Not All Data Augmentation Works: A Typology-Aware Study for Low-Resource Neural Machine Translation in Vietnamese Ethnic Minority Languages

Long S. T. Nguyen[1,3], **Dat T. Truong**[1,3], **Nhan D. Tran**[1,3], **Quynh T. N. Vo**[1,3], **Quy T. Nguyen**[2,3], **Tho T. Quan**[1,3*]

[1]URA Research Group, Ho Chi Minh City University of Technology (HCMUT), Vietnam
[2]University of Social Sciences and Humanities (USSH), Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
{long.nguyencse2023, qttho}@hcmut.edu.vn

Abstract

Neural Machine Translation (NMT) for low-resource and underserved languages remains challenging due to the severe lack of parallel corpora, linguistic tools, and evaluation resources. The issue is evident in Vietnam, where the ethnolinguistic minority languages Tày (Tai–Kadai) and Bahnar (Austroasiatic) hold cultural significance but remain digitally under-represented. Data Augmentation (DA) offers a cost-effective remedy; however, most existing techniques were designed for high-resource analytic languages and are often applied heuristically without assessing their linguistic compatibility. In this work, we present the first systematic study of DA for two minority language pairs, Tày–Vietnamese and Bahnar–Vietnamese, within a three-stage language model pipeline consisting of Vietnamese-based initialization, monolingual adaptation, and supervised fine-tuning. We train two independent encoder–decoder NMT systems to isolate augmentation effects and analyze how linguistic typology shapes augmentation behavior. Our experiments show that meaning-preserving DA methods consistently improve translation adequacy and linguistic faithfulness, whereas several widely used techniques introduce semantic or structural degradation. Through quantitative evaluation and typology-aware linguistic analysis, we derive practical guidelines for selecting DA strategies in extremely low-resource and typologically diverse settings. We additionally release newly digitized high-quality bilingual corpora and trained models to facilitate future research and community-centered NLP development.

Data and Models — https://osf.io/fmq5k/

1 Introduction

Developing reliable *Neural Machine Translation* (NMT) systems for low-resource and underserved languages remains one of the most enduring challenges in *Natural Language Processing* (NLP) (Ranathunga et al. 2023; Her and Kruschwitz 2024). Although multilingual pretraining and transfer learning have brought noticeable improvements for medium-resource languages, many ethnolinguistic minority languages still lack essential resources such as parallel corpora, linguistic tools, and evaluation benchmarks that are necessary for stable NMT or *Language Model* (LM) development (Raja and Vats 2025; Nguyen et al. 2025). This situation is clearly observed in Vietnam, where Tày, a Tai–Kadai language (Holm 2020) spoken by approximately 1.92% of the national population, and Bahnar, an Austroasiatic language (Alves 2019) spoken by about 0.30%, retain high cultural and linguistic value but remain severely under-resourced with limited digitized corpora and almost no community-accessible NLP technologies. As a result, speakers of these languages face inequitable access to digital information, educational materials, and *Artificial Intelligence* (AI)-enabled language services, which positions them as representative underserved communities in modern AI development (Shi et al. 2022).

Transformer architectures and recent advances in LM research have significantly improved translation performance in well-resourced language settings (Khoboko, Marivate, and Sefara 2025). However, even strong encoder–decoder LM-based NMT systems degrade substantially when bilingual supervision is extremely limited because the model lacks sufficient lexical coverage, exposure to morphosyntactic variation, and training signals that support linguistic generalization (Nguyen et al. 2025). Although expanding parallel corpora through field collection would be ideal, this process is costly, time consuming, and dependent on linguistic expertise as well as long-term community engagement. These limitations highlight the importance of *Data Augmentation* (DA) as a practical, low-cost, and scalable strategy for alleviating data scarcity (Li, Hou, and Che 2022; Fadaee, Bisazza, and Monz 2017; Xia et al. 2019). However, most existing DA techniques were developed for high-resource analytic languages and are often reused in low-resource contexts as plug-and-play heuristics without verifying their linguistic compatibility. When applied to typologically different languages, DA operations may introduce various forms of distortion including semantic drift, syntactic violations, morphological corruption, or discourse incoherence. These risks are especially harmful when each training instance carries high informational value. Therefore, the central research question is not whether DA improves low-resource NMT, but which DA techniques are beneficial, under which linguistic conditions, and why.

To investigate this question, we evaluate DA within a three-stage LM development pipeline designed for underserved languages. The pipeline includes: (i) initialization using a Vietnamese pretrained model to take advantage of

typological proximity, (ii) monolingual adaptation on Tày and Bahnar corpora to ground language-specific representation, and (iii) *Supervised Fine-Tuning* (SFT) on parallel data where DA strategies are systematically introduced and analyzed. In order to isolate augmentation effects and avoid cross-language interference, we build and evaluate two independent NMT systems, one for Tày–Vietnamese and one for Bahnar–Vietnamese, rather than adopting a shared multilingual model. This approach treats DA not simply as a method for increasing data volume, but as a controlled alignment mechanism, where both meaning-preserving and structure-altering transformations are systematically examined.

To the best of our knowledge, no previous work has systematically examined DA for Vietnamese ethnic minority languages or analyzed how linguistic typology influences augmentation effectiveness. Tày exhibits an analytic structure with strict *Subject–Verb–Object* (SVO) word order, obligatory classifiers, and clause-final particles. In contrast, Bahnar exhibits agglutinative morphology, productive prefixation, and verbal alternation patterns that encode valency and aspect. These contrasts provide a natural setting for investigating how structural characteristics shape the usefulness and potential drawbacks of DA.

