is a specific benchmark set for assessing machine translation accuracy in multi-domain, n-way parallel scenarios that span 22 Indic languages. It consists of three separate subsets: IN22-Wiki, IN22-Web, and IN22-Conv. The data include sentences from Wikipedia, various domain websites, and everyday conversations. FLORES-200 [2] is a multi-domain general-purpose benchmark created to evaluate translations in 200 languages, encompassing 19 Indic languages. The English sentences are original sources which are translated into other languages.

This work also presents a new dataset (*MHM dataset*) that consists of two custom datasets specifically designed to assess the translation performance for the Hindi-Malayalam language pair. Both datasets comprise lesson content extracted from textbooks covering classes 1 through 12, encompassing

TABLE V
DESCRIPTION OF EVALUATION METRICS FOR TRANSLATION MODELS

| Metric | Type | Level | Strengths | Limitations | Direction |
|---|---|---|---|---|---|
| BLEU | N-gram precision based (0 to 100 range) | Word-level | Fast, standardized, easy to interpret | Insensitive to meaning, penalizes legitimate variations | Higher is better |
| chrF2++ | Character-level F-score (0 to 100 range) | Subword/ character | Handles morphology, robust in low-resource languages | Less intuitive, not suitable for measuring semantic adequacy | Higher is better |
| COMET | Learned regression model (0 to 1 range) | Semantic/ contextual | High correlation with human judgment, captures fluency | Requires pretrained models, computationally expensive | Higher is better |

a broad range of topics and sentence complexities. Translations are performed by experts who have knowledge of both languages. Table IV gives the details of the evaluation datasets.

### E. Evaluation metrics

Translated sentences are evaluated against the reference sentences using BLEU [24], chrF2++ [25], and COMET [26]. They are widely adopted evaluation metrics in machine translation, each reflecting different perspectives on translation quality [11]. Table V provides a brief description of their characteristics.

## IV. Results and Discussions

This section presents the evaluation outcomes and interprets the comparative performance of our translation model prototypes, highlighting its accuracy across three key test benchmarks: IN22 [3], FLORES-200 [2] and MHM. Each models (existing and new) are tested for the language pairs (Hindi and Malayalam) in both directions using the three metrics (BLEU [24], ChrF2++ [25] and COMET [26]).

### A. Results

The bidirectional evaluation on the MHM dataset is detailed through two tables. Tables VI and VII report model performance for the Hindi to Malayalam and Malayalam to Hindi translation directions, respectively, based on all metrics.

On the MHM dataset, MHM-P2 achieves a BLEU of 5.0 and 2.1, outperforming m2m-100 (0.6 and 4.1) in the Hindi to Malayalam direction and achieving comparable COMET scores (0.72 vs. 0.69 and 0.60 vs. 0.59).

In the FLORES-200 benchmark, MHM-P2's chrF2++ scores of 30.6 surpass m2m-100's 20.4 in Hindi to Malayalam direction. The COMET scores of 0.65 and 0.53 remain close to MBART-50's 0.72 and 0.67, despite BLEU being slightly lower.

In the IN22 dataset, MHM-P2 achieves BLEU scores of 2.4 (Hi to Ml) and 7.0 (Ml to Hi), while m2m-100 scores 9.0 in both directions, and MBART-50 achieves only 4.5 and 10.5, showing that MHM-P2 approaches the performance of the latter with significantly fewer parameters. The COMET scores of MHM-P2 surpass the m2m-100 in both directions.

### B. Discussions

IndicTrans2 model has established itself as the strongest model, consistently outperforming other assessed models in both translation directions and across all datasets. The model obtained the greatest chrF2++ scores on Hindi-Malayalam and Malayalam-Hindi translations in the FLORES-200 dataset. The model's sophisticated architecture and large multilingual corpus, fine-tuned for Indian languages, are the major factors behind this. However, such sophistication entails significant computational demand. Figures 1 and 2 show the averaged BLEU, ChrF2++, and COMET scores for all models in the Hindi to Malayalam and Malayalam to Hindi directions, respectively, with all metrics normalized for comparison.

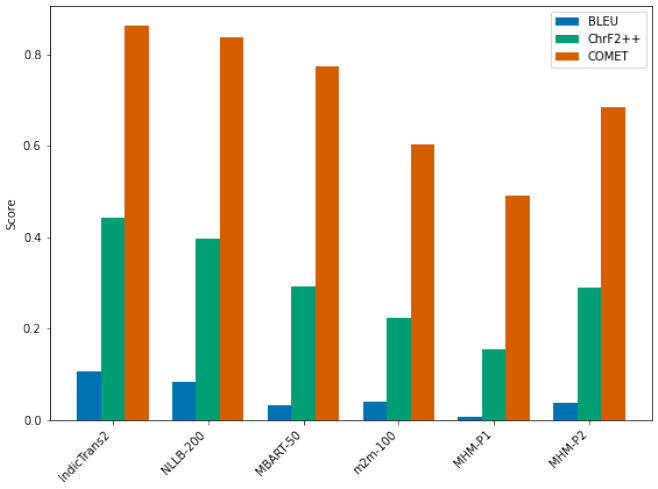


Fig. 1. Normalized comparison of average BLEU, ChrF2++, and COMET metrics for Hindi to Malayalam translation models, aggregated over the IN22, FLORES-200, and MHM datasets

MHM-P2 shows promising performance when evaluated against established multilingual models, particularly given its compact architecture and limited training data. Although the MHM-P1 and MHM-P2 models generally underperform compared to large-scale multilingual systems, their results affirm the viability of lightweight prototypes in low-resource translation tasks. In particular, MHM-P2 achieves COMET scores that are close to those of substantially larger models across similar language pairs, reinforcing its semantic adequacy.

