

# BharatSetu: A $1\text{M} \times 22$ Bridge for Indic Translation

**Haq Nawaz Malik\***

ORCID: 0009-0003-1994-7640

**Nahfid Nissar**

ORCID: 0009-0002-2805-4687

**Faizan Iqbal**

ORCID: 0009-0002-8998-9347

## Abstract

India’s Eighth Schedule recognises 22 languages, yet parallel data is uneven across them, and many language pairs, especially among lower-resource languages such as Bodo, Dogri, Santali and Kashmiri, lack direct bitext. We present BharatSetu 22-Indic 1M, a row-aligned table of 984,764 English sentences, each paired with machine translations into all 22 scheduled languages (21,664,808 Indic cells), generated with the sarvamai/sarvam-translate model family. Because every Indic cell in a row translates the same English sentence, any two columns form a translation pair: 462 directed Indic-to-Indic pairs, or 506 including English, without direct bitext. We call these *pivot-derived pairs* and propose a confidence layer that lets users choose how strictly each pair is connected. The corpus is machine-generated and not human-verified; quality is expected to vary by language, domain and sentence style, and pivot-derived pairs can inherit errors from both sides. We therefore release BharatSetu as a machine layer and propose a human-correction layer keyed to the same row identifiers, so that each corrected column upgrades every pair that uses it, together with an evaluation protocol. We report no translation-quality results.

## 1 Introduction

This section states the resource gap that BharatSetu addresses and defines the contribution made by a single row-aligned table across all scheduled Indian languages. The Eighth Schedule of the Constitution of India lists 22 languages; the Government of India lists Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu as the languages included in the Schedule ([Government of India, Department of Official Language, 2026](#)). Modern machine translation (MT) resources, however, do not cover these

languages with equal density. Widely used Indic corpora and models have improved coverage substantially, but many resources still concentrate on higher-resource languages or expose data as separate bilingual files rather than as one uniform multiway table ([Kakwani et al., 2020](#); [Ramesh et al., 2022](#); [Gala et al., 2023](#)).

The practical problem is not only that some individual languages have limited data. It is also that researchers often lack a common sentence spine across all 22 scheduled languages. Without such a spine, comparisons across languages require reconciling different domains, sentence selections, identifiers, preprocessing conventions and file formats. These differences make it difficult to train one model under controlled conditions, to study transfer from higher-resource to lower-resource languages, or to create pivot-derived pairs that share the same underlying meaning.

BharatSetu 22-Indic 1M addresses this problem by packaging 984,764 English source rows with one aligned machine-translated target column for each of the 22 scheduled languages. In scale terms, this is approximately  $1\text{M}$  English rows  $\times$  22 Indic columns, producing 21,664,808 machine-generated Indic cells, or about 21.7M target-language cells. Each row has an id aligned to the original source corpus, the English source sentence, a source provenance label and 22 full-language-name target columns. The dataset is distributed publicly open-source through the Hugging Face repository [Omarrran/BharatSetu-22Indic-1M](#) ([Malik et al., 2026](#)).

The central claim of this paper is deliberately narrow. BharatSetu is valuable because it combines breadth, uniformity and scale in one table. It is not presented as a human-verified gold corpus, and we do not report BLEU, chrF, COMET or human-evaluation results. The translations were generated with the sarvamai/sarvam-translate model family, and the released corpus should be

understood as a machine layer whose utility must be tested and improved through the evaluation and correction plan described in Sections 7 and 8.

Our contributions are:

- **The need: one table for all 22 languages.** We release BharatSetu, 984,764 English sentences aligned to machine translations in all 22 scheduled languages, including languages with little parallel data such as Bodo, Dogri, Santali and Kashmiri.
- **The use: any Indic language to any other.** Because every language translates the same English sentence, any two columns form a translation pair: 462 Indic-to-Indic directions with no direct bitext, each with a confidence score so users can choose how strict to be.
- **The path to gold quality.** A human-correction layer on the same row identifiers upgrades every pair that uses a corrected column, and we propose how to test its effect on trained models.

