

Enhancing Cultural Awareness for Haitian Creole Machine Translation

Tabitha Megane Bellune, Ngoc Tan Le, Fatiha Sadat

Université du Québec à Montréal, Québec, Canada

firstname.lastname@uqam.ca

Abstract

Natural Language Processing for low-resource languages faces a major challenge: the lack of structured and culturally grounded data. This article presents our contribution to the machine translation task from Haitian Creole to English/French. Our hybrid approach relies on the creation of an original parallel corpus of 17,033 sentence pairs, derived primarily from 52.63 hours of transcriptions from the Atlas Linguistique d’Haïti, combined with the synthetic generation of code-switching data. By comparing a Transformer model trained from scratch with the fine-tuning of the pre-trained NLLB-200 multilingual model, we demonstrate that cultural adaptation is crucial. According to official evaluation scores, our best system records a dramatic increase in its BLEU score of 50.86 on the cultural domain and achieves an average CSI-Match score of 76.99 in the direction of *hat-eng*, and 44.16 BLEU and 73.76 in terms of average CSI-Match in the direction of *hat-fra* respectively, proving its capacity to accurately translate complex idiomatic expressions.

1 Introduction

Natural Language Processing (NLP) has seen major advances in recent years, propelled by the emergence of neural architectures and Large Language Models (LLMs). While these models have vastly improved tasks such as machine translation, these advancements remain deeply unequal. High-resource languages benefit from massive corpora, while low-resource languages face a severe lack of structured and usable data (Wibawa et al., 2024).

Haitian Creole, the primary language of approximately 10 to 12 million speakers and spoken by 90 to 95% of the population of Haiti, perfectly illustrates this paradox (DeGraff, 2007). Despite the existence of oral archives

and informal texts, the language severely lacks standardized data for training translation systems capable of grasping the nuances of its cultural context.

To address the challenges posed by this shared evaluation task (Robinson et al., 2026), our main contributions are as follows:

- The creation of a cultural parallel corpus integrating poems, proverbs, and the extraction of 55 hours of oral data from the Atlas Linguistique d’Haïti (Fattier, 1998).
- The implementation of a hybrid architecture, comparing a Transformer model (Vaswani et al., 2017) trained from scratch (Fairseq) with the adaptation of massive multilingual models (NLLB-200) (Costa-Jussà et al., 2022).
- The generation and evaluation of code-switching data using large language models, via a Token-Level Replacement strategy (Wang et al., 2025).
- The evaluation of Cultural Awareness using the CSI-Match metric (Yao et al., 2024), demonstrating the impact of cultural adaptation on idiomatic expressions.

The remainder of this article is organized as follows:

- Section 2 provides an overview of Creole languages, with a particular focus on Haitian Creole and its linguistic challenges.
- Section 3 presents our proposed methodology for this shared task, and about the preparation of experimental data.

- Section 4 presents the overall performance of all experiments and their evaluation, including an error analysis. It also discusses the impact of cultural adaptation and code-switching.
- Section 5 provides conclusions and future research directions toward a more inclusive and ethically grounded NLP ecosystem.

2 Linguistic Challenges and Limitations of the State of the Art

Natural Language Processing (NLP) inherently privileges high-resource languages, often leaving marginalized and minority languages like Haitian Creole (*Kreyòl Ayisyen*) (DeGraff, 2007) underrepresented in global digital ecosystems (Lent et al., 2022). The current computational paradigm relies heavily on the availability of large-scale, well-annotated textual resources, a dependency that creates systematic inequalities across the world’s languages (Joshi et al., 2020; Blasi et al., 2022).

This framework inherently disadvantages low-resource languages, including many Indigenous and Creole languages shaped by histories of colonization, domination, and cultural erasure (Bird, 2020; Le Ngoc and Sadat, 2021).

