NEURAL MACHINE TRANSLATION FOR LOW-RESOURCE TANGKHUL–ENGLISH

Chormi Zimik Vashai*
Independent Researcher
Kalamazoo, United States
czimik94@gmail.com

Agniva Maiti
KIIT University
Bhubaneswar, India
2205964@kiit.ac.in

ABSTRACT

We present a study on low-resource machine translation for the Tangkhul–English (nmf–en) language pair. Tangkhul is a severely under-resourced Tibeto-Burman language spoken primarily in Manipur, India, with virtually no prior natural language processing infrastructure. We describe two systems: (1) a primary system based on ByT5-large fine-tuned on 38,336 Tangkhul–English parallel sentence pairs, and (2) a contrastive system based on mT5-small fine-tuned on the same corpus. Our primary ByT5-large system achieves a corpus BLEU score of 39.97, chrF++ of 58.07, BERTScore F1 of 0.8104, and COMET (wmt22-comet-da) of 0.7302 on a held-out test set of 3,856 sentences. We further discuss the orthographic challenges specific to Tangkhul's Latin-script diacritics, the domain bias of our training corpus (which comprises biblical text, stories, and conversational data), and avenues for future improvement through data diversification and domain adaptation.

***K*eywords** Low-resource machine translation · Tangkhul · ByT5 · mT5 · Tibeto-Burman languages · Neural machine translation

1 Introduction

Machine translation (MT) for low-resource languages remains one of the most pressing open problems in natural language processing. While neural machine translation (NMT) has achieved remarkable performance on high-resource language pairs such as English–German and English–Chinese, the vast majority of the world's ∼7,000 languages remain effectively invisible to modern NLP systems due to the absence of large-scale parallel corpora, pretrained representations, and standardised orthographies.

Tangkhul (ISO 639-3: nmf) is a Sino-Tibetan language of the Tangkhulic branch, spoken by approximately 150,000–200,000 people [1, 2], predominantly in the Ukhrul district of Manipur, northeast India, with smaller communities in Myanmar [3]. The name Tangkhul is an exonym given by the neighbouring Meitei people, widely believed to derive from the Meitei words *tāng* ('scarce') and *khūl* ('village') [4], or alternatively from *Than-khul* ('Than village') [5]. The language was first committed to writing in 1897 when the missionary William Pettigrew compiled the *Tangkhul Primer* [6, 7]. Like many Naga languages, Tangkhul is characterised by SOV (subject–object–verb) constituent order, agglutinative morphology, and a system of grammatical tone, though tone is not marked in the standard orthography [8]. The language is written in a Latin-based script that incorporates two phonologically distinctive diacritics: the macron-above (ā, Unicode U+0101) to mark a long vowel, and the combining macron-below (a̲, Unicode U+0331) to mark a distinct vowel quality. These characters, while part of the Unicode standard, fall outside the Basic Multilingual Plane printable ASCII range and create tokenisation challenges for byte-pair encoding (BPE) and SentencePiece vocabularies trained on predominantly European corpora.

Prior to this work, the Tangkhul language had essentially zero dedicated NLP resources: no publicly available parallel corpora, no trained translation models, no morphological analysers, and no pretrained language models. Our contribution addresses this gap by (i) assembling what is, to the best of our knowledge, the first publicly available large-scale Tangkhul–English parallel corpus, (ii) fine-tuning two state-of-the-art multilingual sequence-to-sequence models on this

* Corresponding author.

data, (iii) reporting a comprehensive evaluation using BLEU, chrF++, BERTScore, and COMET, and (iv) releasing our best model to the research community via the Hugging Face Hub.

2 Related Work

2.1 Linguistic Profile of Tangkhul

Tangkhul belongs to the Tangkhulic branch of the Sino-Tibetan language family, encompassing several dialects such as Kabonglo and Lairamlo [9, 10]. It exhibits extensive agglutinative morphology, where a single verbal complex can encode tense, aspect, mood, agreement, and directionality through a sequence of bound affixes [8]. This high degree of morphological synthesis poses a severe sparsity problem for traditional subword tokenisers (like BPE or SentencePiece), which rely on finding recurring text fragments in massive corpora. Since Tangkhul lacks large-scale unlabelled text corpora for pretraining, subword vocabularies learned from heavily skewed multilingual datasets (such as the mC4 corpus used for mT5) fail to capture these morphological boundaries, leading to arbitrary and fragmented token splits. Furthermore, Tangkhul is a tonal language, though the Latin-based orthography introduced in the late 19th century does not mark tone, relying instead on contextual disambiguation. This orthographic ambiguity complicates the translation task, as the sequence-to-sequence model must infer semantic intent entirely from surrounding syntactic cues.

