Findings of the First Shared Task for Creole Language Machine Translation at WMT25

Nathaniel R. Robinson[1], Claire Bizon Monroc[2], Rasul Dent[3], Stefan Watson[4], Kenton Murray[1], Raj Dabre[5], Andre Coy[4], Heather Lent[6]

[1]Johns Hopkins University, USA; [2]Mines Paris PSL, France;
[3]Inria Paris, France; [4]University of the West Indies at Mona, Jamaica;
[5]IIT Madras, India; [6]Aalborg University, Denmark

Abstract

Efforts towards better machine translation (MT) for Creole languages have historically been isolated, due to Creole languages' geographic and linguistic diversity. However, most speakers of Creole languages stand to benefit from improved MT for low-resource languages. To galvanize collaboration for Creole MT across the NLP community, we introduce the First Shared Task for Creole Language Machine Translation at WMT25. This Shared Task consists of two systems tracks and one data track, for which we received submissions from five participating teams. Participants experimented with a wide variety of systems and development techniques. Our evaluation campaign gave rise to improvements in MT performance in several languages, and particularly large improvements in new testing genres, though some participants found that reusing subsets of pretraining data for specialized post-training did not yield significant improvements. Our campaign also yielded new test sets for Mauritian Creole and a vast expansion of public training data for two Creole languages of Latin America.

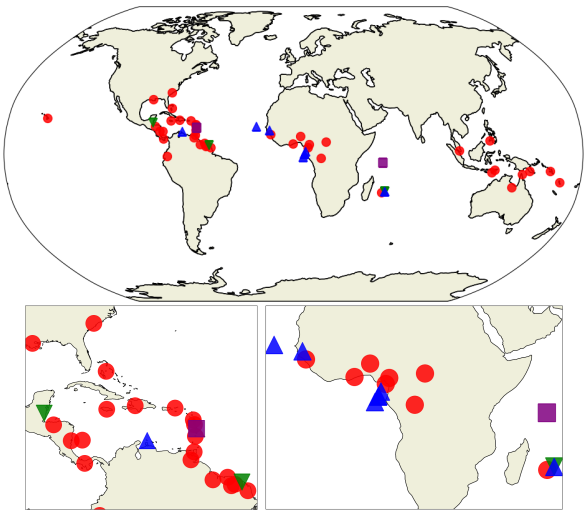

Figure 1: Creole languages included in the shared task, plotted geographically. Purple squares are languages for which we received *constrained* system submissions. Blue up-triangles are those with *unconstrained* system submissions, and green down-triangles are for *data* submissions. Red circles are for remaining Creoles for which we supported submissions but received none.

1 Introduction

Insufficient training data remains a pronounced barrier for creating natural language processing (NLP) systems that cater to lower-resourced languages. This is particularly pronounced for the task of machine translation (MT), due to the importance of aligned bitexts. For Creole languages, a geographically and linguistically diverse group (see Figure 1), the lack of training data brings some challenges common to other low-resource scenarios, but also offers unique opportunities, due to the role of contact in shaping them.

Many Creole-speaking communities have expressed interest in having MT support for their language (Lent et al., 2022). However, efforts to create NLP systems for them have largely been fragmented. This is in part due to the truly expansive scope of the term "Creole". By definition, a Creole language exhibits linguistic influence from an amalgamation of languages, typically both high- and low-resource (Kouwenberg and Singler, 2009). Naturally, the Creole languages of Africa (*e.g.* Nigerian Pidgin and Sango) are often viewed with different historical and cultural lenses than those of the Caribbean (*e.g.* Haitian Creole and Papiamento) or Pacific (*e.g.* Bislama and Tok Pisin).

However, this geographic and cultural fragmentation misses some of the notable commonalities among Creole languages. Many of the Creole languages of Africa and the Caribbean have a shared linguistic history: most of the languages in both groups emerged from European colonialism and slavery in Africa, and languages in both groups are strongly influenced by European and African

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 520–531
November 8–9, 2025.

(typically Niger-Congo) languages in similar ways (Gilman, 1986; Robinson et al., 2024). While Pacific Creole languages are not connected to Africa, they are also influenced by largely the same European colonial languages (*e.g.*, English, French, Portuguese, Dutch, *etc.*), and many show comparable and sometimes even more pronounced grammatical influence from Oceanic languages (Keesing, 1988). Perhaps even more important than their linguistic commonalities, Creole languages across the globe are subject to similar stigmas and play similar linguistic roles in relation to high-resource languages (DeGraff, 2003). Hence, speakers of Creole languages everywhere may have shared experiences in connection to their language and could benefit similarly from improved language technologies.

Only recent publications have addressed the data scarcity of large numbers of Creole languages, a necessary hurdle to create systems that truly serve speakers. (See Section 2.) These recent efforts have been ground-breaking (Lent et al., 2024; Robinson et al., 2024). However, their large scope allowed only for treatment of Creole languages as a conglomerate without particular attention to specific language communities and their distinct needs. To build technologies for the wide array of Creole language speakers, participation from a variety of communities is needed. In an effort to foster collaboration in NLP across these communities, we hold the first Shared Task for Creole Language MT held at WMT25.[1] We detail prior publications that inspired the need for this shared task in Section 2, and we overview the task organization in Section 3. The shared task consists of two subtasks: a call for **System Submissions**—itself consisting of two tracks, namely **Track 1** (Constrained), **Track 2** (Unconstrained)—and a call for **Dataset Submissions**. Five teams participated. Submissions and evaluation results are discussed in Sections 4 and 5, respectively.