Our contributions are summarized as follows.

- We conduct the first systematic comparison of multiple DA techniques for two Vietnamese minority language pairs: Tày–Vietnamese and Bahnar–Vietnamese.
- We provide the first typology-aware linguistic analysis explaining why certain DA methods preserve meaning while others introduce semantic drift or structural distortions.
- We demonstrate that meaning-preserving augmentation consistently improves translation performance over strong encoder–decoder baselines and establish the first dedicated NMT benchmarks for both Tày→Vietnamese and Bahnar→Vietnamese translation directions.
- We propose a linguistically informed guideline for selecting DA strategies in extremely low-resource and typologically diverse conditions.
- We release newly digitized bilingual corpora and reproducible NMT models to support future research and community-centered NLP development.

Overall, our findings show that DA is not universally beneficial and may harm translation performance if applied without typological awareness. We emphasize the importance of linguistically informed, meaning-preserving, and community-aligned data-centric training decisions in the development of equitable and sustainable NLP technologies for underserved populations.

2 Related Works

2.1 NMT for Low-resource Languages

While NMT has achieved strong performance in high-resource settings, its dependence on large-scale parallel corpora poses significant challenges for low-resource languages. Prior studies address these limitations through three main directions (Wang et al. 2021). The first leverages monolingual data using back-translation, joint training, unsupervised learning, and LM pretraining (Sennrich, Haddow, and Birch 2016; He et al. 2016; Lample et al. 2018; Nguyen et al. 2025); however, these methods often underperform on typologically distant languages where alignment signals are weak and lexical transfer is limited. The second direction exploits auxiliary languages via multilingual training, transfer learning, or pivot translation (Johnson et al. 2017; Hujon, Singh, and Amitab 2023; Elmadani and Buys 2024), yet performance is highly sensitive to language relatedness, data balancing, and error propagation across stages. A third line of work incorporates multimodal supervision from images or speech (Chen, Jin, and Fu 2019; Zhang et al. 2021), but such approaches require carefully aligned multimodal datasets and introduce additional modeling complexity. Despite promising progress, most existing methods still assume the availability of sizable, clean monolingual or bilingual corpora, which is unrealistic for extremely low-resource languages such as Tày and Bahnar.

2.2 Data Augmentation for Low-resource NMT

DA has emerged as a practical strategy for mitigating parallel data scarcity, with existing techniques broadly categorized into paraphrasing-based, noising-based, and sampling-based approaches (Li, Hou, and Che 2022). Paraphrasing methods generate lexical or phrasal variants via synonym substitution or embedding-driven rewriting (Miller 1994; Wei and Zou 2019; Wang and Yang 2015), but their effectiveness depends heavily on lexical knowledge resources largely unavailable for minority languages. Noising-based methods, including random swapping, insertion, deletion, or sentence reordering (Wei and Zou 2019; Yan et al. 2019), are simple and scalable but risk degrading syntactic well-formedness and sentence meaning, particularly for structurally rigid languages. Sampling-based approaches such as syntactic transformations, rarity-aware sampling, and bilingual pseudo-data generation (Min et al. 2020; Fadaee, Bisazza, and Monz 2017; Zhang, Ge, and Sun 2020; Xia et al. 2019) are more principled but rely on linguistic metadata and lexicons often unavailable in extreme low-resource settings. Back-translation (Sennrich, Haddow, and Birch 2016) is widely regarded as an effective DA method and recently inspired LM-based synthetic data generation (Fabbri et al. 2021; Mai and Luong 2023), yet these approaches presuppose pretrained translation or foundation models, which do not currently exist for Tày and Bahnar. Importantly, DA research has focused on surface-level diversity rather than evaluating semantic fidelity, syntactic validity, or morphological integrity, a limitation critical for typologically sensitive languages where minor perturbations may alter valency, argument roles, or aspectual meaning. To date, no work has examined the meaning-preservation behavior or typological suitability of DA for Vietnamese minority languages, leaving it unclear which strategies are beneficial, harmful, or conditionally applicable.

3 Selected Data Augmentation Methods

We select and redesign a set of lightweight DA techniques that do not rely on pretrained NMT models, morphologi-

cal analyzers, or external linguistic tools. Unlike most prior work that augments only the source side, we apply all DA methods to both the source (x_i) and target (y_i) sentences. A shared bilingual Tày–Vietnamese and Bahnar–Vietnamese dictionary ensures that all token-level transformations remain lexically aligned across the two sides.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the original bilingual dataset of N aligned sentence pairs. For each source sentence x_i, let $L_i = |x_i|$ be its tokenized length, and define the corpus-level mean length as

$$\bar{L} = \frac{1}{N} \sum_{i=1}^{N} L_i.$$

We write $t_{i,j}$ for the j-th token of x_i ($1 \leqslant j \leqslant L_i$), and use x_i' to denote any augmented variant of x_i. For each augmentation method, let S be the total number of generated samples per side, and let $\bar{L}'$ be the mean length of these augmented sentences. When a method is applied only to a subset of the data, we denote its index set by $\Omega \subseteq \{1, \ldots, N\}$.

3.1 Combining (Thematic Concatenation)

This method enriches contextual content by linearly concatenating multiple sentences sharing the same thematic label. The resulting augmented samples are not strictly meaning-preserving at the sentence level, but topical consistency is maintained.