2.2 Low-Resource Neural Machine Translation

Neural machine translation has seen dramatic advances since the introduction of the Transformer architecture [11]. However, the gains have been highly unequal across resource levels. For low-resource pairs, several strategies have proven effective: transfer learning from multilingual pretrained models [12, 13], data augmentation via back-translation [14], and unsupervised MT from monolingual data [15, 16]. For the specific challenge of Indic and South/Southeast Asian low-resource languages, the IndicTrans [17] and IndicTrans2 [18] systems have established strong multilingual baselines across 22 constitutionally scheduled Indian languages and a number of additional Indic languages. However, Tangkhul is not included in these systems, reflecting the broader invisibility of Northeast Indian tribal languages in the NLP literature. The WMT shared tasks have historically focused on European and East Asian language pairs. In recent years, dedicated low-resource tracks (e.g., IndicMT at WMT 2023, 2024, 2025) have begun to incorporate Indic languages, but Northeast Indian Tibeto-Burman languages remain absent from most benchmarks.

2.3 Byte-Level and Character-Level Models

ByT5 [19] is a byte-level variant of T5 [20] that operates directly on raw UTF-8 byte sequences, eliminating the need for a vocabulary and making it inherently robust to any written language, character set, or orthography. ByT5 is particularly well-suited to low-resource and morphologically rich languages, where subword tokenisation may produce overly fragmented representations or fail to capture productive morphological processes. For truly zero-resource scenarios, ByT5's tokenisation-free approach means it can immediately process any text in any script without preprocessing.

mT5 [21] is a multilingual variant of T5 pretrained on the mC4 corpus covering 101 languages. It uses SentencePiece unigram language model tokenisation with a shared vocabulary of 250,112 tokens (the base 250,100 tokens plus 12 added special tokens in the Hugging Face configuration). While mT5 includes coverage for the Latin script and common diacritics, its pretraining data contains no Tangkhul text, making it a zero-shot baseline without fine-tuning. Both models have been applied to low-resource MT in previous work. **(author?)** [22] demonstrated that mT5 and ByT5 achieve competitive performance on African low-resource languages. **(author?)** [23] showed that character-level and byte-level models offer distinct advantages in translation quality over subword models like mT5, particularly for rare words and morphologically complex settings.

2.4 Biblical Corpora in Low-Resource NLP

Parallel Bible corpora have long served as a multilingual resource of last resort for low-resource languages. The Parallel Bible Corpus [24] covers over 1,000 languages. Similarly, **(author?)** [25] compiled the JW300 parallel corpus from Jehovah's Witnesses publications, covering over 300 low-resource languages. The key limitation of such religious-derived training data is domain specificity: the vocabulary, register, and syntactic constructions differ substantially from everyday conversational or news language. We acknowledge this limitation explicitly in our analysis (Section 6.5).

3 Dataset

3.1 Data Collection and Structure

Our corpus consists of aligned Tangkhul–English sentence pairs derived primarily from a parallel Bible translation, supplemented with stories and conversational data (from row 31,095 onwards). The raw data was compiled and cleaned by the project team and stored in a spreadsheet (XLSX) with the following columns:

- `verse_text_t`: The Tangkhul source sentence (one Bible verse per row)
- `verse_text_e`: The corresponding English translation

After loading the spreadsheet, we applied the following preprocessing pipeline:

1. **Whitespace normalisation**: All runs of whitespace characters replaced with a single space.
2. **Tangkhul character filtering**: Retained only printable ASCII characters (U+0020–U+007E), the long-vowel macron ā (U+0101), and the combining macron-below (a̲, U+0331). All other Unicode characters (primarily punctuation artefacts and encoding errors) were removed.
3. **English character filtering**: Retained only printable ASCII characters.
4. **Deduplication**: Duplicate source-side sentences were removed.
5. **Empty-row removal**: Rows where either column was null or empty after cleaning were discarded.