2 Related Work

Creole languages are a product of intense linguistic contact. They usually draw most of their vocabulary from a single source language that is often referred to as the Creole language's lexifier.[2] Typologically, Creole languages tend to be more isolating than these lexifier relatives. However, attempts to define them as distinct typological class (*e.g.* Bakker et al., 2011) are contested (*e.g.* Fon Sing, 2017). In light of the typological debate, some (*e.g.* Mufwene, 2001) prefer to define Creoles by their history, which is closely associated with the long-term movement of people speaking mutually unintelligible languages.

Today, languages recognized as Creoles are spoken by over 180 million people (Lent et al., 2021). For some speakers, a Creole language is their mother tongue, and these speakers may be monolingual; for others, a Creole language can serve as a lingua franca within the broader, multilingual community. In surveying Creole language speakers, Lent et al. (2022) find that they highly desired MT support for their languages. Bird (2022) further highlights the opportunity for Creole language technology to serve as a bridge between higher and lower-resourced communities. He emphasizes that in contexts where Creoles act as a lingua franca, it may be more viable to develop language technologies for them, rather than going straight to even lower-resourced highly localized languages.

Despite the demand and utility of MT for Creole languages, they occupy a small portion of works in the broader MT community. When they are included, works tend to focus on individual languages. For example, the 2010 earthquake in Haiti prompted the rapid development of MT systems for the French-related Haitian language (Lewis, 2010). Early work by Dabre et al. (2014); Dabre and Sukhoo (2022) on Mauritian Creole acknowledged this trend, though they too developed technologies for a single language. Similarly, MT systems for West African Pidgin (Ogueji and Ahia, 2019) are highly relevant for other English-related Creole languages, like Guyanese Creolese (Clarke et al., 2024), but it has been difficult to coordinate multi-continental efforts.

The first effort towards a joint, large scale MT project for Creoles was proposed in CREOLEVAL (Lent et al., 2024) with the CREOLEM2M model, covering N-way translation between English and 26 Creoles. Building on this work, Robinson et al. (2024) contributed KREYÒL-MT for 41 Creole languages, including the first public, extensively multilingual MT dataset for Creole languages, and focusing on those spoken by the African diaspora. The resulting models have set the current state-of-the-art for Creole language MT.

Even with these advances, efforts to build Creole

[1] https://www2.statmt.org/wmt25/creole-mt.html

[2] This nomenclature is contested, however, as some academics argue it advances the misconception that Creole languages are fundamentally different from other languages (DeGraff, 2003).

technologies are not strongly unified. For instance, while Robinson et al. (2024) focus on Africa and the Americas, the potential relevance of these languages for MT efforts for Pacific Creoles (*e.g.*, Chavacano) (Vicente et al., 2024) remains an open question. Perhaps most pertinently, because both CREOLEVAL and KREYÒL-MT advanced developments for Creole languages as a conglomerate, focus on specific languages and their distinct needs was outside the scope of efforts. In organizing this shared task we hope to foster the beginnings of broader collaboration in Creole NLP. This way, members of specific Creole language communities may have a place to develop their own communities' desired technologies.

3 Shared Task Overview

3.1 Call for Systems Submissions

We solicited MT systems for translation between any number of Creole languages and English or French, for which we have paired data. (We also welcomed submissions for other pairs of a Creole and non-Creole language each, asking participants to justify how translation between such languages would be relevant to the affected community.)

The official train and dev sets provided to participants were the public KREYÒL-MT train and dev splits available on HuggingFace.[3] Participants were not permitted to use the data designated as *test* data from this dataset. We also provided, to any participants who requested them, additional training bitexts with English translations of Haitian, Papiamento, and Sango text from the Church of Jesus Christ of Latter-day Saints, still pending release on LDC.[4] We solicited submissions of systems in both **constrained** and **unconstrained** tracks. To construct our official evaluation sets for both tracks, we selected a random seed (kept private) and shuffled the public KREYÒL-MT test sets.

Constrained Track The purpose of this track was to explore better ways to model Creole MT with limited resources and allow researchers to explore smarter configurations than the simple ones that were used in past Creole language MT models. For instance, Robinson et al. (2024) focused primarily on their presentation of the KREYÒL-MT dataset without justifying all engineering choices used to train the KREYÒL-MT model. Thus, while their model was trained on 41 Creole languages at once, it was not clear whether this model was trained *optimally* for all 41 languages given the resources available. To have a track which is directly comparable to the original KREYÒL-MT model, therefore, we only accepted submissions in one of the 40 Creole languages included in the KREYÒL-MT set, with translation into or out of English and/or French only. The baseline model for this track was the `kreyol-mt-pubtrain` model available on HuggingFace,[5] which is the model trained on the public KREYÒL-MT dataset (Robinson et al., 2024). We note that this model was not trained on any data that was not made available to the participants. Moreover, participants were permitted to use this model however they wished in the constrained track, including as an initialization for fine-tuning, but they were not be permitted to use any other pre-trained models.

Unconstrained Track The purpose of this track was simply to encourage creation of state-of-the-art Creole language MT systems. In the unconstrained track, teams were allowed to use data from any source, and leverage any pre-trained models or LLMs. As the relaxed constraints in this track allowed participants to develop systems for any Creole language, we used two baseline models: `kreyol-mt-pubtrain` (the same as the constrained track)[6] and CREOLEM2M (Lent et al., 2024) (available on HuggingFace,[7] which covers a number of Creole languages not supported by the former[8]). That said, this track allowed for submission of any of the Creole languages supported by either of the baseline models; with translation into/out of English and/or French permitted for languages supported by KREYÒL-MT, and translation into/out of English only permitted for the languages supported by CREOLEM2M, accordingly. We used the same evaluation sets as in the constrained track for any languages supported by KREYÒL-MT, and were prepared to furnish additional eval sets for any other Creole languages.