Let $\mathcal{S} = \{S_1, S_2, \ldots, S_M\}$ be the set of monolingual sentences and $\mathcal{T}$ the set of thematic labels. Each sentence is assigned exactly one theme via $\theta : \mathcal{S} \to \mathcal{T}$. For any theme $\tau \in \mathcal{T}$, define

$$\mathcal{S}_\tau = \{S_i \mid \theta(S_i) = \tau\}, \quad n_\tau = |\mathcal{S}_\tau|.$$

An augmented example is created by selecting any subset $\mathcal{C} \subseteq \mathcal{S}_\tau$ with $|\mathcal{C}| = k \geqslant 2$ and concatenating sentences in their original order. The total number of concatenated samples is

$$S = \sum_{\tau \in \mathcal{T}} \sum_{k=2}^{n_\tau} \binom{n_\tau}{k}.$$

The mean augmented length is

$$\bar{L}' = \frac{1}{S} \sum_{\tau \in \mathcal{T}} \sum_{\substack{\mathcal{C} \subseteq \mathcal{S}_\tau \\ |\mathcal{C}| \geqslant 2}} \left(\sum_{S_i \in \mathcal{C}} L_i \right).$$

Example. If three sentences S_a, S_b, S_c share the same theme, all valid concatenations include

$$S_a \oplus S_b, \quad S_a \oplus S_c, \quad S_b \oplus S_c, \quad S_a \oplus S_b \oplus S_c,$$

where $\oplus$ denotes sentence-level concatenation.

3.2 Swapping (Sentence Reordering)

This method permutes the order of sentences within short paragraphs or compound-sentence structures. Each individual sentence remains unchanged, preserving local semantics, while discourse-level meaning may shift.

Let $P = \{S_1, S_2, \ldots, S_n\}$ be a paragraph of n sentences. A swapping operation is any non-identity permutation,

$$T : P \to P', \quad P' \neq P.$$

The number of distinct augmented variants is

$$g(n) = \sum_{k=2}^{n} \binom{n}{k} (k! - 1).$$

Example. For $P = \{S_a, S_b, S_c\}$, valid reorderings include

$$S_b \to S_a \to S_c, \quad S_c \to S_b \to S_a, \quad S_a \to S_c \to S_b.$$

3.3 Replacement-Based Augmentation

This family introduces lexical variation using predefined token groups. Each token w belongs to a synonym group $\mathcal{G}_{\text{syn}}(w)$ and a theme group $\mathcal{G}_{\text{theme}}(\theta(w))$. For any replacement group

$$\mathcal{G}_k = \{w_1, \ldots, w_r\},$$

valid substitutions are defined as

$$\varphi : w_i \mapsto w_j, \quad w_j \in \mathcal{G}_k \setminus \{w_i\}.$$

The number of possible substitutions is

$$f(r) = r(r-1).$$

Synonym Replacement A proportion $p \in (0, 1]$ of sentences (index set Ω_p) is selected. Eligible tokens are replaced by synonyms,

$$w' \in \mathcal{G}_{\text{syn}}(w), \quad w' \neq w.$$

Let L_i' be the length of the augmented sentence, and define

$$\alpha_i = \frac{L_i' - L_i}{L_i}.$$

Then

$$S = pN, \quad \bar{L}' = \frac{1}{S} \sum_{i \in \Omega_p} L_i' \approx (1 + \bar{\alpha}) \bar{L}, \quad \bar{\alpha} = \frac{1}{|\Omega_p|} \sum_{i \in \Omega_p} \alpha_i.$$

Example. For the sentence

$$S = \text{“The old house was quiet.”},$$

with

$$\mathcal{G}_{\text{syn}} = \{\text{old, ancient, elderly}\},$$

a valid replacement is

$$S' = \text{“The ancient house was quiet.”}.$$

Theme-Based Replacement Similarly, a proportion $q \in (0, 1]$ of sentences (index set Ω_q) undergo theme-consistent substitution:

$$w' \in \mathcal{G}_{\text{theme}}(\theta(w)), \quad w' \neq w.$$

Define

$$\beta_i = \frac{L_i' - L_i}{L_i}.$$

Then

$$S = qN, \quad \bar{L}' = \frac{1}{S} \sum_{i \in \Omega_q} L'_i \approx (1+\bar{\beta})\bar{L}, \quad \bar{\beta} = \frac{1}{|\Omega_q|} \sum_{i \in \Omega_q} \beta_i.$$

Example. Given the theme LOCATION with group

$$\mathcal{G}_{\text{theme}}(\text{LOCATION}) = \{\text{village, forest, riverbank}\},$$

the sentence

$$S = \text{“They travelled to the village.”}$$

may be augmented as

$$S' = \text{“They travelled to the forest.”}.$$

3.4 Sliding Window Segmentation

This method extracts overlapping contiguous segments using a fixed window size $w \in \mathbb{N}$, preserving local syntax and co-occurrence patterns. For a sentence of length L_i, the number of segments is

$$n_i = \max(0,\, L_i - w + 1),$$

and thus

$$S = \sum_{i=1}^{N} n_i, \quad \bar{L}' = w.$$

Formally,

$$\text{Win}(x_i, w) = \{(t_{i,j}, \ldots, t_{i,j+w-1}) \mid 1 \leqslant j \leqslant L_i - w + 1\}.$$

Example. For

$$x = (\text{A}, \text{B}, \text{C}, \text{D}), \quad L = 4, \quad w = 2,$$

the extracted segments are

$$(\text{A}, \text{B}), \quad (\text{B}, \text{C}), \quad (\text{C}, \text{D}).$$