The next section positions BharatSetu against existing multilingual MT corpora, models and benchmarks.

## 2 Related Work

This section identifies the closest resource families and explains why BharatSetu fills a different slot: all 22 scheduled languages aligned to the same English rows in one table. Indic NLP resource building has progressed through monolingual corpora, benchmarks and multilingual models. IndicNLP-Suite introduced monolingual corpora, evaluation benchmarks and pretrained multilingual language models for 11 Indian languages (Kakwani et al., 2020). Samanantar collected large English–Indic parallel corpora for 11 Indic languages and used an English-centric corpus to derive Indic–Indic pairs through pivoting (Ramesh et al., 2022). These resources established the importance of scale and shared infrastructure for Indic NLP, but they do not provide a single 22-language scheduled-language table with one row per English source sentence.

Recent model work has moved closer to full scheduled-language coverage. IndicTrans2 targets all 22 scheduled Indian languages and emphasises accessible MT models and data for those languages (Gala et al., 2023). Sarvam Translate is documented by Sarvam AI as a translation model supporting English–Indian-language

translation across the official Indian-language set, and the BharatSetu generation process used the sarvamai/sarvam-translate model family (Sarvam AI, 2026). BharatSetu is complementary to such systems: it is a derived data resource for training, evaluation design, auditing and correction workflows, not a new translation model.

Global multilingual MT research has also demonstrated the importance of many-language coverage and shared benchmarks. M2M-100 moved beyond English-centric translation by training a many-to-many multilingual model (Fan et al., 2021). mBART showed that multilingual sequence-to-sequence pretraining can support MT, including low-resource settings (Liu et al., 2020). NLLB-200 and FLORES-200 scaled model and evaluation coverage to 200 languages and more than 40,000 translation directions, with human-translated evaluation data and safety analysis (NLLB Team et al., 2024). More recent Omnilingual MT work extends the ambition of multilingual MT beyond the 200-language scale (Omnilingual MT Team et al., 2026). BharatSetu differs from these efforts in scope: it focuses on the administratively defined set of 22 scheduled Indian languages and on a uniform English-source row alignment rather than on a universal MT system.

Synthetic and machine-translated corpora have a long role in low-resource MT, especially when direct bitext is scarce. Back-translation and self-training are common ways to use model-generated text for training, but they require careful evaluation because machine-generated data can propagate errors and domain bias (Sennrich et al., 2016; Edunov et al., 2018). BharatSetu follows this tradition only as a machine layer. Its intended contribution is a structured starting point for experiments and correction, not a claim that synthetic translations are equivalent to human reference translations. Section 3 therefore describes the construction process before Section 4 specifies the table users will inspect.

## 3 Dataset Construction

This section describes how BharatSetu preserves row alignment from English source rows through translation, merge and release packaging. The construction process starts from an English source corpus assembled from sources such as literature and news, alongside resources such as Wikipedia-derived WikiMatrix, OPUS, C4 and Tatoeba

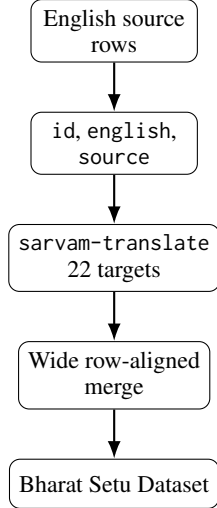


Figure 1: BharatSetu construction pipeline. The release keeps source-row identity, machine-generated target columns and packaging metadata separate so that provenance, translation quality and integrity can be audited independently.

(Schwenk et al., 2021; Tiedemann, 2012; Dodge et al., 2021; Tiedemann, 2020). Each original row is represented by an id, an English sentence and a source provenance label. The id field is retained so that translations can be traced back to the original source-corpus row and so that future correction passes can update a row without changing its identity.

For each English row, translations were generated into the 22 scheduled-language target columns using the sarvamai/sarvam-translate model family. The resulting files were merged into a wide table with one row per English sentence. This wide format is important because it keeps all target-language outputs for a sentence adjacent, making cross-language comparison, correction and pair extraction a row-local operation.