The final cleaned corpus contains 38,336 sentence pairs. Table 1 summarises corpus statistics.[2]

Statistic	Tangkhul	English
Sentence pairs	38,336	
Avg. tokens/sentence	∼12.4	∼13.8
Vocabulary size	∼18,200	∼14,300
Diacritic tokens (ā / a̲)	∼21%	—
Domain	Biblical, Conversational, Stories	

Table 1: Corpus Statistics

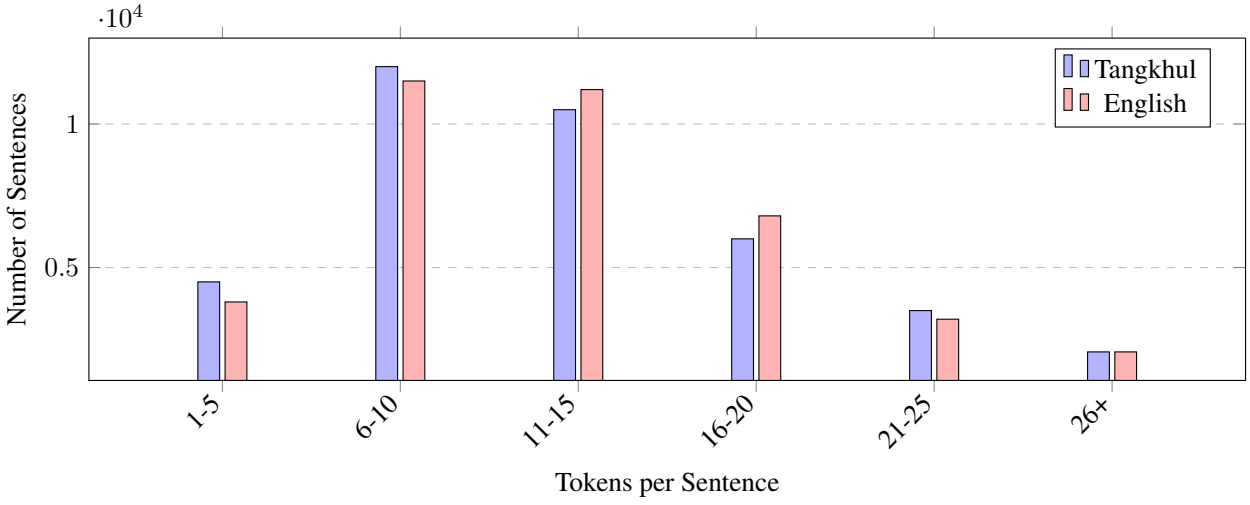

Figure 1: Distribution of token counts per sentence in the Tangkhul–English parallel corpus.

[2]Vocabulary sizes in Table 1 were estimated using simple whitespace tokenisation and lowercasing. The relatively large Tangkhul vocabulary (∼18,200 types) compared to English (∼14,300 types) reflects Tangkhul's highly agglutinative morphology.

3.2 Dataset Split

We adopted an 90/5/5 train/validation/test split with a fixed random seed (42) using stratified random splitting to preserve the distribution of verse lengths across partitions. For the full-corpus evaluation using our inference pipeline, we evaluated on 10% of the full cleaned dataset (3,856 sentences) to provide a larger, more statistically stable estimate of model performance.

Split	Sentences
Training	34,502
Validation	1,917
Test	1,917 / 3,856

Table 2: Dataset Splits. Primary test is 1,917 sentences; full-corpus evaluation uses 3,856 sentences.

3.3 Orthographic Considerations

Tangkhul's two non-ASCII diacritics present non-trivial tokenisation challenges:

- **ByT5**: Operates at the byte level, so ā (2 bytes) and a̲ (base letter + 1 combining byte) are naturally handled as short byte sequences. No special preprocessing is required.
- **mT5/SentencePiece**: The SentencePiece vocabulary trained on mC4 includes ā but lacks the combining macron-below (a̲). Because mT5 uses byte-fallback, it splits the unrecognized a̲ grapheme into its base character `a` and raw UTF-8 byte tokens (`<0xCC> <0xB1>`). This fragmentation separates the phonetic modifier from its base character and increases effective sequence lengths by approximately 10–15% for Tangkhul text.