Haitian Creole possesses analytic syntactic structures (SVO) that are independent of French. Unlike inflectional languages, tense, aspect, and modality are expressed through preverbal particles (*e.g.*, *te* for the past, *ap* for the progressive). Furthermore, the lack of grammatical gender—illustrated by the epicene pronoun *li* ("*li renmen fè manje*" meaning both "*he*" and "*she*" likes to cook)—poses contextual ambiguity problems for traditional systems.

Recent initiatives, such as the Kreyòl-MT corpus (14.5 million sentence pairs across 41 Creoles), have significantly advanced the field (Robinson et al., 2024). However, these general-domain corpora often lack local sociocultural grounding. Moreover, industrial translation solutions struggle to process dialectal variations and the frequent phenomenon of language alternation (code-switching) with En-

glish or French. Our methodology integrates specifically Haitian data to bridge this gap.

3 Methodology and Collected Data

To submit a robust system to this shared task, we developed a hybrid pipeline ranging from speech processing to the synthetic generation of textual data.

3.1 Constitution of the Hybrid Training Corpus

The training of our models relies on the combination of two sources:

- (1) The general corpus (CreoleMT or Kreyòl-MT): Used to establish the baseline of the models¹.
- (2) The novel cultural corpus: We built a set of 19,148 culturally anchored sentence pairs. A major technical contribution of this work is the exploitation of 73 recordings (55 hours) from the Atlas Linguistique d’Haïti (ALH) (Fattier, 1998).

Processing Pipeline

We selected 73 recordings from the Linguistic Atlas of Haiti (ALH), representing approximately 55 hours of raw audio in Haitian Creole. We extracted this data using a waterfall approach: the audios were first automatically transcribed, then the transcriptions were cleaned and used as textual data to enrich the cultural translation corpus. Since the ALH audios are long, averaging around 45 minutes, we performed a preliminary segmentation before transcription, resulting in an average of 135 segments.

These audio files, converted in wav format, were segmented into 20-second blocks, then transcribed using the *ZeeshanGeoPk/haitian-speech-to-text*² ASR model. The resulting Creole segments were translated into English using translation APIs, automatically filtered (by length ratio and language detection), and manually evaluated on a random sample to mitigate

¹<https://huggingface.co/datasets/jhu-clsp/kreyol-mt>

²<https://huggingface.co/ZeeShanGeoPk/haitian-speech-to-text>

Processing Step	Statistics
ALH Selected Audios	73
Raw audio duration	52.63 h
Chunks of 20 s	9,473
Noisy segments generated	~40,000
Segments after translation and preprocessing	~ 17,033 pairs

Table 1: Statistics of the ALH corpus after processing of audio recordings.

translationese biases. This rigorous process generated 17,033 exploitable field sentence pairs (see Table 1).

Cultural data were added to the general Kreyòl-MT data. Validation cultural data were also added to the development corpus to track model performance during training.

The final mixed corpus used for training comprises 55,536 sentence pairs (approximately 30% cultural data and 70% general data).

3.2 Synthetic Generation of Code-Switching

To evaluate the models’ robustness when faced with bilingual speakers, we utilized a Large Language Model (LLM) to generate code-switching data. We applied a Token-Level Replacement approach, prompting the LLM to replace 1 to 3 Creole words with their English equivalents while maintaining fluency and contextual meaning. Code-switched phrases are generated using a LLM³ from the following prompt:

Prompt: You are a Haitian Creole-English code-switching generator.

Task: Generate a natural Haitian Creole-English code-switched sentence using TOKEN-LEVEL REPLACEMENT.

Instructions:

- Keep the sentence mostly in Haitian Creole.
- Replace only one to three words depending on the length of the sentence with English.
- Do NOT translate the whole sentence.
- Keep the sentence fluent and natural.
- Preserve the original meaning.
- Output only the code-switched sentence.

The code-switching was used as a data augmentation strategy. Sentences containing a

mixture of Creole and English or French were generated and then added to the whole Kreyòl-MT corpus. The code-switching file was then concatenated with the general data during training. This approach allows us to assess whether exposure to multilingual sentences improves the robustness of the model.

