WMT 2025 CreoleMT Systems Description : Martinican Creole and French

Ludovic Mompelat
Department of Modern Languages
University of Miami
Miami, FL, USA
lvm861@miami.edu

Abstract

This paper describes our submissions to the constrained subtask of the WMT25 Creole Machine Translation shared task. We participated with a bidirectional Martinican Creole ↔ French system. Our work explores training-time strategies tailored for low-resource MT, including LoRA fine-tuning, curriculum sampling, gradual unfreezing, and multitask learning. We report competitive results against the baseline on both translation directions.

1 Introduction

The WMT25 Creole MT shared task is the first shared task focused on machine translation involving Creole languages. As part of the WMT25 Creole MT shared task (Robinson et al., 2025), we submitted constrained systems for Martinican Creole ↔ French. This track explores how to improve Creole MT under limited resource conditions. Participants were restricted to using only the official training data provided by the organizers and were allowed to initialize their models from the baseline `kreyol-mt-pubtrain` model on HuggingFace.

Our submission focused on the Martinican Creole ↔ French language pair (ISO: `mart1259-fra`). We trained and evaluated a bidirectional system for both `mart1259`→`fra` and `fra`→`mart1259` directions.

Motivation Our participation stems from ongoing work on NLP tools for Creole languages, including Martinican Creole and Haitian Creole. Despite progress in NLP, Creole languages remain underrepresented and under-resourced, making them ideal candidates for constrained MT settings. We aimed to investigate how far one can push model performance without access to expanded corpora by instead exploring training strategies and fine-tuning configurations tailored to low-resource language modeling.

This work is part of a broader effort to develop NLP tools for Creole researchers, educators, and communities. In particular, we aim to build usable MT systems, including domain-specific applications (e.g., medical MT for Haitian Creole). We explore various configurations: custom tokenizer, curriculum sampling, LoRA fine-tuning, gradual unfreezing, denoising, label smoothing, weighted BLEU scoring, multitask learning, and mixed training strategies.

2 Task Overview and Constraints

The WMT25 constrained track permits only the use of official datasets provided by the organizers. No additional monolingual or parallel data was allowed for training, tuning, or backtranslation. The baseline system is the publicly available `kreyol-mt-pubtrain` model on HuggingFace.

2.1 Data Used

We used the official datasets for the `mart1259-fra` pair, cleaned and released by the organizers:

- **Dictionnaire créole martiniquais–français** by Raphaël Confiant (2007)[1].
- CAPES corrections for Guadeloupean Creole from the Kapes Kreyol project.

2.2 Baseline Scores

mart1259→fra. On the organizers' blind test set, our system scores **25.31 BLEU** and **49.08 chrF++**, below the baseline (**28.27 BLEU**, **50.43 chrF++**). While LoRA plus curriculum sampling proved stable during development, these settings did not surpass the strong baseline under official evaluation.

fra→mart1259. On the blind test set, our system achieves **25.76 BLEU** and **48.74 chrF++**, close to but slightly below the baseline (**26.49 BLEU**, **48.69 chrF++**). This direction is comparatively tighter,

[1] https://www.potomitan.info/dictionnaire/

Proceedings of the Tenth Conference on Machine Translation, pages 1198–1200
November 8-9, 2025 ©2025 Association for Computational Linguistics

suggesting our configuration is broadly competitive but does not yield consistent gains over baseline.

3 Data Preprocessing and Setup

We experimented with a 70/30 training/validation split (3,094 / 1,338) to increase the size of the development set for more stable evaluation but reverted back to the original 90/10 split as it yielded slightly better results. Training was conducted separately for each translation direction rather than using concatenated bidirectional data or multitask learning, as this yielded more consistent results in our preliminary experiments given the dataset provided.

3.1 Tokenization and Language Codes

We retained the baseline's SentencePiece model and language code scheme (e.g., `mart1259`, `fra`) to ensure vocabulary alignment. Custom `forced_bos_token_id` was applied during decoding to enforce direction.

3.2 Curriculum Sampling and Difficulty Weighting

We computed difficulty scores for each example and implemented curriculum sampling in early epochs, gradually introducing harder examples. Weighted BLEU was used for dev set evaluation to better align training progress with final performance.

4 Model Architecture and Training

All systems were initialized from the publicly available `jhu-clsp/kreyol-mt-pubtrain` model on HuggingFace. We applied Low-Rank Adaptation (LoRA) for efficient fine-tuning, targeting the attention on `q_proj` and `v_proj`. Our final configuration used rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.1, though we also explored $r \in \{8, 32\}$ and $\alpha \in \{64\}$ in preliminary runs.

Training was performed with a custom `Seq2SeqTrainer` that integrates curriculum sampling (ordering examples by difficulty), gradual unfreezing of encoder and decoder layers, and multitask loss weighting to balance direction-specific performance. We applied label smoothing of 0.1 to improve generalization and used weighted BLEU scoring on the development set as the main model selection criterion.