3.5 Insertion (Contextual Place–Time Expansion)

This method inserts place or time units from a controlled set $\mathcal{G}_{\text{pt}} = \{v_1, \ldots, v_r\}$ before sentence-ending punctuation. For

$$x_i = (t_{i,1}, \ldots, t_{i,L_i}, \alpha),$$

insertion produces

$$x'_i = (t_{i,1}, \ldots, t_{i,L_i}, v_j, \alpha), \quad v_j \in \mathcal{G}_{\text{pt}}.$$

If multiple units may be inserted, any non-empty subset is valid:

$$R_i = 2^r - 1, \quad S = \sum_{i=1}^{N} R_i.$$

Example. For

$$x = (\text{A}, \text{B}, \text{C}, \alpha), \quad \mathcal{G}_{\text{pt}} = \{\text{u}, \text{v}\},$$

valid variants include

$$(\text{A}, \text{B}, \text{C}, \text{u}, \alpha), \quad (\text{A}, \text{B}, \text{C}, \text{v}, \alpha), \quad (\text{A}, \text{B}, \text{C}, \text{u}, \text{v}, \alpha).$$

3.6 Deletion (Exhaustive Token Removal)

This method removes exactly one token from every possible position:

$$x_i = (t_{i,1}, \ldots, t_{i,L_i}) \quad \Rightarrow \quad D_i = L_i, \quad S = \sum_{i=1}^{N} D_i.$$

Each deletion yields a sequence of length $L_i - 1$, giving

$$\bar{L}' = \frac{\sum_{i=1}^{N} L_i(L_i - 1)}{\sum_{i=1}^{N} L_i}.$$

Example. For

$$x = (\text{A}, \text{B}, \text{C}),$$

deletion variants are

$$(\text{B}, \text{C}), \quad (\text{A}, \text{C}), \quad (\text{A}, \text{B}),$$

corresponding to removals at positions 1, 2, and 3.

4 Training Pipeline for Low-resource NMT

We adopt a three-stage training strategy, following (Nguyen et al. 2025), to adapt a pretrained encoder–decoder Transformer model for extremely low-resource translation in the Tày–Vietnamese and Bahnar–Vietnamese language pairs. Our system is based on *BART*, a denoising autoencoder with a bidirectional encoder and an autoregressive decoder (Lewis et al. 2020). During pre-training, parts of the input are masked or permuted, and the model learns to reconstruct the original text. This denoising objective makes BART well suited for noisy or morphologically diverse data, and its encoder–decoder structure aligns naturally with translation, where the encoder models the source language and the decoder produces the target-language output. We use a three-stage pipeline because each stage addresses a different challenge in low-resource NMT: Vietnamese-based initialization supplies a strong linguistic prior, continual LM pre-training adapts the model to the minority language, and supervised fine-tuning learns the final translation mapping. Separate models are trained for each translation direction.

Vietnamese-based Initialization We initialize our models using BARTPho (Tran, Le, and Nguyen 2022), which is pretrained on 145 million Vietnamese sentences. Instead of training from scratch, we directly load the BARTPho checkpoint and retain its corruption functions, including random masking and sentence permutation, to preserve exposure to Vietnamese syntactic and discourse patterns. This step compensates for the lack of large Tày or Bahnar corpora by providing a strong Vietnamese linguistic foundation, and BARTPho serves as a typologically compatible and efficient starting point for subsequent LM adaptation; an overview of the BART architecture and its denoising objective is shown in Figure 1.

Continual Pre-training on Tày and Bahnar The second stage adapts the Vietnamese-initialized model to each minority language by performing continual LM pre-training

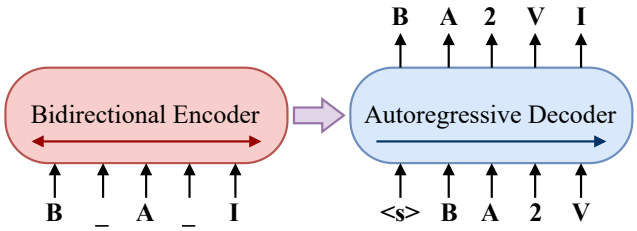

Figure 1: The architecture of BART and its denoising-based training objective.

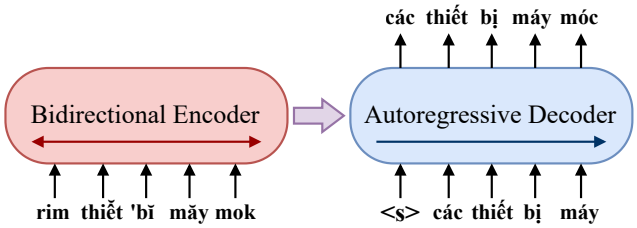

Figure 2: Supervised fine-tuning for the Tày–Vietnamese and Bahnar–Vietnamese translation tasks.

on monolingual Tày or Bahnar text. We reuse the denoising objective from BARTPho pre-training, in which some input tokens are masked and the model reconstructs the full sequence. Although the available corpora are small, this stage allows the model to internalize language-specific lexical distributions, function-word behavior, and morphosyntactic patterns that differ markedly from Vietnamese. It also narrows the distribution gap between the Vietnamese-based initialization and the minority-language data used during supervised translation.