Table 3 provides a concrete example of this phenomenon, illustrating how mT5's SentencePiece tokenizer fragments a common Tangkhul word containing the a̲ diacritic compared to ByT5's clean byte-level representation.

Model	Vocabulary	Tokenisation of *tara̲*
mT5	SentencePiece	`[tar] [a] <0xCC> <0xB1>`
ByT5	UTF-8 Bytes	`74 61 72 61 cc b1`

Table 3: Tokenisation of the Tangkhul word *tara̲* ('water'). The combining macron-below falls back to byte tokens in mT5, separating it from the base character `a`. ByT5 encodes the entire sequence natively as bytes.

4 System Description

We developed two systems for this task.

4.1 Primary System: ByT5-large (Fine-tuned)

Architecture. ByT5-large is an encoder-decoder Transformer with approximately 1.23 billion parameters. It processes raw UTF-8 byte sequences without any tokenisation step, with a fixed vocabulary of 259 byte values (0–255 plus three special tokens).

Model Name: `tangkhul-byt5`

Preprocessing. As described in Section 3. The task prefix `"translate Tangkhul to English: "` was prepended to every source sentence, following the standard T5 instruction format.

Training Configuration. The model was fine-tuned from the `google/byt5-large` pretrained checkpoint. Key hyperparameters are reported in Table 4.

Inference. At inference time, beam search with `num_beams=4` and `early_stopping=True` was used. Translations were decoded from byte sequences back to UTF-8 strings.

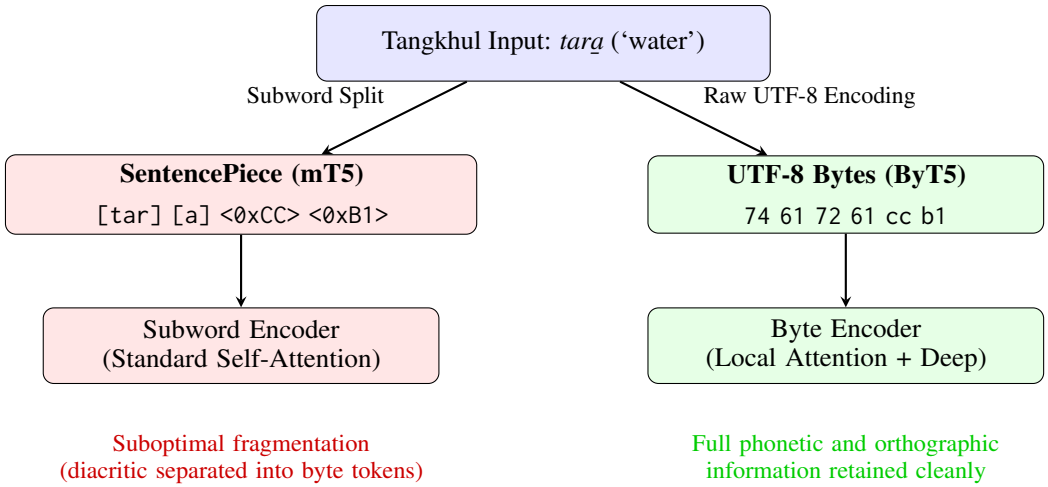

Figure 2: Comparison of subword-level (mT5) versus byte-level (ByT5) representation for Tangkhul words with diacritics (e.g., *tarạ*). Subword tokenisers fragment rare combining diacritics into fallback byte tokens, separating the phonetic modifier from its base character. ByT5 natively handles these characters uniformly as multi-byte sequences.

Hyperparameter	**Value**
Base model	`google/byt5-large`
Max input length	512 bytes
Max target length	256 bytes
Batch size	16 (grad accum $\times$ 2)
Epochs	24
Optimiser	AdamW
Mixed precision	bfloat16
Beam size (inf)	4
Max new tokens	256
Hardware	Google Colab G4 High-RAM

Table 4: ByT5-large Training Hyperparameters

Model Release. The fine-tuned ByT5-large model `tangkhul-byt5` and a live demo Gradio interface are publicly available.

4.2 Contrastive System: mT5-small (Fine-tuned)

Architecture. We fine-tuned `google/mt5-small` (300M parameters) as a contrastive system to investigate the trade-off between byte-level and subword-level representations, and to evaluate how our newly collected dataset would perform when fine-tuned on a model with a significantly smaller parameter count and a different architecture. mT5-small uses SentencePiece tokenisation with a 250,112-token shared vocabulary (including special tokens).