These results suggest that such bilingual prototype models can serve as effective baselines and a strong starting point for future refinement. Furthermore, their minimal computational requirements make them especially suitable for rapid iteration and deployment in environments where large-scale models are infeasible.

TABLE VI
COMPARISON OF BLEU, CHRF2++, AND COMET SCORES FOR HINDI TO MALAYALAM TRANSLATION MODELS EVALUATED ON THE IN22, FLORES-200, AND MHM DATASETS

| Model | IN22 | | | FLORES-200 | | | MHM | | |
|---|---|---|---|---|---|---|---|---|---|
| | BLEU | ChrF2++ | COMET | BLEU | ChrF2++ | COMET | BLEU | ChrF2++ | COMET |
| IndicTrans2 | 8.3 | 39.3 | 0.86 | 12.6 | 47.0 | 0.86 | 11.5 | 46.3 | 0.87 |
| NLLB-200 | 7.3 | 36.5 | 0.83 | 8.8 | 40.3 | 0.83 | 9.1 | 42.4 | 0.85 |
| MBART-50 | 4.5 | 29.3 | 0.78 | 2.8 | 26.8 | 0.72 | 2.4 | 31.7 | 0.82 |
| m2m-100 | 9.0 | 28.1 | 0.59 | 2.8 | 20.4 | 0.53 | 0.6 | 18.7 | 0.69 |
| MHM-P1 | 1.4 | 19.0 | 0.48 | 0.2 | 12.5 | 0.40 | 0.7 | 15.4 | 0.59 |
| MHM-P2 | 2.4 | 27.2 | 0.68 | 4.4 | 30.6 | 0.65 | 5.0 | 29.3 | 0.72 |

TABLE VII
COMPARISON OF BLEU, CHRF2++, AND COMET SCORES FOR MALAYALAM TO HINDI TRANSLATION MODELS EVALUATED ON THE IN22, FLORES-200, AND MHM DATASETS

| Model | IN22 | | | FLORES-200 | | | MHM | | |
|---|---|---|---|---|---|---|---|---|---|
| | BLEU | ChrF2++ | COMET | BLEU | ChrF2++ | COMET | BLEU | ChrF2++ | COMET |
| IndicTrans2 | 19.0 | 41.1 | 0.65 | 24.7 | 49.0 | 0.75 | 12.9 | 28.5 | 0.79 |
| NLLB-200 | 20.1 | 42.6 | 0.76 | 23.2 | 46.9 | 0.35 | 25.8 | 44.7 | 0.80 |
| MBART-50 | 10.5 | 32.0 | 0.57 | 12.7 | 35.3 | 0.67 | 7.0 | 20.5 | 0.73 |
| m2m-100 | 9.0 | 28.1 | 0.48 | 12.3 | 34.7 | 0.58 | 4.1 | 16.0 | 0.59 |
| MHM-P1 | 4.4 | 18.1 | 0.37 | 1.8 | 12.4 | 0.35 | 0.4 | 8.8 | 0.50 |
| MHM-P2 | 7.0 | 23.9 | 0.49 | 9.5 | 30.3 | 0.53 | 2.1 | 15.0 | 0.60 |

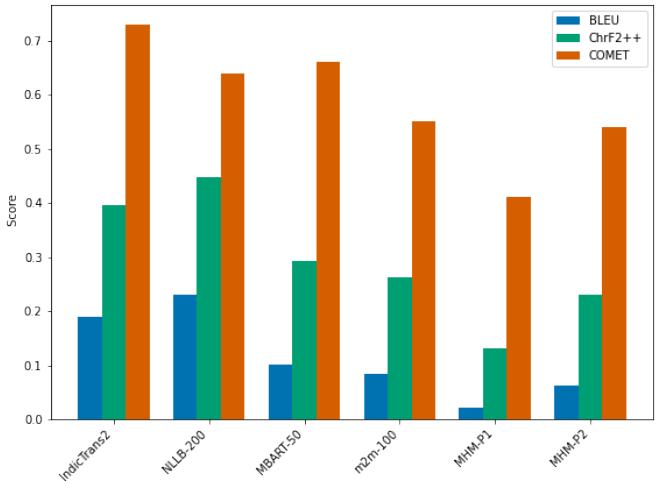


Fig. 2. Normalized comparison of average BLEU, ChrF2++, and COMET metrics for Malayalam to Hindi translation models, aggregated over the IN22, FLORES-200, and MHM datasets

## V. CONCLUSION

It is quite difficult to accurately translate Hindi and Malayalam, since they have numerous subtle variations, complex grammatical structures, and nuances that vary according to context. Most of the time, translation models use English as an intermediary language, which results in missing the cultural and semantic nuances involved and leading to errors and loss of meaning.