[3] https://huggingface.co/datasets/jhu-clsp/kreyol-mt

[4] Further details: https://huggingface.co/datasets/jhu-clsp/kreyol-mt/blob/main/README.md

[5] https://huggingface.co/jhu-clsp/kreyol-mt-pubtrain

[6] Due to a miscommunication on our part, the EHOW team used `kreyol-mt`, the KREYÒL-MT model trained on both public and private data, as their baseline model throughout their experiments. Hence we used `kreyol-mt` as the baseline for their submissions.

[7] https://huggingface.co/AAU-NLP/CreoleVal-CreoleM2M

[8] https://github.com/hclent/CreoleVal/

For this unconstrained track, we were also prepared to accommodate submissions for Creole languages not supported by either baseline model, in which case we would require the participants to submit an evaluation set of their own creation, and we would employ the baseline models in a zero-shot setting to calculate baseline scores. However, this circumstance did not arise.

3.2 Call for Data Submissions

We also solicited contributions to Creole language MT training and evaluation sets. We requested data submissions to be in bitext formats with translations into *any* other language – not only English or French. (Again, we stipulated that submissions should justify why the non-Creole language would be relevant for the Creole language-speaking community).

We thus established the following requirements for any submitted datasets:

- Participants must show that 100% of translations were either translated or post-edited by competent native or proficient speakers of the source and target languages.
- Participants must prepare a data card[9] with each submitted dataset.
- Participants must be able to show that one of the languages in each submitted bitext is considered a Creole language, by citing adequate academic sources or other sufficiently convincing means.
- If submitting training data, we strongly encouraged participants to also develop an accompanying MT system and evaluate this system on a test set (either a test set of their own creation, which must be submitted along with the training data, or a previously published test set). Ideally, participants should show significant ($p < 0.05$) improvements in chrF++ (Popović, 2017) over the previous state-of-the-art open-source MT system for the given language pair. (To do this they must identify the previous SOTA model and make a compelling case for why it would be considered SOTA.) We committed to provide software to assist meeting this requirement as needed. If participants were not able to meet this requirement, we required that they provide other convincing evidence of the utility of their training set.
- If submitting a test set, participants must use it to evaluate performance of an MT model and provide compelling evidence that the model's performance on the test set aligns with conventional wisdom regarding the model's performance in the translation direction.

[9] See https://oldi.org/guidelines

3.3 Support for participants

During the course of the shared task, we aimed to support participants wherever possible. To this end, we provided tutoring, paper workshopping, and a dataset for manual translation (i.e. FLORES-200 English (NLLB Team et al., 2022)).

4 Shared Task Submissions

4.1 Constrained system submissions

We received two submissions for the constrained track, in which participants were only permitted to use provided training data: **KREY-ALL** (Ayasi, 2025) for Seychellois Creole translation into English, and **LUDOVIC MOMPELAT** (Mompelat, 2025) for translation between Martinican Creole and French.

4.1.1 KREY-ALL

KREY-ALL investigated joint training on typologically related Creole languages. They focused on Seychellois Creole (`crs`), and selected four additional languages known to be structurally similar, due to both shared francophone vocabulary and historical migration patterns: Mauritian Creole (`mfe`), French Guianese Creole (`gcr`), Louisiana Creole (`lou`) and Réunion Creole (`rcf`) (Papen, 1978).

Translation data from these languages were used in conjunction with Seychellois Creole data, with two tagging strategies: "All Kreyol" where all languages used the same language tag tag (`crs`), and "Specialized" where each language used its own tag. For each strategy, both full and partial (last 4-6 layers) fine-tuning were compared.

KREY-ALL found that using the same language tag for all languages and fine-tuning all model parameters were most effective. They also found that, although Mauritian Creole data was an order of magnitude more plentiful than Seychellois data, upsampling the Seychellois segments by more than 5x relative to the other languages was ineffective. Analysis of the model's embeddings revealed a distinct cluster for each language, with Seychellois having close proximity and overlap with Mauritian. However, there was no discernible Indian Ocean group, as Mauritan also overlapped with Louisiana

Creole, while Seychellois showed commonalities with Guianese. The other Indian Ocean language, Réunion Creole, did not overlap with any of the four.

4.1.2 LUDOVIC MOMPELAT

LUDOVIC MOMPELAT (LM) submitted two MT systems: from Martinican Creole (mart1259) to French and vice-versa. Their approach was to fine-tune kreyol-mt-pubtrain using LoRA (Hu et al., 2021). LM experimented with a different train/dev data ratio (70/30 vs. 90/10) and explored a few values of LoRA rank and scaling factor. Their final system for into-French MT used weighted BLEU scores for a curriculum sampling training set-up with difficulty ranking. Models for both translation directions were trained with label smoothing and a weighted BLEU criterion for checkpoint selection, and both employed a newly trained tokenizer with new language tags.

4.2 Unconstrained system submissions

Two teams submitted to the unconstrained track, in which participants were allowed to use external data and pre-trained models: **EDINHELSOW** (EHOW) (Rowe et al., 2025) and **KOZKREOLMRU** (Rajcoomar, 2025).