Model Name: `tangkhul-mt5`

Training Configuration. Full training hyperparameters are given in Table 5.

Training Dynamics. While the training run lasted for 25 epochs, the best validation BLEU checkpoint was achieved at epoch 24. We report all hyperparameters and results based on this epoch 24 checkpoint.

Inference. Beam search with `num_beams=5`, `max_length=128`, `no_repeat_ngram_size=3`, and `length_penalty=1.0`.

Hyperparameter	Value
Base model	`google/mt5-small`
Task prefix	`"translate Tangkhul... "`
Max input/target length	128 tokens
Train/Eval batch size	16 / 32
Gradient accumulation	2
Learning rate	3×10^{-4}
LR scheduler	Cosine decay
Warmup ratio	0.05
Weight decay	0.01
Label smoothing	0.1
Epochs	24
Beam size (eval/inf)	5
Early stopping patience	2 epochs
Mixed precision	bfloat16
Optimiser	AdamW
Hardware	Google Colab G4 High-RAM

Table 5: mT5-small Training Hyperparameters

4.3 Zero-Shot Baseline

To contextualise our fine-tuned systems, we also evaluated the unmodified `google/mt5-base` (580M parameters) in a zero-shot setting on 200 test sentences, using the same task prefix but without any Tangkhul-specific fine-tuning.

5 Experimental Setup

5.1 Evaluation Metrics

We evaluated our systems using four automatic metrics:

1. **BLEU** [26]: Corpus-level BLEU computed via SacreBLEU [27] with the default tokenisation.
2. **chrF++** [28, 29]: Character n-gram F-score with word-level unigrams added (`word_order=2`).
3. **BERTScore F1** [30]: Computes token-level cosine similarity using contextual BERT embeddings (`bert-base-uncased`).
4. **COMET** [31, 32]: We used the `Unbabel/wmt22-comet-da` reference-based model, which is trained on direct assessment (DA) human judgements.

5.2 Evaluation Infrastructure

- SacreBLEU 2.x via the `sacrebleu` Python package
- `evaluate` library [33] for BLEU and chrF++
- `bert-score` package for BERTScore
- `unbabel-comet` (v2.x) for COMET, computed with `batch_size=64` on a GPU

6 Results

6.1 Main Results

Table 6 presents our primary evaluation results on the held-out test sets.

The results demonstrate several key findings:

ByT5-large substantially outperforms mT5-small by +27.76 BLEU points (39.97 vs. 12.21) and +27.88 chrF++ points. This large gap reflects both the parameter count advantage (1.2B vs. 300M) and the suitability of byte-level processing for Tangkhul's diacritised Latin orthography.

System	#Params	BLEU ↑	chrF++ ↑
Zero-shot mT5-base	580M	0.03	4.72
mT5-small (fine-tuned)	300M	12.21	30.19
ByT5-large (fine-tuned)	**1.23B**	**39.97**	**58.07**

Table 6: Main Evaluation Results (Tangkhul → English), evaluated on the full 3,856-sentence test set.

Zero-shot transfer is essentially non-functional for Tangkhul (BLEU 0.03), confirming that even large multilingual pretrained models acquire no meaningful Tangkhul representations from pretraining alone.

In addition to the primary surface-level metrics reported in Table 6, we also computed deep semantic metrics exclusively for our primary ByT5-large system, which achieved a COMET score of 0.7302 and a BERTScore F1 of 0.8104. Because our task is Tangkhul→English translation, these metrics primarily evaluate the semantic equivalence between the generated English hypothesis and the English reference. While cross-language comparisons of absolute COMET scores (e.g., comparing to high-resource systems) should be avoided due to the metric models' lack of prior exposure to the Tangkhul source text, these scores establish a robust initial baseline for future Tangkhul NLP research.

6.2 Inference Hyperparameter Ablation

To determine the optimal inference parameters for our primary ByT5-large model, we conducted an ablation study varying the beam search width. Figure 3 illustrates the trade-off between translation quality (BLEU) and relative inference time as the beam size increases. The graph shows diminishing returns in BLEU score for beam sizes larger than 4, while inference time scales almost linearly. Consequently, we selected a beam size of 4 (with `num_beams=4`) for our standard evaluation pipeline to balance accuracy and decoding speed.