4 Experimentations and Evaluations

4.1 Data Preprocessing and Setup

We experimented with the train, dev and test splits from Kreyòl-MT Data⁴ (667k-646-1,2k sentences respectively).

We used the baseline’s *SentencePiece* model⁵ and language code schema (e.g. hat, eng, fra) in order to ensure the vocabulary mapping.

4.2 Translation Architectures

All systems were initialized from the public model *jhu-clsp/kreyol-mt* available on HuggingFace.

We compared two learning paradigms:

- Model trained from scratch: A *Transformer* (6 encoder/decoder layers) trained specifically on our data using the *Fairseq* (Ott et al., 2019) library⁶.
- Adaptation of massive models: Fine-tuning of the pre-trained multilingual models based on *NLLB-200* (Costa-Jussà et al., 2022).

In terms of models, we performed the experimental plan, in the both direction of hat-eng and hat-fra, with the following configurations:

- E1: MT-Kreyòl (baseline)

⁴<https://huggingface.co/datasets/jhu-clsp/kreyol-mt>

⁵<https://github.com/google/sentencepiece>

⁶<https://github.com/facebookresearch/fairseq>

³[gpt-5.4-mini:https://openai.com/fr-FR/index/introducing-gpt-5-4-mini-and-nano/](https://openai.com/fr-FR/index/introducing-gpt-5-4-mini-and-nano/)

System	BLEU	chrF++	TER	BLEURT	Avg CSI match
E1: Baseline	43.60	65.00	45.00	0.086	49.97
E2: Ours (primary)	47.56	66.19	37.83	0.3232	75.77
E3: Ours (contrastive)	50.86	68.30	35.04	0.4127	76.99

Table 2: Performance of the *hat-eng* direction blind-test results (official evaluation scores).

System	BLEU	chrF++	TER	BLEURT	Avg CSI match
E1: Baseline	41.35	59.94	67.70	0.0368	70.97
E2: Ours (primary)	42.83	63.84	62.19	0.1204	73.56
E3: Ours (contrastive)	44.16	64.73	61.35	0.1288	73.76

Table 3: Performance of the *hat-fra* direction blind-test results (official evaluation scores).

- E2: MT-Kreyòl + code-switching data (primary)
- E3: MT-Kreyòl + code-switching data + cultural data (contrastive)

whereas E1 serves as a reference model based on *Transformer fairseq*, E2 allows us to evaluate the effect of code-switching, while E3 allows us to evaluate the combined effect of cultural data and code-switching. Both E2 and E3 are based on *NLLB-200* model.

We also performed a fine-tuning based on the *facebook/nllb-200-distilled-600M*⁷ model. The pre-trained model was then trained on Kreyòl-MT data as well as on various experimental setups. The trained models were used to generate the submitted translation files and evaluation outputs.

Then, our systems were evaluated using several standard metrics such as **BLEU** (Papineni et al., 2002), **chrF++** (Popović, 2016), **TER** (Snover et al., 2006), **BLEURT** (Sellam et al., 2020), to which we added **CSI-Match** to measure cultural awareness (Yao et al., 2024).

Unlike lexical metrics, **BLEURT** aims to more correctly capture the semantic similarity between the produced translation and the reference. Therefore, it allows us to assess if the overall meaning of the source phrase is correctly conveyed, even in situations when the translation differs lexically from the reference. **CSI-Match** is a crucial assessment indicator for gauging the model’s comprehension of cultural data.

Table 2 and Table 3 shown the performance of our systems in both direction of Cre-

ole Haitian to English and Creole Haitian to French.