4.2.1 EDINHELSOW

EHOW submitted systems that translated between English and seven lusophone Creole languages: Angolar (aoa), Annobonese (fab), Guinea-Bissau Creole (pov), Kabuverdianu (kea), Papiamento (pap), Principense (pre) and Sãotomense (cri). The team conducted an incredibly thorough analysis of numerous techniques for enhancing Creole MT, and even leveraged the linguistic relationship between Portuguese and related Creoles.

Notably, EHOW collected additional parallel and monolingual data to supplement the provided data sources. These additional parallel data came from various sources: online Bible translations (pap and pov); the Jehovah's Witnesses' Watchtower magazine (kea, pap, and pov); text sourced from an online educational sentence generator (pap); and gloss text from a dictionary (pov). They used the monolingual corpora to create synthetic parallel data generated through back-translation, using the kreyol-mt model. For some of their experiments, the English sentences from the pap, kea, pov and cri bitexts were also forward-translated for Sequence-Level Distillation (Kim and Rush, 2016) of kreyol-mt. Again for some experiments, training data were further augmented using 112k high-quality English-Portuguese sentences, extracted from the Tatoeba Translation Challenge203 Dataset (Tiedemann, 2020).

Using these curated data sources, the EHOW team fine-tuned various pretrained multilingual base models: two sizes of NLLB (NLLB Team et al., 2022), three configurations of mBART (Tang et al., 2020), and kreyol-mt. EHOW experimented with inclusion of the Portuguese and distillation data mentioned in the previous paragraph, as well as with initializing language token embeddings with the Portuguese token embedding, to explore 14 combinations of training practice. They then merged six combinations of the resulting models to produce final systems. EHOW's *primary* submission for each language pair was the overall best performing merged model for each generalized direction (XX→eng or eng→XX), while *contrastive1* submissions were the best trained model (merged or otherwise) for each language pair. They found that post-editing of system outputs using LLMs and bilingual lexicons was typically not helpful, but they submitted some systems that incorporated this practice as *contrastive2*.

4.2.2 KOZKREOLMRU systems

KOZKREOLMRU took a unique three-step approach to translation between Mauritian Creole and English. Their first step was continuous pretraining of Llama 3.1-8B over 500k monolingual Mauritian Creole tokens (18k lines) sourced from (Dabre and Sukhoo, 2022), with an additional 100k monolingual tokens each of English and French data. Step two was then fine-tuning the model for MT. This was done on $\tilde{4}$0k lines of bitext, sourced again from Dabre and Sukhoo (2022); Robinson et al. (2024); 4.9k lines of synthetic data from prompting a Claude model with text from MMLU (Hendrycks et al.); and 300 lines of bitext from community translation of English Claude outputs. The final step was parameter-efficient fine-tuning (PEFT) via LoRA (Hu et al., 2021) over newly contributed Mauritian Creole translations of FLORES-200 (NLLB Team et al., 2022). In some eperiments, only the dev set translations were used for PEFT, while the devtest set was reserved for testing. In others, KOZKREOLMRU used all FLORES-200 data for PEFT and evaluated on a newly created bitext from the LALIT newspaper. (See Section 4.3 for details of these datasets.) The

KozKreolMRU team performed ablations, to isolate the effects of monolingual continuous pretraining, vanilla fine-tuning, and PEFT.

4.3 Data submissions

Two of our participating teams submitted datasets: **KozKreolMRU** (Rajcoomar, 2025) submitted two dev/test sets for Mauritian Creole↔English MT; and **JHU** (Robinson, 2025) submitted train, dev, and test sets for Belizean Kriol↔English and French Guianese Creole↔French MT. See data cards for these submissions on GitHub.[10]

4.3.1 KozKreolMRU data

KozKreolMRU submitted two dev/test sets to evaluate translation between Mauritian Creole and English. The first consists of Mauritian Creole translations of FLORES-200 (NLLB Team et al., 2022), containing 997 `dev` lines and 1012 `devtest` lines. This dataset is particularly useful because (1) it is automatically aligned with the other language sets contained in FLORES-200 and (2) FLORES is a common benchmark to judge MT model proficiency. The second dataset is a small test set consisting of 102 sentence pairs sourced from the LALIT newspaper.

4.3.2 JHU

JHU submitted three datasets for two language pairs. The first consists of 5.5k lines of Belizean Kriol and English translations from a Belizean textbook, both automatically and manually aligned after document processing. The second dataset comes from an online Bible translation and pairs 879 French Guianese Creole lines with French translations. The third dataset consists of 792 sentences from French Guianese Creole fables, aligned with French translations (mostly manually after web-scraping the raw data). All of these datasets were divided into train, dev, and test splits with a 90-10-10 ratio. Together they increase the amount of publicly available bitext by 2,300% for Belizean Kriol↔English and 370% for French Guianese Creole↔French. JHU demonstrated improvements ranging from +3.2 chrF++ to +33.3 chrF++ on the submitted test sets via fine-tuning `kreyol-mt-pubtrain` on the submitted train sets.

5 Evaluation Results

We designed a shared evaluation process for the two system tracks. All language pairs in the received submissions were supported by Kreyòl-MT, so only our shuffled Kreyòl-MT public test sets were used for official evaluation.[11]

For every language pair and direction, we computed the chrF++ (Popović, 2017) and BLEU (Papineni et al., 2002) scores of the submission using the default parameters of the `sacrebleu` library. We compared these results to the baseline scores of `kreyol-mt-pubtrain`, `kreyol-mt`, and CreoleM2M as stipulated in Section 3.1. Note that no two teams submitted a system for the same language pair, and therefore scores could only be compared with baselines. Results for the constrained track are reported in Table 1, and for the unconstrained track in Table 2. We discuss the results for each track separately below.