Fine-tuning for Translation In the final stage, we fine-tune two separate NMT systems, one for Tày→Vietnamese and one for Bahnar→Vietnamese. The encoder receives an uncorrupted Tày or Bahnar sentence, and the decoder generates the Vietnamese translation using the standard sequence-to-sequence cross-entropy objective, as shown in Figure 2. Each DA method described in Section 3 is applied independently to create a distinct augmented training set, and a separate model is fine-tuned for each method. This design ensures that the effect of every augmentation strategy is evaluated in isolation, without interference from others, enabling a controlled comparison across two typologically distinct languages.

5 Experimentations

5.1 Datasets

We constructed two high-quality parallel corpora for the Tày–Vietnamese and Bahnar–Vietnamese language pairs through a combination of fieldwork, community collaboration, and careful linguistic processing.

Tày To build the Tày–Vietnamese dataset, we conducted fieldwork in remote Tày-speaking regions of northern Vietnam. With community permission, we collected locally used Tày-language textbooks and Tày–Vietnamese dictionaries provided by two regional universities. All materials were digitized, manually cleaned, and standardized into a unified Latin-based orthographic format reflecting contemporary Tày conventions. A trained annotation team then performed sentence segmentation, orthographic normalization, and bilingual alignment. The resulting parallel corpus was reviewed by native Tày speakers to ensure linguistic accuracy before being used in NMT training and data augmentation experiments.

Bahnar For the Bahnar–Vietnamese dataset, we conducted extensive field surveys across Bahnar-speaking communities in Vietnam's Central Highlands. Our sources included direct elicitation sessions with native speakers, printed materials such as religious books, newspapers, and song lyrics, and archival documents including local bulletins and historical texts. After digitization, an annotation team standardized spelling, resolved dialectal variation, and aligned Bahnar sentences with their Vietnamese equivalents. The final corpus, together with a derived Bahnar–Vietnamese lexicon used for dictionary-based augmentation, was validated by native Bahnar speakers before being integrated into the training pipeline.

Dataset Statistics Table 1 summarizes key statistics for both corpora on the minority-language and Vietnamese sides across the training and test splits.

5.2 Evaluation Metrics

We evaluate translation quality using the *BiLingual Evaluation Understudy* (BLEU) (Papineni et al. 2002) and the *Metric for Evaluation of Translation with Explicit ORdering* (METEOR) (Banerjee and Lavie 2005). Both metrics capture lexical and structural correspondence between model outputs and human references, making them suitable for the Tày–Vietnamese and Bahnar–Vietnamese translation tasks.

5.3 Baselines

The baseline systems are trained using the same three-stage pipeline described in Section 4. Each model starts from the Vietnamese-pretrained BARTPho checkpoint, undergoes continual pre-training on monolingual Tày or Bahnar data, and is then fine-tuned solely on the original human-authored parallel corpus. No data augmentation is applied at any stage. These models therefore serve as strong, fully trained low-resource NMT systems that capture the full benefit of the LM-based pipeline while providing a clean, augmentation-free reference point. All DA-enhanced systems are compared directly against these baselines to quantify the effect of each augmentation method.

We evaluate every augmentation technique introduced in Section 3, covering both individual transformations and composite variants. The methods considered in our experiments are

- *Insertion* + *Swap* (Ins + Swap),
- *Sentence Reordering* (Swap),
- *Insertion* + *Original* (Ins + Orig),

Table 1: Statistics of the Tày–Vietnamese and Bahnar–Vietnamese datasets (minority-language side vs. Vietnamese side, for train and test splits), computed using the model tokenizer.

	Tày–Vietnamese		**Bahnar–Vietnamese**	
Statistic	**Minority**	**Vietnamese**	**Minority**	**Vietnamese**
Sentences (train / test)	20,554 / 2,295		51,930 / 2,001	
Total tokens (train / test)	154,058 / 17,084	112,223 / 12,468	2,407,456 / 117,694	1,380,616 / 61,165
Avg. length (train / test)	7.50 / 7.44	5.46 / 5.43	46.36 / 58.82	26.59 / 30.57
Max length (train / test)	285 / 202	193 / 135	470 / 418	317 / 182

Table 2: Statistics of the augmented training sets for all DA methods across the Tày–Vietnamese and Bahnar–Vietnamese parallel corpora. Each row reports the total number of sentences and the average sentence length (in tokens) after augmentation.

Dataset / Metric	**Side**	**Base**	**Ins + Swap**	**Swap**	**Ins + Orig**	**Sliding**	**Combine**	**Syn**	**Theme**	**Delete**	**Del + Orig**
Tày–Vietnamese Parallel Corpus											
Sentences	–	20,554	22,412	22,464	41,108	63,209	205,515	74,939	57,979	67,197	87,751
Avg. length	Tày	7.50	28.48	24.86	9.34	5.03	14.22	14.83	13.52	17.02	14.79
	Vietnamese	5.46	21.10	18.48	6.81	3.47	10.36	10.13	9.79	11.70	10.24
Bahnar–Vietnamese Parallel Corpus											
Sentences	–	51,930	1,385,628	2,342,835	83,108	10,863,844	169,543	682,091	682,091	836,006	887,936
Avg. length	Bahnar	46.36	83.22	49.61	64.02	31.21	99.91	71.95	71.95	73.52	71.93
	Vietnamese	26.59	56.34	44.72	36.67	18.79	62.70	41.18	41.18	41.50	40.58

- *Sliding Window Segmentation* (Sliding),
- *Thematic Concatenation* (Combine),
- *Synonym Replacement* (Syn),
- *Theme-Based Replacement* (Theme),
- *Exhaustive Deletion* (Delete), and
- *Deletion + Original* (Del + Orig).