4.3 Global Performances and Impact of Cultural Adaptation

When establishing the baseline on the general corpus, *NLLB-200* vastly outperformed *fairseq* especially in the *hat-eng* direction (BLEU score of 50.86 vs. 43.60 respectively, see Table 2). The impact of our fine-tuning on the cultural domain proved to be decisive:

- The **BLEU** score of the *NLLB-200* model jumped from 43.60 to 50.86 (+7.26 points), proving the vital necessity of specialization.
- The Translation Edit Rate (**TER**) dropped from 45.00 to 35.04, confirming that the adapted model’s translations require significantly less human post-editing.
- The primary model also saw its performance double, going from an initial **CSI-match** of 49.97 to 76.99 on the cultural corpus in the direction of *hat-eng*, and up to 73.76 on the cultural corpus in the direction of *hat-fra*. The results in Table 2 and Table 3 show that models adapted with cultural data achieve better **CSI-Match** scores than baseline models. The cultural *NLLB-200*-based model exhibits the best performance.

4.4 Evaluation of Cultural Awareness and Code-Switching

Translation success for Creole languages often relies on figurative meaning. Evaluated on the **CSI-Match** metric, the adapted *NLLB-200*

⁷<https://huggingface.co/facebook/nllb-200-distilled-600M>

model obtained the best average score of 76.99 (see Table 2).

Qualitative analysis perfectly illustrates this gain. When translating the Haitian proverb "*Fò ou jete pou ou ranmase*", baseline models produce meaningless literal translations ("*Throw or pick up*"). In contrast, our culturally adapted model successfully restores the true idiomatic semantics: "*You reap what you sow*" (see Table 5).

4.5 Discussion

While these results demonstrate the efficacy of our pipeline, NLP for low-resource languages still faces persistent barriers. Our synthetic code-switching generation method, though innovative, cannot fully capture the real, pragmatic, and spontaneous sociolinguistic complexity of bilingual speakers. Additionally, the use of automatic translations as an intermediate step to exploit ALH oral data may introduce residual biases.

5 Conclusion and Future Work

In conclusion, our participation demonstrates that cultural adaptation—including fine-tuning on field data, proverbs, and oral traditions—is not only relevant but essential for improving machine the *NLLB-200* model, combined with our hybrid corpus, opens promising avenues for further research.

For future work, we recommend the direct and systematic involvement of Haitian communities, particularly native annotators, in data validation and in the exploration of regional dialectal variants, such as Cap-Haïtien Creole. An ethical and community-centered approach to technology is essential for preserving and digitally modeling Haiti's linguistic identity.

6 Limitations

Despite the growing interest in low-resource language processing, Haitian Creole remains significantly underrepresented in the NLP landscape. Existing lexical resources are often unstructured (*e.g.*, PDFs or printed dictionaries) and fail to capture the dialectal diversity of the language.

Moreover, the limited involvement of Haitian communities in data collection and annotation processes undermines both the cultural validity and linguistic relevance of existing resources. Most corpora are developed without participatory methodologies, often overlooking key linguistic phenomena such as regional variation, code-switching, and pragmatic usage.

Finally, many experimental results related to Haitian Creole remain unpublished or scattered across non-academic platforms, making reproducibility and comparative evaluation difficult. These limitations highlight the urgent need for ethical, representative, and community-grounded resource development to better address the sociolinguistic complexity of Haitian Creole.

References

- Steven Bird. 2020. [Decolonising speech and language technology](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. [Systematic inequalities in language technology performance across the world's languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.
- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Michel DeGraff. 2007. Kreyòl ayisyen, or haitian creole (creole french). *Comparative creole syntax: Parallel outlines of*, 18:101–126.
- Dominique Fattier. 1998. *Contribution à l'étude de la genèse d'un créole: l'Atlas linguistique d'Haïti, cartes et commentaires*, 6 vol. Ph.D. thesis, Université de Provence, Villeneuve d'Ascq. Ph.D. Dissertation.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