5.1 Constrained system results

direction	Team	Model	chrF++	BLEU
		creole → XX		
crs→eng	**Krey-All**	contrastive2	59.0	34.5
	Krey-All	contrastive1	58.9	34.2
	baseline	`kreyol-mt-pubtrain`	57.7	33.8
	Krey-All	primary	58.4	33.7
mart1259 →fra	*baseline*	`kreyol-mt-pubtrain`	50.4	28.3
	LM	primary	49.1	25.3
		XX → creole		
fra→ mart1259	*baseline*	`kreyol-mt-pubtrain`	48.7	26.5
	LM	primary	48.7	25.8

Table 1: Results of primary, *contrastive1*, and *contrastive2* submissions to the constrained track. Systems are ordered by BLEU score.

ChrF++ and BLEU scores are reported in Table 1 for both teams, alongside the `kreyol-mt-pubtrain` baseline. The LM systems score on par with or slightly below the baseline, while *constrative1* and *constrative2* submissions of Krey-All show marginal improvements.

The *primary* **Krey-All** submission corresponds to approaches with data all merged under the Seychellois Creole language tag (`crs`); *contrastive1* represents use of language-specific tags, and *constrative2* indicates partial parameter freezing (with the last 4 encoders, all decoder layers and

[10] https://github.com/n8rob/creolemt_wmt25

[11] This led to complications with the unconstrained systems track. Because Robinson et al. (2024) originally shuffled and split their private and public test sets independently, the `kreyol-mt` model trained on the private set has been contaminated with some of the segments in the public Kreyòl-MT test sets we used to construct our eval sets. Yet as a publicly available model, `kreyol-mt` was permitted for unconstrained submissions. This is what prompted our reevaluation of EHOW submissions; see Section 5.2.1.

the shared embeddings fixed during training). The best model for crs→eng was *constrative2*, however all four submitted systems score within one BLEU point of each other. This, combined with the observation that KREY-ALL came up with a different system ranking (one in which *constrative2* performed worst of all, (Ayasi, 2025)) by using a different shuffle of the same test set, suggests that score differences are not significant.

Recall from Section 4.1.2 that **LM**'s into-French system employed curriculum learning with BLEU-based difficulty ranking. The system for opposite translation direction did not employ this, but used different train/dev split (70-30 instead of 90-10). In our official evaluation, both submissions under-performed the baseline, albeit by less than 2.0 chrF++. LM's own reporting a slight improvement over the baseline for `mart1259→fra` on a shuffling of the same test set, again gives the impression that score differences are not significant.

Both LM and KREY-ALL fine-tuned the `kreyol-mt-pubtrain` model for a single language pair. These were valuable experiments because such attempts had not been made previously. The original work of (Robinson et al., 2024) focused primarily on the KREYÒL-MT dataset and did not include extensive experiments regarding how to engineer optimal MT systems from the available data. One question left open by their work was whether, if in the absence of additional data or models, significant improvements could be achieved by post-training on a select subset of training data. Results from these two studies (Ayasi, 2025; Mompelat, 2025) now indicate no such significant improvements, suggesting that either very different methods would be needed to improve performance, or that researchers may find more promise in developing new datasets and using external models (rather than post-training on subsets of `kreyol-mt-pubtrain`'s own data).

5.2 Unconstrained system results

ChrF++ and BLEU scores are reported in Table 2 for both teams, alongside baselines.[12] The **KOZKREOLMRU** systems are unique in that they are not based on any system that was trained on the full KREYÒL-MT dataset. Both KOZKREOLMRU systems out-performed CREOLEM2M for `mfe↔eng` MT, but both under-performed `kreyol-mt-pubtrain`, with the out-of-English direction performingly comparitively worse. (This is consistent with Rajcoomar's (2025) own finding that monolingual pre-training and two steps of fine-tuning has different effectiveness