Figure 1 summarises the construction pipeline. The figure intentionally distinguishes source provenance, machine translation and packaging, because these layers support different future checks.

The release contains sharded Parquet files under data/train-\*.parquet, a single-file Parquet export, JSONL and XLSX copies, metadata.json and generation/merge\_manifest.json. These files serve different users: Parquet supports efficient training and filtering, JSONL supports streaming and line-based tooling, XLSX supports manual inspection, and metadata files support release review.

No human verification is claimed for the trans-

| Column                  | Description                                                                                    |
|-------------------------|------------------------------------------------------------------------------------------------|
| id                      | Identifier aligned to the original source-corpus row.                                          |
| english source          | English source sentence. Provenance label from the input corpus.                               |
| Target language columns | One machine-generated translation column per target language, named by the full language name. |

Table 1: BharatSetu row schema. Each row contains source identity, English text, source provenance and 22 target-language columns.

lations. This construction choice affects all downstream claims. BharatSetu can support experiments, sampling and correction workflows immediately, but it should not be used as a gold reference corpus without additional validation. The next section specifies the schema and language coverage that downstream users will see.

## 4 Dataset Description

This section defines the table, columns and language coverage in BharatSetu 22-Indic 1M. Table 1 shows the schema. The required invariant is that each row contains one English source sentence and the 22 generated translations aligned to that sentence.

Table 2 lists the target languages and codes supplied in the fact sheet. Script conventions are not asserted as dataset facts because several scheduled languages have multiple orthographic traditions or release-specific script choices. Script handling should therefore be treated as a release-level audit item rather than an assumed property of the model output.

The corpus contains 984,764 English source rows and 22 target-language columns, yielding 21,664,808 machine-generated Indic cells. The public dataset is hosted on [Hugging Face](#) (Malik et al., 2026) and distributed in sharded Parquet, single-file Parquet, JSONL and XLSX formats.

The row count and target count imply the pair structure used later. There are 22 English-to-Indic directions and 22 Indic-to-English directions if the English column is used as a target, for 44 English-connected directions. For Indic-to-Indic use, each source language can be paired with each of the other 21 target languages, giving  $22 \times 21 = 462$  directed pivot-derived pairs. The total directed directions available from the aligned table are therefore  $44 + 462 = 506$ . Section 5 explains why this

| Language  | Code | Layer   | Language | Code | Layer   |
|-----------|------|---------|----------|------|---------|
| Assamese  | as   | machine | Bengali  | bn   | machine |
| Bodo      | brx  | machine | Dogri    | doi  | machine |
| Gujarati  | gu   | machine | Hindi    | hi   | machine |
| Kannada   | kn   | machine | Kashmiri | ks   | machine |
| Konkani   | kok  | machine | Maithili | mai  | machine |
| Malayalam | ml   | machine | Manipuri | mni  | machine |
| Marathi   | mr   | machine | Nepali   | ne   | machine |
| Odia      | or   | machine | Punjabi  | pa   | machine |
| Sanskrit  | sa   | machine | Santali  | sat  | machine |
| Sindhi    | sd   | machine | Tamil    | ta   | machine |
| Telugu    | te   | machine | Urdu     | ur   | machine |

Table 2: Language coverage of BharatSetu. The 22 listed languages correspond to the Eighth Schedule set. Script conventions in the generated columns should be audited by the authors before public release.

arithmetic matters for actual research workflows rather than only for dataset description.

## 5 Why Row-Aligned Multiway Data Solves Real Problems

Row alignment makes one table useful for training, pivoting, transfer, analysis and distillation across all scheduled languages. A bilingual corpus supports one source–target direction; a collection of unrelated bilingual corpora supports several, but not on shared domains or sentences. A row-aligned table supports controlled extraction of many directions from one sentence spine.

The first use case is English-to-22 and 22-to-English training. Every row has an English source and one target column per scheduled language, so a user can build 22 English-to-Indic training files with identical row identifiers and provenance labels, and reverse each for Indic-to-English training. The Indic side is machine-generated, whereas the English side is original source text. Using the same rows for every language reduces confounding by row selection when languages are compared.