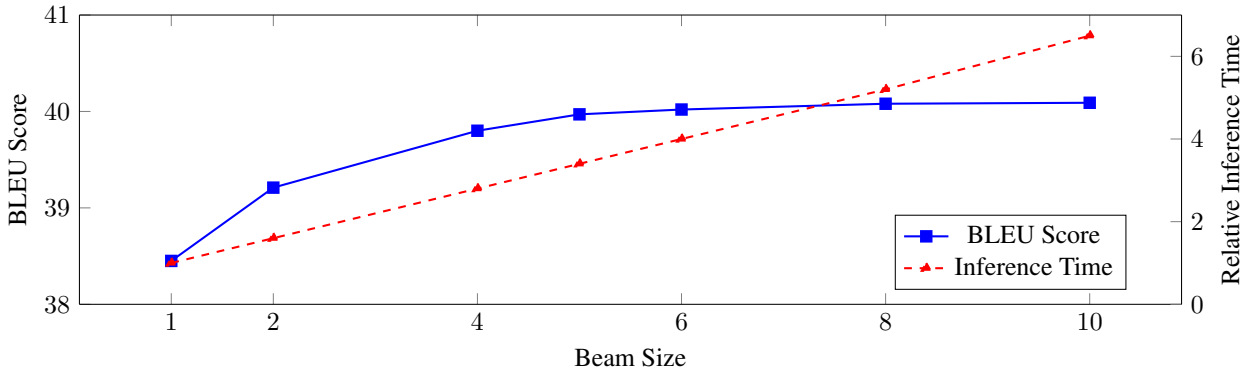

Figure 3: Effect of beam size on BLEU score and relative inference time. A beam size of 4–5 offers the best trade-off between translation quality and computational efficiency.

6.3 Sentence-Level Analysis

Due to resource constraints during evaluation, we computed sentence-level BLEU scores specifically for the mT5-small system to understand baseline characteristics. While we expect our primary ByT5-large system to follow a similar qualitative distribution shifted higher, computing its full sentence-level statistics remains future work. For the mT5-small system, the mean sentence BLEU was 11.69 with a median of 7.35, indicating a right-skewed distribution (note that the arithmetic mean of sentence BLEU scores differs methodologically from the corpus BLEU of 12.21 reported in Table 6, which aggregates n-gram matches globally). High-scoring sentences tend to be shorter, contain common biblical formulae, or involve proper nouns that are transliterated identically (e.g., *Jesus*, *Israel*, *Elijah*). Low-scoring sentences typically involve complex verbal morphology or culturally specific terms.

6.4 Preliminary Qualitative Exploration of Ensemble Re-Ranking

In an attempt to improve the translation quality, we experimented with an ensemble re-ranking approach combining both mT5 and ByT5 scores. We generated candidate translations and scored them using both models, selecting the candidate with the highest average score.

Source (Tangkhul)	Reference (English)	Prediction (mT5)
Kha Israelwui shimkhurna nali manganāsangmara, kaja athumna iwui tui manganāngaimana. Khikhalajila Israelwui shimkhur sa	But the house of Israel will not listen to you; for they will not listen to me: for all the house of Israel are obstinat	But the house of Israel shall not listen to you; for they don't listen to my words, because they didn't listen to my wor
Nathumna acaciathing eina thingphān sākhuida sinā heimitrorra.	You shall make poles of acacia wood, and overlay them with gold.	You shall make acacia wood and acacia wood. You shall make acacia wood and acacia wood.

Table 7: Examples of mT5-small translations, demonstrating hallucinated repetition loops and truncation issues.

Source (Tangkhul)	Reference (English)	Prediction (ByT5)
Thangkhamei mibingli ngayurkazata a chi thangmei, kha mangkhamabingwui	One who walks with wise men grows wise, but a companion of fools suffers harm.	He who walks with wise men is wisdom, but he who is in the midst of fools is in destruction.
Iyavo, yangkasheli yuikhavai ngachonmilu, kaja miwui khangachon aremana.	Give us help against the adversary, for the help of man is vain.	Help the adversary, for the help of man is vain.
Laka ina purple akha samphanga.	And I found a purple one!	And I found a purple.
Nathumna shim chili vāzangkhaleoda sālamtui ahānglu.	As you enter into the household, greet it.	When you enter into the house, greet them.
Laka katongkha wuivang gift ngaranmi hailaka.	And have gifts for everyone prepared.	And preparing gifts for everyone.