direction	Team	Model	chrF++	BLEU
		creole → XX		
aoa→eng	**EHOW**	contrastive1	34.9	30.0
	EHOW	primary	19.3	17.6
	baseline	`kreyol-mt`	10.5	4.4
cri→eng	*baseline*	`kreyol-mt`	82.8	79.9
	EHOW	primary	82.0	78.2
fab→eng	**EHOW**	contrastive2	28.2	15.4
	EHOW	contrastive1	27.7	14.7
	EHOW	primary	12.0	1.2
	baseline	`kreyol-mt`	10.0	0.3
kea→eng	*baseline*	`kreyol-mt`	93.7	90.1
	EHOW	primary	93.4	90.0
mfe→eng	**KOZKREOL**	primary	46.7	25.6
	baseline	`kreyol-mt-pubtrain`	46.3	25.0
	baseline	CreoleM2M	33.5	12.8
pap→eng	**EHOW**	contrastive1	84.0	74.8
	EHOW	primary	76.3	64.8
	baseline	`kreyol-mt`	74.6	62.1
	baseline	CreoleM2M	56.7	37.1
pov→eng	*baseline*	`kreyol-mt`	87.7	82.8
	EHOW	contrastive1	81.0	74.0
	EHOW	primary	81.0	74.0
pre→eng	**EHOW**	contrastive2	56.6	40.3
	EHOW	contrastive1	55.3	40.5
	EHOW	primary	24.1	9.3
	baseline	`kreyol-mt`	9.9	0.3
		XX → creole		
eng→aoa	**EHOW**	contrastive2	33.2	24.4
	EHOW	contrastive1	33.0	23.5
	EHOW	primary	27.8	21.4
	baseline	`kreyol-mt`	8.7	12.4
eng→cri	*baseline*	`kreyol-mt`	80.2	76.5
	EHOW	contrastive1	78.1	73.6
	EHOW	primary	25.4	7.3
eng→fab	**EHOW**	contrastive2	26.0	5.1
	EHOW	contrastive1	25.5	7.7
	EHOW	primary	16.1	2.6
	baseline	`kreyol-mt`	6.6	0.8
eng→kea	*baseline*	`kreyol-mt`	91.4	87.5
	EHOW	contrastive1	90.1	85.5
	EHOW	primary	41.5	17.9
eng→mfe	*baseline*	`kreyol-mt-pubtrain`	49.7	28.7
	KOZKREOL	primary	43.1	18.6
	baseline	CreoleM2M	32.7	10.0
eng→pap	**EHOW**	contrastive1	76.7	62.2
	EHOW	primary	72.1	53.0
	baseline	`kreyol-mt`	65.6	48.8
	baseline	CreoleM2M	51.4	29.5
eng→pov	*baseline*	`kreyol-mt`	92.0	89.9
	EHOW	contrastive1	74.1	67.6
	EHOW	primary	31.7	12.3
eng→pre	**EHOW**	contrastive2	44.6	21.8
	EHOW	contrastive1	42.4	22.8
	EHOW	primary	26.4	5.4
	baseline	`kreyol-mt`	9.1	1.2

Table 2: Results of the **EHOW** and **KOZKREOLMRU** (abbreviated **KOZKREOL**) submissions for the unconstrained track. Systems are ordered by chrF++ score.

[12] Note that CREOLEM2M supports only two of the submitted language pairs: `mfe↔eng` and `pap↔eng`.

	mfe→eng		eng→mfe	
	FLORES	LALIT	FLORES	LALIT
baseline	57.3	50.8	49.1	46.2
primary	**67.7**	**70.2**	**57.7**	**68.9**

Table 3: Comparison of KOZKREOLMRU chrF scores with `kreyol-mt` on FLORES and LALIT test sets. Here *primary* is understood to refer to the version of the *primary* submitted systems that was not trained on FLORES `devtest` data, in the columns for FLORES. These are averaged sentence chrF scores across each set, consistent with Rajcoomar (2025). High scores are **bold**.

depending on language direction.) But despite its under-performance on the official evaluation, the KOZKREOLMRU systems significantly outperform `kreyol-mt-pubtrain` on the KOZKREOLMRU submitted test sets (by 8.0 chrF minimum). See Table 3, which compares chrF (Popović, 2015) scores of Rajcoomar's (2025) models with `kreyol-mt-pubtrain` performance on both FLORES and LALIT test sets.

Scores for **EHOW** are also in Table 2. As noted in Section 3.1, the EHOW team used the `kreyol-mt` model (trained on both public and private KREYÒL-MT data). For a fair comparison, we use this model as their baseline. However, since `kreyol-mt` was trained on portions of the public test sets for `cri`, `kea`, `pap`, `pre`, and `pov`, this results in inflation of scores for both EHOW systems and their baselines.

Recall from Section 4.2.1 that the EHOW *primary* and *contrastive1* submissions consisted mostly of merged models (the best model for each overall direction, and the best for each language pair, respectively). In cases where LLM post-editing improved on the *contrastive1* result, this was submitted as *contrastive2*. Scores tend to improve from *primary*→*contrastive1*→*contrastive2* for most language pairs. EHOW systems outperform the baseline in both translation directions for `pap`, `pre`, `fab`, and `aoa`. The baseline scores higher on `cri`, `kea`, and `pov`. However, note that, as mentioned in the previous paragraph, these results and the others for `cri`, `kea`, `pap`, `pre`, and `pov` may be obfuscated by dataset contamination, making it difficult to draw clear conclusions from EHOW scores. To address this issue, we set up a second round of evaluations for a subset of the EHOW submissions.

5.2.1 Reevaluation of EHOW systems

Table 4 contains the results for our accurate reevaluation of EHOW *primary* and *contrastive1* systems, avoiding data contamination. Avoiding this contamination was a challenge. The team fine-tuned some models from a `kreyol-mt` initialization, which was trained on some segments in the test set used to evaluate `kreyol-mt-pubtrain`. And for fine-tuning they used the set used to train `kreyol-mt-pubtrain`, which contains some overlap with the test set used to evaluate `kreyol-mt`. Noting that each KREYÒL-MT model's corresponding train and test sets had no overlap with each other, we decided to use the intersection of both `kreyol-mt` and `kreyol-mt-pubtrain` test sets to ensure we would not evaluate on any segments used in training. However, this intersection was incredibly small for some language pairs. Hence, we augmented any resulting test sets with fewer than 20 aligned sentences, by adding sentences that had originally been filtered out of the test sets during Robinson et al.'s (2024) test set cleaning processes. These extra segments were removed due to length or noise, so we cleaned them manually. In these decontaminated, augmented test sets, the smallest set was for `pov-eng` (with 23 aligned sentences), just as in the original test set (in which the set for this same language pair had 33 aligned sentences).