The final training configuration included a batch size of 32, maximum source and target lengths of 128 tokens, 50 training epochs, a constant learning rate of 2×10^{-5} with 500 warmup steps, and the AdamW optimizer. Evaluation used a beam size of 4 and a maximum of 128 new tokens during generation.

5 Evaluation

5.1 Final Systems

For `mart1259`→`fra`, our final constrained submission fine-tunes the publicly available `jhu-clsp/kreyol-mt-pubtrain` model, an mBART-50 style encoder–decoder Transformer, using only the official training set with the *original* 90/10 split. This direction uses Low-Rank Adaptation (LoRA) and curriculum sampling with difficulty weighting, as these settings yielded the most stable gains over the baseline.

For `fra`→`mart1259`, preliminary experiments showed that a 70/30 training/validation split improved dev set stability, while curriculum sampling provided no measurable benefit. We therefore retained LoRA but removed curriculum sampling for this direction.

In both cases, we retain the original SentencePiece tokenizer and language tags, forcing the target language with `decoder_start_token_id`. Training uses a batch size of 32, label smoothing (0.05), AdamW with a learning rate of 2×10^{-5}, cosine scheduling, and early stopping based on BLEU on the dev set. Generation uses beam search (beam size 4, max length 128 tokens).

Parameter	Value
Batch size	32
Max source/target length	128 / 128
Epochs	50
Learning rate	2e−5
LoRA rank r	16
LoRA α	32
LoRA dropout	0.05
Target modules	q_proj, v_proj
Label smoothing	0.1
Optimizer	AdamW
Warmup steps	500
Scheduler	constant
Beam size (eval)	4
Max new tokens (eval)	80

Table 1: Training and generation hyperparameters for both final constrained submissions.

No additional data or back-translation was

used. We fine-tune the publicly released `jhu-clsp/kreyol-mt-pubtrain` model with parameter-efficient adaptation (LoRA). Decoding uses beam search (beam size 4; max length 128), and we constrain the target language with `decoder_start_token_id`.

We report **BLEU** (Papineni et al., 2002) and **chrF++** (Popović, 2016). These are the *official* blind-test scores provided by the organizers (Robinson et al., 2025).

System	BLEU	chrF++
Baseline (official)	28.27	50.43
Ours (primary, official)	25.31	49.08

Table 2: `mart1259`→`fra` official blind-test results.

On the organizers' blind test set, our system is below the baseline (25.31 vs. 28.27 BLEU; 49.08 vs. 50.43 chrF++). While LoRA with curriculum sampling was stable in development, it did not surpass the strong baseline under official evaluation for this direction.

System	BLEU	chrF++
Baseline (official)	26.49	48.69
Ours (primary, official)	25.76	48.74

Table 3: `fra`→`mart1259` official blind-test results.

For `fra`→`mart1259`, our system is close to but slightly below the baseline in BLEU (25.76 vs. 26.49), with nearly identical chrF++ (48.74 vs. 48.69). The 70/30 split without curriculum sampling remained a stable configuration for this direction, but it did not yield a clear improvement over the baseline on the official evaluation.

Overall, direction-specific tuning—retaining curriculum sampling for `mart1259`→`fra` and removing it for `fra`→`mart1259`—produced systems that are broadly competitive but ultimately *do not surpass* the baseline under the constrained track conditions. In particular, the gap is larger for `mart1259`→`fra` and smaller for the reverse direction. A plausible explanation is the limited size and diversity of the training data: with few examples, systems may be less sensitive to changes in fine-tuning strategy or adapter placement, and estimates of improvements can have high variance. We therefore treat these direction-specific effects as suggestive rather than definitive.

6 Conclusion and Future Work

Under the constrained track, our LoRA-based systems with direction-specific training choices are close to but ultimately below the strong baseline on the organizers' blind test set. The gap is larger for `mart1259`→`fra` and smaller for `fra`→`mart1259`, where BLEU and chrF++ are nearly tied. One plausible factor is data scarcity: with limited and relatively homogeneous training material, core architectural or training-time changes may yield muted or unstable gains, and measurement noise can obscure small effects.

We will (i) explore alternative parameter-efficient adapters and layer targeting paired with data-scaling, (ii) tune curriculum schedules per direction and analyze where they help/hurt, (iii) align preprocessing/normalization and decoding with the organizers' pipeline, (iv) investigate tokenizer and segmentation choices for Martinican Creole, and (v) add robust automatic and human error analyses (e.g., code-switching, diacritics, OOVs). In the unconstrained setting, we also plan to study back-translation and multilingual transfer to further close the gap.

References

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.

Maja Popović. 2016. chrf deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504.

Nathaniel R. Robinson, Claire Bizon Monroc, Rasul Dent, Stefan Watson, Raj Dabre, Kenton Murray, Andre Coy, and Heather Lent. 2025. Findings of the first shared task for creole language machine translation at wmt25. In *Proceedings of the Tenth Conference on Machine Translation*.