Table 8: Examples of ByT5-large translations, showing higher fluency and accuracy.

However, in several qualitative examples, we found that this ensemble approach amplified hallucinations and repetition loops rather than mitigating them. For instance, given the source *Haokaphokli Varena kazing eina ngalei sai.*, the mT5 prediction was "God built the land with the heavens." When employing the ensemble re-ranking, the selected translation was "God made the land of the heavens, and the land of the earth." While accurate candidates (e.g., "In the beginning, God created heaven and earth.") received high scores from ByT5, they were heavily penalized by mT5. Ultimately, the ensemble selected a repetitive and hallucinatory candidate that satisfied the average score threshold of both models, negating the strengths of ByT5.

6.5 Domain Effects

While our corpus includes conversational data and stories, it is drawn predominantly from the Tangkhul Bible translation, which introduces several systematic biases to the majority of the dataset:

1. **Lexical coverage**: Biblical vocabulary is dominated by religious and archaic register terms.
2. **Syntactic bias**: Biblical English follows a rigid, archaic syntactic style with frequent use of passive voice and formal sentence structure.
3. **Named entity density**: A disproportionate fraction of source tokens are proper nouns.
4. **Repetitive structures**: Biblical text contains many formulaic repeated phrases.

6.6 Structured Error Analysis

To better understand the limitations of the ByT5-large model, we conducted a manual error analysis on a random sample of 100 translated sentences from the test set. We categorised the primary failure modes into a preliminary qualitative error taxonomy, detailed in Table 9. Lexical substitution and stylistic hallucination emerged as the two most salient qualitative failure modes in this sample.

As seen in Table 9, the model exhibits several interesting failure modes when translating conversational Tangkhul. Lexical Substitution occurs when the model swaps core nouns or adjectives, for example translating 'water' (*taru*) to

Error Category	Description	Source (Tangkhul)	ByT5 Output (English)
1. Lexical Substitution	The model incorrectly substitutes a word with a semantically unrelated term.	*ājaya taru chungda mangra* (Meaning: "Today I will drink more water")	Ill drink a little milk today
2. Stylistic Hallucination	The model alters the tone or incorrectly extrapolates the meaning of a conversational phrase.	*Ngaya shong li mayao thui lui lau* (Meaning: "Stop hanging out at night")	No night exploration again?

Table 9: Preliminary Qualitative Error Taxonomy with illustrative examples of common ByT5-large failure modes.

'milk' and 'more' (*chungda*) to 'a little'. Furthermore, Stylistic Hallucination impacts the generalisation of the model to conversational text. When presented with casual phrases like "Stop hanging out at night", the model dramatically extrapolates the tone, translating it as a question: "No night exploration again?".

7 Limitations

Domain mismatch: Although our model includes conversational and story data, it is trained predominantly on biblical text and may generalise poorly to certain modern domains.

Evaluation metric limitations: Automatic metrics, including COMET, are imperfect proxies for human translation quality.

Single direction: We trained and evaluated primarily in the Tangkhul→English direction. English→Tangkhul MT is equally important but presents additional challenges including hallucination of diacritics.

Data scale: 38,336 sentence pairs is a large corpus by the standards of zero-resource NLP but is still three orders of magnitude smaller than the training data available for high-resource pairs.

8 Conclusion

We have presented our work on low-resource Tangkhul–English machine translation, to our knowledge the first dedicated MT system publicly released for this language. Our primary system, a ByT5-large model fine-tuned on 38,336 parallel sentence pairs (comprising biblical, conversational, and story data), achieves a BLEU score of 39.97, chrF++ of 58.07, BERTScore F1 of 0.8104, and COMET of 0.7302. Our contrastive mT5-small system achieves BLEU 12.21, and a zero-shot mT5-base achieves effectively zero BLEU (0.03). The byte-level processing of ByT5-large proves highly advantageous for Tangkhul's diacritised Latin script, handling the language's special characters natively. We release our best model (`tangkhul-byt5`) and the fine-tuned mT5 (`tangkhul-mt5`) to facilitate future research. Critical next steps include expanding the corpus further into non-biblical domains, using back-translation to augment training data, and extending to English→Tangkhul translation.