The second use case is Indic-to-Indic translation through an English pivot (Figure 2). Panel (a) shows the construction: each English sentence is translated into the 22 Indic targets, giving one English cell and 22 aligned Indic cells per row. Panel (b) shows the consequence: any two Indic cells in a row can be paired, for example Kashmiri–Santali or Bodo–Tamil, with no further translation step. These are *pivot-derived pairs*, not independently human-translated bitext. Their semantic link is mediated by the shared English source, and because both sides are machine translations, errors on either side can carry into the pair; pair quality is therefore a question to be evaluated. For 22 Indic targets there are  $22 \times 21 = 462$  directed Indic-to-

Indic pairs; adding the 44 English directions (22 in each direction) gives 506, which equals  $23 \times 22$  for 23 aligned columns.

The third use case is transfer for lower-resource scheduled languages. Bodo, Dogri, Santali, Kashmiri and Manipuri are often harder to serve with large parallel corpora than Hindi, Bengali, Tamil or Telugu. A uniform table lets researchers test whether multilingual training transfers representations from higher-resource to lower-resource columns under a shared row distribution; this is an open empirical question, not a result of the corpus.

The fourth use case is controlled cross-lingual analysis. Because every target cell in a row aligns to one English sentence, researchers can compare how languages handle named entities, numerals, punctuation, honorifics, morphology and word order. The same structure supports language-identification experiments, transliteration studies, script-conversion audits and consistency checks for multilingual generation, with the assurance that each row shares one source meaning even when target quality varies.

The fifth use case is synthetic-data distillation. A model trained on BharatSetu can be compared with the original translation model family, an IndicTrans2 baseline, or a general multilingual MT model such as NLLB-200, to ask whether a compact or domain-adapted model recovers useful performance from a uniform synthetic table, and whether it inherits or amplifies errors from the machine layer. This leads to the human correction roadmap in Section 6.

## 6 Human Correction Layer: Roadmap

This section proposes a correction design that improves the machine layer without misrepresenting

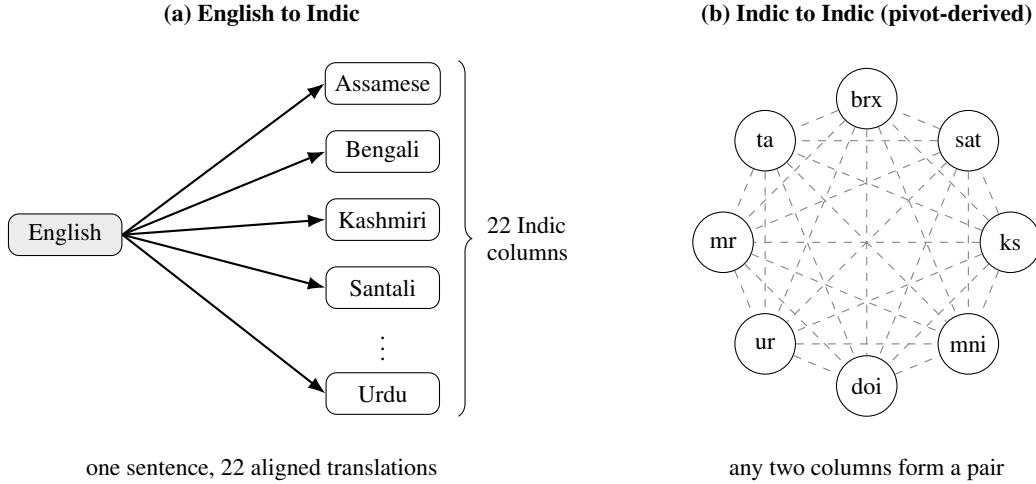


Figure 2: The two stages of the row-aligned design. (a) English to 22 Indic columns. (b) Pivot-derived Indic-to-Indic pairs. Eight of the 22 languages are shown in (b) by ISO code; dashed lines apply in both directions.

the current release as human-verified. The simplest model is layered. Tier 0 is the released machine layer: 984,764 rows, 22 target columns and no human verification claim. Tier 1 is a sampled human-verified evaluation layer for selected languages and domains. Tier 2 is a set of fully corrected priority-language columns or row subsets. Tier 3 is a versioned release in which corrected cells coexist with machine cells and preserve the same row identifiers.