Formal definitions and token-level length formulas for these methods appear in Section 3, and the corresponding augmented dataset statistics are reported in Table 2.

5.4 Experimental Setup

All baseline models are trained for 15 epochs on the original training splits shown in Table 1 using a single A100 40GB GPU. We employ the AdamW optimizer (Loshchilov and Hutter 2019) with a learning rate of 2×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, a batch size of 256, and gradient accumulation over 2 steps. Early stopping based on validation BLEU is applied with a patience of 3 epochs.

To assess the impact of data augmentation in extremely low-resource conditions, we fine-tune a separate model for each DA method using the corresponding baseline checkpoint as initialization. This setup ensures that all DA variants inherit the same Vietnamese-based initialization and the same language-specific continual pre-training, enabling a strictly controlled and fair comparison across augmentation techniques. All hyperparameters are kept identical across DA experiments.

5.5 Results and Analysis

Table 3 summarizes the performance of all augmentation methods on the Tày–Vietnamese and Bahnar–Vietnamese test sets. The baseline corresponds to continual pre-training followed by supervised fine-tuning without augmentation. Most methods outperform the baseline for both language pairs, although the magnitude of improvement varies considerably. Notably, augmentation techniques that produce very large corpora, such as Sliding Window, do not necessarily yield the best results. This suggests that data quantity alone is not a reliable indicator of translation quality, and that the linguistic nature of each transformation plays a decisive role.

Tày–Vietnamese Almost all augmentation methods improve upon the baseline BLEU score of 17.13. Swap provides a modest gain, while Sliding Window (26.43 BLEU) and Combine (26.84 BLEU) deliver stronger improvements. Lexical approaches perform well. Synonym Replacement reaches 28.20 BLEU, Theme-Based Replacement reaches 28.66 BLEU, and Deletion reaches 28.91 BLEU. The best overall performance comes from Deletion + Original, which attains 31.86 BLEU and 0.3236 METEOR. These results indicate that compact and semantically faithful modifications can be more effective than broader structural expansions.

Bahnar–Vietnamese Bahnar exhibits a similar pattern, although with tighter typological constraints. Several methods substantially outperform the baseline BLEU of 10.41. Synonym Replacement reaches 21.68 BLEU, the strongest among single methods, and both Theme-Based Replacement and Deletion provide clear gains. Composite strategies fur-

Table 3: Translation performance of all DA methods on the Tày–Vietnamese and Bahnar–Vietnamese test sets.

	Tày–Vietnamese		Bahnar–Vietnamese	
DA Method	**BLEU**	**METEOR**	**BLEU**	**METEOR**
Original (Baseline)	17.13	0.1918	10.41	0.2822
Insert + Swap	12.02	0.1658	7.56	0.1905
Swap	18.64	0.2047	13.74	0.2758
Insert + Original	25.80	0.2874	12.18	0.2921
Sliding Window	26.43	0.2863	16.37	0.2640
Combine	26.84	0.3122	16.63	0.3170
Synonym Replacement	28.20	0.2905	21.68	0.3459
Theme-Based Replacement	28.66	0.3294	20.19	0.3210
Deletion	28.91	0.2847	19.45	0.3323
Deletion + Original	**31.86**	**0.3236**	**22.37**	**0.3581**

ther reinforce these improvements when applied carefully. Deletion + Original again achieves the best result, with 22.37 BLEU and 0.3581 METEOR, more than doubling the baseline. In contrast, highly disruptive transformations such as Insert + Swap consistently underperform.

General Trends Across both languages, two clear patterns emerge: (1) Deletion + Original yields the strongest improvements, and (2) Insert + Swap is consistently the weakest, suggesting that highly disruptive perturbations introduce noise the model cannot effectively use. More broadly, these results show that raw data volume alone is a poor predictor of translation quality; what matters is how each augmentation interacts with the structural properties of the language. These observations motivate a typology-aware analysis of why augmentation behaves differently across languages, presented in Section 6.

6 Linguistic Analysis

Although Tày and Bahnar are spoken within Vietnam, they belong to distinct genealogical families and exhibit markedly different morphosyntactic structures. These differences play a central role in determining how stable or fragile each augmentation method becomes.

6.1 Language Genealogy

Tày as a Central Tai Language Tày belongs to the Central Tai subgroup of the Tai–Kadai family and is closely related to Nùng and several Zhuang varieties. Prior research (Li 1977; Liao 2017) shows that Central Tai languages preserve conservative tone systems, monosyllabic morphemes, and a fixed SVO order. Tày also shares several contact-induced similarities with Vietnamese, such as analytic morphology, classifier-based noun phrase structure, and transparent token-to-morpheme alignment. These properties make surface-level augmentation relatively safe because common operations such as swapping, deleting, or replacing tokens rarely disrupt core grammatical relations.

Bahnar as a Central Bahnaric Language Bahnar belongs to the Central Bahnaric subgroup of the Austroasiatic family and displays hallmark Mon–Khmer morphosyntax, including productive prefixation and infixation, multisyllabic prosody, and valence-changing constructions such as the causative *po*, the passive or reciprocal *to*, and the completive *jo*. The verbal system comprises multiple licensed verb classes with strict compatibility constraints (Alves 2019). Negation and aspect markers occupy fixed pre-verbal positions, and predicate structure is encoded primarily through morphological alternation rather than word order. This means that inserting or swapping tokens within a verb complex, or separating a prefix from its stem, easily produces ungrammatical forms. Bahnar therefore requires augmentation methods that preserve internal verb morphology and valence relationships.

