Word-Level Transliteration for English–Indic Code-Mixed Grammar Correction using mT5

Rucha Ambaliya, Mahika Dugar, Dr. Pruthwik Mishra

December 21, 2025

Problem Statement

- Indic languages have rich morphology and multiple scripts.
- Real-world text frequently contains English–Indic code-mixing.
- Grammar correction models are trained on clean, monolingual text.
- Mixed-script and Romanized tokens degrade model performance.
- Limited annotated data makes grammar correction challenging.

Related Work

Grammar Error Correction (GEC) for English:

Early GEC research focused on English using statistical and neural machine translation models. The CoNLL-2014 shared task established standard evaluation using the M^2 scorer, while later work framed GEC as a low-resource translation task. The GLEU metric was introduced for evaluating grammatical corrections based on n-gram overlap.

GEC for Indic Languages:

Research for Indic languages is limited and mainly focuses on transliteration and normalization. Both rule-based and neural transliteration systems have been proposed to handle script variations using character-level and subword representations.

Code-Mixed Language Processing:

Prior work on Hindi-English, Bangla-English, and Tamil-English text primarily targets sentiment analysis, employing lexicon-based methods, Naive Bayes classifiers, and subword-level LSTMs, showing that code-mixing significantly impacts model performance.

Data Provided (IndicGEC 2025)

The IndicGEC 2025 shared task provides grammar correction datasets for five Indic languages in the form of input–output sentence pairs.

Language	Train	Dev	Test
Hindi	599	107	236
Bangla	659	102	330
Malayalam	312	50	102
Tamil	91	16	65
Telugu	603	100	315

Observations:

- Datasets are extremely small.
- Dev and test sets contain very few English tokens.
- Insufficient for training robust multilingual GEC models.

Baseline Evaluation – Dev Set (GLEU, mT5-small)

Language	With Transliteration	Without Transliteration
Hindi	17.74	18.13
Bangla	17.0	17.0
Malayalam	20.05	20.05
Tamil	4.99	4.99
Telugu	12.21	12.21

Baseline Evaluation – Test Set (GLEU, mT5-small)

Language	With Transliteration	Without Transliteration
Hindi	15.62	15.56
Bangla	18.08	18.08
Malayalam	27.07	27.07
Tamil	0.46	0.46
Telugu	12.39	12.16

Issues with Baseline

- GLEU scores are consistently low across languages.
- Tamil performance is extremely poor.
- Transliteration shows negligible impact due to lack of English tokens.
- Models often copy input instead of correcting errors.
- Primary limitation is data scarcity.

Data Augmentation Strategy

To overcome data limitations, we construct an augmented training corpus using IndicCorpV2.

Filtering Criteria:

- Sentences containing only the target language script.
- Sentence length between 5 and 15 words.

Final Dataset:

- 10,000 sentences per language.
- 70% augmented, 30% original sentences.
- Balanced distribution of error types.

Data Augmentation (Examples)

We inject synthetic grammatical errors using character-level and word-level perturbations.

Character-Level Augmentation:

- **Insertion:** घर → घरर
- **Deletion:** रमत → रत
- **Swap:** खाना → नाखा

Word-Level Augmentation:

- **Insertion:** मैं स्कूल गया। → मैं गया स्कूल गया।
- **Deletion:** मैं स्कूल गया। → मैं गया।
- **Swap:** मैं स्कूल गया। → स्कूल मैं गया।

- Each sentence receives one random error.
- Augmented sentence is paired with the original sentence for training.

Training Configuration and Inference Flow

Model Configuration:

- Model: mT5-small
- Learning Rate: 2e-4
- Batch Size: 2
- Epochs: 21
- Max Seq Length: 128
- Gradient Accumulation Value: 4

Training Setup:

- Trained on 10,000 augmented Hindi sentence pairs
- Each pair consists of (synthetic error → correct sentence)
- Objective: learn grammatical correction patterns

Inference Strategy:

- Zero-shot inference on Bangla, Malayalam, Tamil, and Telugu
- No language-specific fine-tuning applied

Inference Flow:

- Transliterate English tokens into target language script
- Apply grammar correction using trained mT5 model
- Generate corrected sentence
- Compute GLEU score against reference

Evaluation After Augmentation – Dev Set (GLEU)

Language	With Transliteration	Without Transliteration
Hindi	83.25	83.25
Bangla	86.94	86.94
Malayalam	89.79	89.79
Tamil	73.07	73.07
Telugu	85.18	85.18

Evaluation After Augmentation – Test Set (GLEU)

Language	With Transliteration	Without Transliteration
Hindi	79.47	78.98
Bangla	81.83	81.83
Malayalam	89.77	89.77
Tamil	84.48	84.48
Telugu	85.03	85.03

Observations

- Data augmentation yields large GLEU improvements.
- Transliteration provides a small gain for Hindi.
- Other languages unaffected due to lack of English tokens.
- Performance gains driven primarily by augmentation.

Future Scope & Conclusion

Future Work:

- Language-specific fine-tuning for all Indic languages.
- Larger models (mT5-base, mT5-large).
- Improved handling of numerals and out of vocabulary transliteration.
- Use of additional evaluation metrics and human evaluation.

Conclusion:

- Data augmentation is critical for Indic grammar correction.
- Word-level transliteration supports code-mixed inputs.
- The proposed pipeline is effective and extensible.