The key advantage of a correction layer is that each corrected cell has network effects. If the Tamil column is corrected for a row, then English–Tamil improves directly and every pivot-derived pair involving Tamil receives a better Tamil side. If a new corrected Kashmiri subset is added, it improves English–Kashmiri, Kashmiri–English and all Kashmiri–other-language pivot-derived pairs for those rows. The benefit scales by connection rather than by one isolated bilingual file.

Human correction should be implemented as an auditable annotation process. The authors should define language-specific guidelines, annotator qualifications, compensation policy, adjudication rules, version identifiers and inter-annotator agreement measures. For languages with script variation, the guidelines must specify the expected script or permit multiple scripts with explicit labels. Corrected columns should not overwrite the machine columns silently; they should be stored as additional versioned fields or release variants so that experiments can compare the machine layer against the human correction layer.

The main research hypotheses are as follows.

First, training on the machine layer may outperform having no in-domain or no aligned data for lower-resource scheduled languages. Second, training on a corrected subset may improve quality relative to training on the machine layer alone. Third, correcting a small number of strategically chosen languages may improve pivot-derived pairs that use those languages. These hypotheses require measurement before they can be asserted.

## 7 Recommended Evaluation Protocol for Users

This section reports no results. It gives a protocol that we recommend to anyone who uses Bharat-Setu, whether for training, evaluation or analysis, because the corpus is machine-generated and its quality has not been measured. The protocol combines data-integrity checks, automatic MT metrics, human evaluation and correction-ablation experiments. Where reference translations exist, users should score with BLEU, chrF and COMET (Papineni et al., 2002; Popović, 2015; Rei et al., 2020), and report metric choices with caveats for morphology, script variation and domain mismatch. Human evaluation should cover adequacy, fluency, error typology and harmful-content or bias checks, following error-analysis practice such as MQM where qualified annotators are available (Freitag et al., 2021).

Table 3 summarises the recommended protocol. We advise running every experiment on a fixed dataset version and documented model checkpoints.

Users who train models should run four com-

| Evaluation target        | Directions                                                | Data condition                                            | Metrics or judgments                                                 |
|--------------------------|-----------------------------------------------------------|-----------------------------------------------------------|----------------------------------------------------------------------|
| Machine-layer MT quality | English→22 and 22→English                                 | Released machine layer; fixed sample per language         | chrF, BLEU, COMET where references exist; human adequacy and fluency |
| Pivot-derived pairs      | Selected low-resource and high-resource non-English pairs | Pairs derived from shared English rows                    | Human adequacy; error-propagation categories                         |
| Confidence layer         | Same non-English pairs                                    | Tier 0, Tier 1 and Tier 2 subsets of one pair             | Metric deltas on a fixed human-checked test set                      |
| Correction benefit       | Priority languages and sampled rows                       | Machine layer vs. human-corrected subset                  | Metric deltas and paired human preference                            |
| Robustness               | Domain, length and script buckets                         | Stratified samples using source, length and script labels | Error rates by bucket; qualitative audit                             |
| Responsible-use audit    | All languages, sampled                                    | Released and corrected variants                           | Bias, toxicity, sensitive named-entity and dialect checks            |

Table 3: Recommended evaluation protocol for users of BharatSetu. No results are reported in this paper.

parisons. The first compares models trained with no BharatSetu data against models trained on the released machine layer. The second compares machine-layer training against training on a human-corrected subset. The third compares Tier 0, Tier 1 and Tier 2 subsets of the same language pair (Section 6). The fourth tests whether correcting one language column improves the pivot-derived pairs that use it. Every experiment should record the repository revision, source rows, language columns, preprocessing, tokenizer, model checkpoints, metric versions and random seeds. This record matters because the corpus is synthetic and because a small change in filtering or correction can change how results are interpreted.