6.2 Implications for Multilingual DA

The genealogical and typological distance between Tày and Bahnar leads to fundamentally different expectations for augmentation design. Tày has an analytic structure that allows flexible token-level manipulation, whereas Bahnar's agglutinative morphology imposes strict boundaries on what perturbations remain grammatical. Table 4 summarizes the core contrasts and their implications for augmentation.

6.3 Why Augmentation Behaves Differently

Surface-Level DA Works Well for Tày The analytic structure of Tày creates a high degree of tolerance for surface-level modifications. The substantial gain from Deletion + Original (31.86 BLEU compared with 17.13 for the baseline) shows that the model benefits from exposure to both reduced and full versions of the same clause. Because Tày permits the omission of peripheral modifiers without altering core argument structure, deletion rarely produces ungrammatical sentences. Lexical replacements also tend to preserve classifier–noun compatibility and part-of-speech alignment. Insert + Swap is the main exception because inserting or moving tokens may interrupt clause-final particles or aspect markers that require fixed adjacency.

Table 4: Genealogical and typological differences between Tày (Tai–Kadai) and Bahnar (Austroasiatic), and their implications for augmentation.

Feature	Tày (Tai–Kadai)	Bahnar (Austroasiatic)	Implications
Morphology	Analytic and isolating	Agglutinative with prefixation and infixation	Token-level edits generally safe for Tày; Bahnar requires morphology-aware operations
Tone	5–7 tones	Minimal or absent	Tone-sensitive augmentation applies only to Tày
Word order	Fixed SVO	SVO with valence-modifying morphology	Bahnar augmentation must maintain verb-complex structure
Noun phrase structure	Quantifier + Classifier + Noun	Noun + Modifier	Classifier-sensitive operations apply only to Tày
Verb classes	Uniform verb behavior	Multiple licensed verb classes	Bahnar augmentation must respect verb-class compatibility
Morphosyntax	Transparent dependencies	Non-local dependencies across prefixes, stems, aspect, and negation	Insert or swap operations safe for Tày but fragile for Bahnar

Morphology-Aware DA Is Required for Bahnar Bahnar's agglutinative morphology demands much greater care. Prefixes, aspect markers, and negation must remain in the pre-verbal position. Perturbations that violate these constraints, such as Insert + Swap, lead to ungrammatical patterns and account for the sharp drop to 7.56 BLEU from a 10.41 baseline. In contrast, Deletion + Original (22.37 BLEU) and controlled lexical replacement preserve the integrity of verb complexes while still offering useful variation. These methods introduce diversity without crossing morphological or valence boundaries.

6.4 Implications for Low-Resource NLP

The analysis suggests several broader principles:

- Analytic languages such as Tày, Vietnamese, and Lao benefit from surface-level augmentation because token edits rarely disrupt core grammatical relations.
- Morphologically rich languages such as Bahnar, Khmer, and Sedang require augmentation strategies that respect prefixation, verb-class licensing, and valence-related morphology.
- A universal augmentation pipeline is unlikely to succeed across diverse language families without explicit grammatical modeling.
- Genealogical distance strongly predicts how sensitive a language will be to particular augmentation strategies and how robust the generated variants will be.

6.5 Discussion

Taken together, our findings highlight that augmentation effectiveness is driven primarily by typological fit: the same transformation can be beneficial in one language yet detrimental in another, even under comparable data scales. Analytic languages such as Tày, where morpheme boundaries align cleanly with token boundaries, tolerate surface-level perturbations and therefore benefit from techniques like Deletion + Original or controlled lexical replacement. In contrast, Bahnar's agglutinative morphology, strict prefixation patterns, and valence-related dependencies make the language highly sensitive to token insertion and reordering, causing methods such as Insert + Swap to break verb-complex integrity and introduce ungrammatical forms. This contrast underscores the need for typology-aware augmentation frameworks that adapt perturbation rules to each language's morphological and syntactic constraints rather than applying uniform operations across languages. Incorporating these principles may improve the robustness, fairness, and real-world usability of NMT systems for typologically diverse and underserved language communities.

7 Conclusion

This work provides the first systematic evaluation of data augmentation for Tày and Bahnar within a controlled NMT pipeline, establishing new bilingual benchmarks and demonstrating that meaning-preserving transformations offer consistent gains when they maintain semantic coherence and respect each language's structural constraints. While the study offers clear empirical and typological insights, it focuses primarily on surface-level operations and is restricted by limited monolingual data, which limits the modeling of deeper morphological and syntactic phenomena and leaves automatic metrics only partially reflective of linguistic adequacy. Future work may explore morphology-aware augmentation, incorporate syntactic cues from symbolic or LM- based resources, and broaden cross-linguistic evaluation, alongside community-driven data expansion and multimodal supervision to strengthen NMT systems for underserved and typologically diverse language communities.

References

Alves, M. J. 2019. Morphology in Austroasiatic Languages.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Goldstein, J.; Lavie, A.; Lin, C.-Y.; and Voss, C., eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

Chen, S.; Jin, Q.; and Fu, J. 2019. From Words to Sentences: A Progressive Learning Approach for Zero-resource Machine Translation with Visual Pivots. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 4932–4938. International Joint Conferences on Artificial Intelligence Organization.

Elmadani, K. N.; and Buys, J. 2024. Neural Machine Translation between Low-Resource Languages with Synthetic Pivoting. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 12144–12158. Torino, Italia: ELRA and ICCL.