- Tan Le Ngoc and Fatiha Sadat. 2021. [Revitalisation des langues autochtones via le prétraitement et la traduction automatique neuronale: le cas de l’inuktitut \(revitalization and preservation of indigenous languages through natural language processing\)](#). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 273–274, Lille, France. ATALA.
- Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. [What a creole wants, what a creole needs](#). In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 6439–6449, Marseille, France. European Language Resources Association (ELRA). Licensed under CC-BY-NC-4.0.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. [fairseq: A fast, extensible toolkit for sequence modeling](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.
- 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 Robinson, Ruth-Ann Armstrong, Claire Bizon Monroc, Rasul Dent, Pranav Gupta, Raj Dabre, Andre Coy, and Kenton Murray. 2026. Findings of the second shared task for creole language machine translation, at WMT 2026. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Nathaniel R Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Bizon Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome A Etori, and 1 others. 2024. Kreyolmt: Building mt for latin american, caribbean and colonial african creole languages. *arXiv preprint arXiv:2405.05376*.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7881–7892.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Zhijun Wang, Jiahuan Li, Hao Zhou, Rongxiang Weng, Jingang Wang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. Investigating and scaling up code-switching for multilingual language model pre-training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11032–11046.
- Aji Prasetya Wibawa, Fachrul Kurniawan, and 1 others. 2024. Advancements in natural language processing: Implications, challenges, and future directions. *Telematics and Informatics Reports*, 16:100173.
- Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. [Benchmarking machine translation with cultural awareness](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

A Hyper-parameters

The detailed training configuration is shown in the table 4.

Parameters	Value
Architecture	Transformer
Encoder/Decoder Layers	6 / 6
Embedding Size	512
Attention Heads	8
Feed-Forward Size	2048
Dropout	0.3
Attention Dropout	0.1
Activation Dropout	0.1
Loss Function	Label-Smoothed Cross-Entropy
Optimizer	Adam
Learning Rate	0.0005
Warmup Updates	4000
Max Tokens	2048
Max Epochs	30
Patience	5
GPUs Used	4

Table 4: Transformer fairseq-based model training configuration.

B Examples of translation on cultural data

As shown in the updated Table 5 below, our model (E3) produced: ‘*The strength you throw away is for you to pick up.*’ While this fails to generate the figurative English idiom (‘*You reap what you sow*’), it represents a major syntactic advance over E1 (‘*Forgot your throw to*

System	Translation output	Characterization
Source	<i>Fò ou jete pou ou ranmase.</i>	Haitian Creole Proverb
Reference	You reap what you sow.	Human Figurative Reference
E1 (baseline)	Forgot your throw to you pick up.	Ungrammatical / Fragmented
E2 (CS only)	Throw or pick up.	Extremely literal / Simplified
E3 (CS + cultural)	The strength you throw away is for you to pick up.	Cohesive literal translation

Table 5: Qualitative comparison of translations produced by our models.

you pick up’) and E2 (*Throw or pick up*’). We explain the linguistic root of this choice: the Creole preverbal modal particle *’fò*’ (meaning *’must/have to*’, from French *il faut*) is homophonous with the adjective *’fò*’ (meaning *’strong/strength*’, from French *fort*). Our model translated *’fò*’ as *’strength*’, demonstrating localized literal alignment while highlighting the ongoing challenge of figurative proverb translation.

C Data preparation for train-dev-test

We provide more configuration about data preparation for all models.

Model	Configuration	Split	Phrases	Mots CR	Mots EN	Tokens CR	Tokens EN
E1	baseline	train	665 033	8 477 485	7 839 120	14 947 629	13 648 756
		dev	646	9 077	8 286	14 924	13 451
		test	1 682	30 335	27 422	44 577	38 808
E2	Code switching	train	731 536	9 293 614	8 623 183	16 364 373	15 015 221
		dev	646	9 145	8 405	14 924	13 451
		test	1 682	30 335	27 422	44 577	38 808
E3	Code switching + Cultural Data	train	746 854	9 396 508	8 733 045	16 493 710	15 154 134
		dev	2 561	22 023	22 012	31 310	31 008
		test	1 682	30 335	27 422	44 577	38 808

Table 6: Data distribution for the *hat-eng* systems of the WMT26 Shared Task where E1: Kreyòl-MT, E2: Kreyòl-MT + Code Switching, and E3: Kreyòl-MT + Code Switching + Cultural Data