Before using the data, we also recommend dataset-integrity checks: each release file should have 984,764 rows and the expected 25 columns, no row should be shuffled across target columns, and shard counts should match `generation/merge_manifest.json`. These checks establish whether the resource is structurally safe to use, not translation quality; Section 8 states what users still cannot infer once they pass.

## 8 Limitations and Ethics

This section states the limits that constrain all claims about BharatSetu. The corpus is machine-translated and not human-verified. Translation errors may include mistranslation, omission, hallucination, named-entity corruption, register mismatch, grammatical errors and script inconsistency. Quality may vary by language, domain and sentence style. These limitations are especially important for lower-resource languages, where model performance may be less stable and where users may

have fewer independent references for checking outputs.

The row-aligned design creates pivot-derived pairs, but it does not create direct human-translated bitext between non-English languages. A Tamil–Kashmiri pair derived from a shared English row may be useful for training or analysis, yet it can inherit errors from both generated columns and from English-centric semantic framing. Researchers should label such data as pivot-derived and should avoid using it as a gold reference for evaluation unless a human correction layer has been added.

The resource may also inherit social bias, dialect imbalance and domain skew from both the English source corpus and the translation model. Scheduled-language labels do not fully capture dialect diversity, register variation or script choice within each language. Responsible use therefore requires evaluation with native speakers, transparent error reporting and care in high-stakes settings such as legal, medical, educational or public-service translation.

## 9 Conclusion

India’s 22 scheduled languages have uneven parallel data, and most language pairs among them have no direct bitext. BharatSetu addresses this by placing 984,764 English sentences and their machine translations into all 22 languages in one row-aligned table. Because every Indic cell shares an English source, any two columns form a pivot-derived pair, which connects any Indic language to any other without direct data. A confidence layer lets users choose how strictly each pair is connected. The corpus is machine-generated and not human-verified, and pivot-derived pairs can inherit

errors from both sides, so we report no quality results. We release BharatSetu as a machine layer and propose a versioned human-correction layer and an evaluation protocol. Each corrected column would upgrade every pair that uses it, so the table can improve step by step into stronger evidence for multilingual Indic machine translation. Future work is to run the protocol, starting with the lower-resource languages.

## References

- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. [Documenting large webtext corpora: A case study on the colossal clean crawled corpus](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305. Association for Computational Linguistics.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. [Understanding back-translation at scale](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500. Association for Computational Linguistics.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. [Beyond english-centric multilingual machine translation](#). *Journal of Machine Learning Research*, 22(107):1–48.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Jay Gala, Pranjal A. Chitale, A. K. Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Jyoti Nawale, Anupama Sujatha, Ratish Pudupully, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. [IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages](#). *arXiv preprint arXiv:2305.16307*.
- Government of India, Department of Official Language. 2026. [Languages included in the eighth schedule of the indian constitution](#). Web page. Accessed 2026-10-11.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Haq Nawaz Malik, Nahfid Nissar, and Faizan Iqbal. 2026. [BharatSetu 22-Indic 1M](#). Hugging Face dataset repository.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, et al. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630:841–846.
- Omnilingual MT Team, Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, et al. 2026. [Omnilingual mt: Machine translation for 1,600 languages](#). *arXiv preprint arXiv:2603.16309*.
- 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. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh Bheemraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. [Samanantar: The largest publicly available parallel corpora collection for 11 indian languages](#). *Transactions of the Association for Computational Linguistics*, 10:145–162.
- Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702. Association for Computational Linguistics.
- Sarvam AI. 2026. [Sarvam-Translate](#). Model card and API documentation. Documentation also available at <https://docs.sarvam.ai/api/getting-started/models/sarvam-translate>; accessed 2026-10-11.

- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. [WikiMatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96. Association for Computational Linguistics.
- Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218. European Language Resources Association.
- Jörg Tiedemann. 2020. [The Tatoeba translation challenge – realistic data sets for low resource and multi-lingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182. Association for Computational Linguistics.