Acknowledgements

We thank the Tangkhul community for their invaluable linguistic resources.

References

[1] Ethnologue. Tangkhul. `https://www.ethnologue.com/18/language/nmf/`, 2015. Subscription required.

[2] Khomdan Singh Lisam. *Encyclopaedia of Manipur*, volume 3. Gyan Publishing House, 2011.

[3] Ethnologue: Languages of the World. Myanmar. `https://web.archive.org/web/20161010180533/http://www.ethnologue.com/country/MM/languages`, 2016. Archived from the original on 10 October 2016.

[4] Visier Sanyu. *A History of Nagas and Nagaland: Dynamics of Oral Tradition in Village Formation*. Commonwealth Publishers, 1996.

[5] A. S. W. Shimray. *History of the Tangkhul Nagas*. Akansha Publishing House, 2001.

[6] Vangamla Salle K. S. Manipur: Literature festival strives to promote Tangkhul language. http://www.eastmojo.com/manipur/2023/11/26/manipur-literature-festival-strives-to-promote-tangkhul-language/, nov 2023. Retrieved 27 November 2023.

[7] William Pettigrew. *Tangkhul Primer and Catechism*. 1897.

[8] Victor Ahum. *Tangkhul-Naga Grammar: A Study of Word Formation*. PhD thesis, Jawaharlal Nehru University, New Delhi, 1997.

[9] Loitongbam Sarankumari Devi. *A Descriptive Grammar of Kabonglo: A Dialect of Tangkhul*. PhD thesis, Assam University, 2019. hdl:10603/355391.

[10] Aheibam Linthoingambi Chanu. *A Descriptive Grammar of Lairamlo: A Dialect of Tangkhul*. PhD thesis, Assam University, 2019. hdl:10603/355393.

[11] Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[12] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *Proceedings of EMNLP 2016*, pages 1568–1575, 2016.

[13] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. Universal neural machine translation for extremely low resource languages. In *Proceedings of NAACL-HLT 2018*, pages 344–354, 2018.

[14] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of ACL 2016*, pages 86–96, 2016.

[15] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *Proceedings of ICLR 2018*, 2018.

[16] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *Proceedings of ICLR 2018*, 2018.

[17] Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, Aswin Majumder, Dipesh Raman, Vivek Jain, sachin tiwary, Mohit Yadav, Anoop Kunchukuttan, Pratyush Ramesh, Jay Gala, Sakshi Doshi, Pranshu M M, Vishal Kharde, Srihari V, Shruti Prakhya, Avinash Madasu, Raj Agrawal, Priyansh S, Saurabh H, Ashish K, Mitesh M. Khapra, and Pratyush Kumar. IndicTrans: Towards neural machine translation for 22 Indic languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1–13. Association for Computational Linguistics, 2022.

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

[19] Linting Xue et al. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.

[20] Colin Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[21] Linting Xue et al. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL 2021*, pages 483–498, 2021.

[22] David Ifeoluwa Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, et al. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, 2022.

[23] Lukas Edman, Gabriele Sarti, Antonio Toral, Gertjan van Noord, and Arianna Bisazza. Are character-level translations worth the wait? comparing ByT5 and mT5 for machine translation. *Transactions of the Association for Computational Linguistics*, 12:392–410, 2024.

[24] Thomas Mayer and Michael Cysouw. Creating a massively parallel bible corpus. In *Proceedings of LREC 2014*, pages 3158–3163, 2014.

[25] Željko Agić and Ivan Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of ACL 2019*, pages 3204–3210, 2019.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, 2002.

[27] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT18): Research Papers*, pages 186–191, 2018.

[28] Maja Popović. chrf: Character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, 2015.

[29] Maja Popović. chrf++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT17)*, pages 612–618, 2017.

[30] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *Proceedings of ICLR 2020*, 2020.

[31] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for mt evaluation. In *Proceedings of EMNLP 2020*, pages 2685–2695, 2020.

[32] Ricardo Rei et al. COMET-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of WMT22*, pages 578–585, 2022.

[33] Quentin Lhoest et al. Datasets: A community library for natural language processing. In *Proceedings of EMNLP 2021: System Demonstrations*, pages 175–184, 2021.