When we evaluate on decontaminated test sets, EHOW systems outperform the baseline in chrF++ for every language pair, except those involving `pap` and `pov`. See Table 4. (Note that we can conclude from Table 2 already that EHOW outperformed the baseline on `aoa` and `fab` directions, not shown in Table 4.)

It is worth mentioning here that the EHOW team found superior performance of their systems over the baseline for directions involving `pap` and `pov` when they used their own test sets—which better match the distribution of the additional training data the team curated and used for fine-tuning. See Table 5 for chrF (Popović, 2015) scores. The EHOW in-house test sets are not completely free from contamination, since they contain some of the synthetic data segments tha the team produced using the `kreyol-mt` model itself. However, this would ostensibly give the baseline mode an advantage, rather than the competitor models; and the EHOW test sets for `pap-eng` and `pov-eng` only contain 13% and 15% synthetic segments, respec-

direction	Model	chrF++	BLEU
	creole → eng		
cri→eng	primary	39.2	22.0
	kreyol-mt	37.2	22.6
kea→eng	primary	61.0	43.9
	kreyol-mt	56.5	36.1
pap→eng	kreyol-mt	67.8	54.2
	primary	66.6	52.6
	contrastive1	63.1	47.2
pov→eng	primary	51.0	41.8
	kreyol-mt	43.0	27.4
	contrastive1	40.4	19.4
pre→eng	contrastive1	59.7	54.5
	primary	26.3	9.9
	kreyol-mt	5.8	0.1
	eng → creole		
eng→cri	primary	35.7	24.3
	kreyol-mt	28.6	16.5
	contrastive1	27.6	12.2
eng→kea	contrastive1	50.0	27.8
	kreyol-mt	50.0	26.7
	primary	43.6	22.3
eng→pap	kreyol-mt	59.3	41.1
	contrastive1	58.4	41.4
	primary	46.8	27.2
eng→pov	kreyol-mt	29.3	8.3
	contrastive1	27.5	7.5
	primary	25.9	3.4
eng→pre	contrastive1	31.2	14.0
	primary	25.9	10.6
	kreyol-mt	9.0	1.3

Table 4: Results from reevaluating EHOW *primary* and *contrastive1* submissions on a decontaminated test set. Systems are ordered by chrF++ score.

tively. Hence we infer that the effects of this data contamination are minor, and can conclude with reasonable confidence that EHOW's own models for these language pairs would outperform the baseline model in the genres represented in the extra data they used.

Given all of this, we observe that that both unconstrained submissions (EHOW and KOZKREOLMRU) show a common pattern: employing new training datasets and different pre-trained models can expand Creole MT performance to new genres, even in cases when it does not significantly improve performance on pre-existing test sets or distributions.

6 Conclusions

The first shared task for Creole language MT convened submissions for a variety of Creole lan-

	XX→eng		eng→XX	
	pap	pov	pap	pov
baseline	39.5	29.8	38.8	20.1
primary	45.8	28.6	26.9	**44.2**
contrastive1	**67.6**	**46.2**	**49.5**	18.4

Table 5: Comparison of EHOW submitted models to the kreyol-mt baseline chrF on private EHOW test sets. Highest scores are **bold**.

guages situated across the Caribbean, South America, and Africa. This convergence was made possible by an important acknowledgment: despite their differences, Creole languages may benefit from a united approach in developing new language technology solutions. Indeed, Creole languages are the fruit of similar sociohistorical developments, leading to shared linguistic patterns, and also have in common a paucity of corpora and MT systems. We received and evaluated four submissions for new MT systems, and two dataset submissions, representing 12 Creole languages total.

We note several noteworthy observations from the contributions of the participants:

- **Creative data curation**: We observed a wide variety of approaches, including human translation, data augmentation, data up-sampling, back-translation, and synthetic data generation.
- **Harnessing linguistic information**: Submissions demonstrated the utility of linguistic considerations and relationships between languages (both between Creole languages with shared history and with relative languages).
- **Data-conscious methods**: Similarly, participants leveraged an assortment of algorithmic approaches to overcome data scarcity, including LoRa PEFT, partial freezing of layers, and model merging.
- **Adressal of directional challenges**: Participants noted that translation into a high-resource language tended to yield better results than translation into a Creole language.

Our evaluation led to some central lessons and takeaways. It was found that constrained systems, which attempted to boost MT performance for a particular language by post-training the kreyol-mt-pubtrain model on a subset of its own training data (pertaining to the language in question), did not result in significant performance improvements, even when authors searched across other training tactics and hyperparameters to maxi-

mize performance. This suggests that developing new datasets and using pre-trained models may be a more promising direction. Accordingly, our participants in the unconstrained systems track showed that such methods are often effective at improving results for some language pairs, and that even in cases where performance on the original test domain does not improve, new datasets and models can bring about expansion to new testing genres.

Takeaways and Future Ambitions The variety of training approaches developed and evaluated for this shared task provides valuable insights into the training of Creole MT systems. New data collected during the campaign can also be incorporated into future Creole MT datasets and models. In particular, this will allow us to train a new baseline model for the next iteration of the shared task. Re-training the baseline model will also give us the chance to ensure that all updated versions of KREYÒL-MT models online can be evaluated with the same test sets without contamination (a critical point for interpreting results).

We are proud that a number of this year's shared task organizers represent various Creole language-speaking communities. Our committee includes two L1 speakers of Jamaican Patois, one L1 speaker of Martinican Creole, one L2 speaker of Louisiana Creole, and one L2 speaker of Haitian. In the future, we hope to incorporate members from a broader diversity of Creole language communities, so that our efforts better serve the realistic needs of these communities.