Fabbri, A.; Han, S.; Li, H.; Li, H.; Ghazvininejad, M.; Joty, S.; Radev, D.; and Mehdad, Y. 2021. Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 704–717. Online: Association for Computational Linguistics.

Fadaee, M.; Bisazza, A.; and Monz, C. 2017. Data Augmentation for Low-Resource Neural Machine Translation. In Barzilay, R.; and Kan, M.-Y., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 567–573. Vancouver, Canada: Association for Computational Linguistics.

He, D.; Xia, Y.; Qin, T.; Wang, L.; Yu, N.; Liu, T.-Y.; and Ma, W.-Y. 2016. Dual learning for machine translation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 820–828. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.

Her, W.-h.; and Kruschwitz, U. 2024. Investigating Neural Machine Translation for Low-Resource Languages: Using Bavarian as a Case Study. In Melero, M.; Sakti, S.; and Soria, C., eds., *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, 155–167. Torino, Italia: ELRA and ICCL.

Holm, D. 2020. The Tày and Zhuang vernacular scripts: Preliminary comparisons. *Journal of Chinese Writing Systems*, 4(3): 197–213.

Hujon, A. V.; Singh, T. D.; and Amitab, K. 2023. Transfer Learning Based Neural Machine Translation of English-Khasi on Low-Resource Settings. *Procedia Computer Science*, 218: 1–8. International Conference on Machine Learning and Data Engineering.

Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; Hughes, M.; and Dean, J. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5: 339–351.

Khoboko, P. W.; Marivate, V.; and Sefara, J. 2025. Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models. *Machine Learning with Applications*, 20: 100649.

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018. Unsupervised Machine Translation Using Monolingual Corpora Only. In *International Conference on Learning Representations*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Li, B.; Hou, Y.; and Che, W. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3: 71–90.

Li, F. K. 1977. A Handbook of Comparative Tai. *Oceanic Linguistics Special Publications*, i–389.

Liao, H. 2017. Proto-Tai reconstruction of 'maternal grandmother' revisited. *Language and Linguistics*, 18(1): 116–140.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Mai, H.-H.; and Luong, N. H. 2023. Data Augmentation with GPT-3.5 for Vietnamese Natural Language Inference. In *2023 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 435–440.

Miller, G. A. 1994. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Min, J.; McCoy, R. T.; Das, D.; Pitler, E.; and Linzen, T. 2020. Syntactic Data Augmentation Increases Robustness to Inference Heuristics. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2339–2352. Online: Association for Computational Linguistics.

Nguyen, L.; Le, T.; Nguyen, H.; Vo, Q.; Nguyen, P.; and Quan, T. 2025. Serving the Underserved: Leveraging BART-Bahnar Language Model for Bahnaric-Vietnamese Translation. In Truong, S.; Putri, R. A.; Nguyen, D.; Wang, A.; Ho, D.; Oh, A.; and Koyejo, S., eds., *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, 32–41. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-242-8.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Raja, R.; and Vats, A. 2025. Parallel Corpora for Machine Translation in Low-Resource Indic Languages: A Comprehensive Review. In Ojha, A. K.; Liu, C.-h.; Vylomova, E.; Pirinen, F.; Washington, J.; Oco, N.; and Zhao, X., eds., *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT*

2025), 129–143. Albuquerque, New Mexico, U.S.A.: Association for Computational Linguistics. ISBN 979-8-89176-230-5.

Ranathunga, S.; Lee, E.-S. A.; Prifti Skenduli, M.; Shekhar, R.; Alam, M.; and Kaur, R. 2023. Neural Machine Translation for Low-resource Languages: A Survey. *ACM Comput. Surv.*, 55(11).

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Improving Neural Machine Translation Models with Monolingual Data. In Erk, K.; and Smith, N. A., eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 86–96. Berlin, Germany: Association for Computational Linguistics.

Shi, S.; Wu, X.; Su, R.; and Huang, H. 2022. Low-resource Neural Machine Translation: Methods and Trends. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

Tran, N. L.; Le, D. M.; and Nguyen, D. Q. 2022. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association*.

Wang, R.; Tan, X.; Luo, R.; Qin, T.; and Liu, T.-Y. 2021. A Survey on Low-Resource Neural Machine Translation. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 4636–4643. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Wang, W. Y.; and Yang, D. 2015. That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets. In Màrquez, L.; Callison-Burch, C.; and Su, J., eds., *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2557–2563. Lisbon, Portugal: Association for Computational Linguistics.

Wei, J.; and Zou, K. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6382–6388. Hong Kong, China: Association for Computational Linguistics.

Xia, M.; Kong, X.; Anastasopoulos, A.; and Neubig, G. 2019. Generalized Data Augmentation for Low-Resource Translation. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5786–5796. Florence, Italy: Association for Computational Linguistics.

Yan, G.; Li, Y.; Zhang, S.; and Chen, Z. 2019. Data Augmentation for Deep Learning of Judgment Documents. In Cui, Z.; Pan, J.; Zhang, S.; Xiao, L.; and Yang, J., eds., *Intelligence Science and Big Data Engineering. Big Data and Machine Learning*, 232–242. Cham: Springer International Publishing. ISBN 978-3-030-36204-1.

Zhang, C.; Tan, X.; Ren, Y.; Qin, T.; Zhang, K.; and Liu, T.-Y. 2021. UWSpeech: Speech to Speech Translation for Unwritten Languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16): 14319–14327.

Zhang, Y.; Ge, T.; and Sun, X. 2020. Parallel Data Augmentation for Formality Style Transfer. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3221–3228. Online: Association for Computational Linguistics.