This work provided an in-depth review of the most recent translation models, evaluation metrics, and performance analysis for Hindi-Malayalam language pairs. Transformer-based architectures work better when generic multilingual models are appropriately calibrated using high-quality parallel corpora. The prototype models based on the transformer-based approach showed promise; however, the results indicate that the models need to grasp the semantics and syntax of these languages even better. The work provides a strong foundation for improving translation quality by modelling languages directly, without using English as a pivot.

Future work will prioritise expanding and diversifying the training dataset while maintaining the lightweight, energy-efficient design. This approach supports both sustainable development and equitable access to high-quality translation tools. By strengthening resource modelling for low-resource languages, this work contributes to the broader goals of being responsible and inclusive.

## ACKNOWLEDGMENT

The authors utilised AI tools to aid in language editing. All contents were reviewed, validated, and approved by all authors.

## REFERENCES

[1] F. Stahlberg, "Neural machine translation: A review," *Journal of Artificial Intelligence Research*, vol. 69, pp. 343–418, oct 2020.

[2] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "No Language Left Behind: Scaling Human-Centered Machine Translation," aug 2022.

[3] AI4Bharat, J. Gala, P. A. Chitale, R. AK, S. Doddapaneni, V. Gumma, A. Kumar, J. Nawale, A. Sujatha, R. Puduppully, V. Raghavan, P. Kumar, M. M. Khapra, R. Dabre, and A. Kunchukuttan, "IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages," may 2023.

[4] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Stroudsburg, PA, USA), pp. 6282–6293, Association for Computational Linguistics, apr 2020.

[5] J. Pang, F. Ye, L. Wang, D. Yu, D. F. Wong, S. Shi, and Z. Tu, "Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models," *Transactions of the Association for Computational Linguistics*, vol. 13, pp. 73–95, jan 2024.

[6] F. Richter, "The Most Spoken Languages: On the Internet and in Real Life," 2024.

[7] S. Bala Das, D. Panda, T. Kumar Mishra, B. Kr. Patra, and A. Ekbal, "Multilingual Neural Machine Translation for Indic to Indic Languages," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, pp. 1–32, may 2024.

[8] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond English-Centric Multilingual Machine Translation," *Journal of Machine Learning Research*, vol. 22, pp. 1–48, oct 2020.

[9] E. Encyclopaedia Britannica, "Indian languages — Definition & Facts — Britannica," 2025.

[10] M. Singh, R. Kumar, and I. Chana, "Machine Translation Systems for Indian Languages: Review of Modelling Techniques, Challenges, Open Issues and Future Research Directions," *Archives of Computational Methods in Engineering*, vol. 28, pp. 2165–2193, jun 2021.

[11] T. O. Tafa, S. Z. M. Hashim, M. S. Othman, H. Alhussian, M. Nasser, S. J. Abdulkadir, S. H. Huspi, S. O. Adeyemo, and Y. A. Bena, "Machine Translation Performance for LowResource Languages: A Systematic Literature Review," *IEEE Access*, 2025.

[12] J. Pang, F. Ye, L. Wang, D. Yu, D. F. Wong, S. Shi, and Z. Tu, "Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models," *Transactions of the Association for Computational Linguistics*, vol. 13, pp. 73–95, dec 2024.

[13] R. Raja and A. Vats, "Parallel Corpora for Machine Translation in Low-resource Indic Languages: A Comprehensive Review," pp. 129–143, apr 2025.

[14] S. B. Das, D. Panda, T. K. Mishra, and B. K. Patra, "Statistical machine translation for Indic languages," *Natural Language Processing*, vol. 31, pp. 328–345, mar 2025.

[15] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, dec 2020.

[16] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning," aug 2020.

[17] H. Schwenk, G. Wenzek, S. Edunov, E. Grave, A. Joulin, and A. Fan, "CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB," *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, vol. 1, pp. 6490–6500, nov 2019.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 2017-Decem, pp. 5999–6009, aug 2017.

[19] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pp. 923–929, 2016.

[20] A. Joglekar and S. Umesh, "Shiksha: A Technical Domain focused Translation Dataset and Model for Indian Languages," dec 2024.

[21] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Stroudsburg, PA, USA), pp. 66–71, Association for Computational Linguistics, aug 2018.

[22] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, vol. 3, pp. 1715–1725, 2016.

[23] V. M R and P. Chandran, "Analysis of Subword based Word Representations Case Study: Fasttext Malayalam," in *2022 IEEE 19th India Council International Conference (INDICON)*, pp. 1–6, IEEE, nov 2022.

[24] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 2002-July, pp. 311–318, 2002.

[25] M. Popović, "Chrf: Character n-gram f-score for automatic mt evaluation," in *10th Workshop on Statistical Machine Translation, WMT 2015 at the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015 - Proceedings*, pp. 392–395, Association for Computational Linguistics (ACL), 2015.

[26] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 2685–2702, Association for Computational Linguistics (ACL), 2020.