Limitations

As noted throughout this paper, our primary limitation stemmed from issues with data contamination between the train and test sets for `kreyol-mt` and `kreyol-mt-pubtrain models`. This was an organizational failure on our part that will be rectified in future iterations of this work. A limitation not so easily overcome is simply that of genre homogeneity in the datasets for low-resource languages. As demonstrated by both unconstrained track participants, it was much easier to out-perform pre-trained baseline models on novel test sets than on existing test sets, likely due to correlations of genre and topic between training data and testing data. Though this is a significant limitation, it is one of the very problems that this shared task is intended to rectify. (Due to this year's progress, we now have more diverse datasets for Mauritian Creole, Belizean Kriol, and French Guianese Creole.)

Ethical Statement

Given the historical and ongoing marginalization of many Creole languages and their population of speakers (DeGraff, 2003), we stress that community engagement is crucial. To ensure resulting research in machine translation is in accordance with community wants and needs (Lent et al., 2022), with the goal of preserving community autonomy (Bird, 2020). That said, one common limitation of working in the low-resource space is over-reliance on religious domain data; we acknowledge the presence of data which may not be culturally relevant to Creole language speakers (Hershcovich et al., 2022; Mager et al., 2023).

Acknowledgments

This work received funding from the Inria "Défi"-type project COLaF. We would like to thank the WMT organizers for making a place for Creole languages and for the opportunity to host a shared task. We also thank all of the teams who submitted systems and datasets.

References

Ananya Ayasi. 2025. Krey-All WMT 2025 CreoleMT System Description: Language Agnostic Strategies for Low-Resource Translation. In *Proceedings of the Tenth Conference on Machine Translation*.

Peter Bakker, Aymeric Daval-Markussen, Mikael Parkvall, and Ingo Plag. 2011. Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole Languages*, 26(1):5–42.

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.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Christopher Clarke, Roland Daynauth, Jason Mars, Charlene Wilkinson, and Hubert Devonish. 2024. GuyLingo: The Republic of Guyana Creole Corpora. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 792–798, Mexico

City, Mexico. Association for Computational Linguistics.

Raj Dabre and Aneerav Sukhoo. 2022. Kreol-MorisienMT: A dataset for mauritian creole machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 22–29, Online only. Association for Computational Linguistics.

Raj Dabre, Aneerav Sukhoo, and Pushpak Bhattacharyya. 2014. Anou Tradir: Experiences In Building Statistical Machine Translation Systems For Mauritian Languages – Creole, English, French. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 82–88, Goa, India. NLP Association of India.

Michel DeGraff. 2003. Against creole exceptionalism. *Language*, 79(2):391–410.

Guillaume Fon Sing. 2017. Creoles are not typologically distinct from non-Creoles. *Language Ecology*, 1(1):44–74.

Charles Gilman. 1986. African Areal Characteristics: Sprachbund, not Substrate? *Journal of Pidgin and Creole Languages*, 1(1):33–50.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Roger M. Keesing. 1988. *Melanesian Pidgin and the Oceanic Substrate*. Stanford University Press.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1317–1327.

Silvia Kouwenberg and John Victor Singler. 2009. *The handbook of pidgin and creole studies*. John Wiley & Sons.

Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. 2021. On language models for creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 58–71, Online. Association for Computational Linguistics.

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 Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, and 2 others. 2024. CreoleVal: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

William Lewis. 2010. Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, St Raphael, France.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics.

Ludovic Mompelat. 2025. Ludovic mompelat wmt 2025 creolemt systems description : Martinican creole and french. In *Proceedings of the Tenth Conference on Machine Translation*.

Salikoko S. Mufwene. 2001. Creolization is a social, not a structural, process. In Ingrid Neumann-Holzschuh and Edgar W. Schneider, editors, *Degrees of Restructuring in Creole Languages*, Creole Language Library, page 65. John Benjamins Publishing Company.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Kelechi Ogueji and Orevaoghene Ahia. 2019. Pidginunmt: Unsupervised neural machine translation from west african pidgin to english. *ArXiv*, abs/1912.03444.

Robert Antoine Papen. 1978. *The French-based Creoles of the Indian Ocean: an Analysis and Comparison.* University of California, San Diego.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Hemkeshsing Y. Rajcoomar. 2025. KozKreolMRU WMT 2025 CreoleMT System Description: Koz Kreol: Multi-Stage Training English-Mauritian Creole MT. In *Proceedings of the Tenth Conference on Machine Translation*.

Nathaniel Robinson. 2025. JHU WMT 2025 CreoleMT System Description: Data for Belizean Kriol and French Guianese Creole MT. In *Proceedings of the Tenth Conference on Machine Translation*.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Stutzman, Bismarck Odoom, Sanjeev Khudanpur, Stephen Richardson, and Kenton Murray. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and colonial African creole languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.

Jacqueline Rowe, Ona de Gibert, Mateusz Klimaszewski, Coleman Haley, Alexandra Birch, and Yves Scherrer. 2025. EdinHelsOW WMT 2025 CreoleMT System Description: Improving Lusophone Creole Translation through Data Augmentation. In *Proceedings of the Tenth Conference on Machine Translation*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Jörg Tiedemann. 2020. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*.

Aileen Joan Vicente, Theresse Faith Amamampang, Dunn Dexter Lahaylahay, and Charibeth Cheng. 2024. ChavacanoMT: A Corpus and Evaluation of Neural Machine Translation for Philippine Creole Spanish.