JHU WMT 2025 CreoleMT System Description: Data for Belizean Kriol and French Guianese Creole MT

Nathaniel R. Robinson
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD, USA
nrobin38@jhu.edu

Abstract

This document details the Johns Hopkins University's submission to the 2025 WMT Shared Task for Creole Language Translation. We submitted exclusively to the data subtask, contributing machine translation bitext corpora for Belizean Kriol with English translations, and French Guianese Creole with French translations. These datasets contain 5,530 and 1,671 parallel lines of text, respectively, thus amounting to an 2,300% increase in publicly available lines of bitext for Belizean Creole with English, and an 370% such increase for French Guianese Creole with French. Experiments demonstrate genre-dependent improvements on our proposed test sets when the relevant state-of-the-art model is fine-tuned on our proposed train sets, with improvements across genres of up to 33.3 chrF++.

1 Introduction

The vast majority of countries and territories throughout Central and South American are hispanophone. Two notable exceptions to this trend are anglophone Belize and francophone French Guiana, pictured in Figure 1. These are both home to a duplicity of Creole languages that exist alongside English, French, and other regional languages. Many thousands of Belize residents speak Belizean Kriol and Garifuna. And French Guiana is home to French Guianese Creole, Saramaccan, and Ndyuka (Robinson et al., 2024).

According to the Statistical Institute of Belize' 2022 Population and Housing Census, Belizean Creole is spoken by 181k people in Belize (or 49% of the overall population), making it the third-most-spoken language of the country, after English (at 278k, or 75.5% of the population) and Spanish (at 199k, or 54%).

French Guiana is home to 292k people[1] and a large diversity of both Creole languages (French Guianese Creole, Haitian, French Antillean Creole, Ndyuka, Saramaccan, Sranan Tongo, and Guyanese Creolese); European languages (French, Spanish, English, Dutch, and Portuguese); Amerindian languages (Palikur, Teko, Wayampi, Wayana, and Arawak); and East Asian languages (Hmong, Cantonese, Hakka Chinese) (Léglise, 2013). According to Léglise,[2] French Guianese Creole is the mother tongue of about a third of the population. Ndyuka, an English-related Creole language, and its dialects Aluku and Paramaccan (Hammarström et al., 2023) are also spoken by roughly a third of the population. Multiple of the languages Léglise listed are spoken by immigrant communities: Haitian by 10-20% of the population, French as a mother tongue by roughly 10%, Brazilian Portuguese by 5-10%, and French Antillean Creole languages by roughly 5%. Léglise estimates that less than 5% speak indigenous Amerindian languages.

Figure 1: The greater Caribbean area with Belize and French Guiana indicated in red, made with https://www.mapchart.net

Because English and French are prominent in Belize and French Guiana, respectively, and because these European languages are used by populations that do not speak Belizean Kriol or French Guianese Creole in Belize and French Guiana, we justify the choice of these languages for translation

[1] Per the French Institut national de la statistique et des études économiques

[2] See https://leglise.cnrs.fr/?lang=en.

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 1191–1197

into and out of our Creole languages of focus.

In this work, we contribute:

- A bitext with 5,530 Belizean Kriol and English translations from a Belizean textbook, *Kriol-Inglish Dikshineri/English-Kriol Dictionary* (Herrera et al., 2009)
- A bitext with 879 French Guianese Creole and French translations from a web-sourced Bible text
- A bitext with 792 French Guianese Creole and French translations from a collection of French Guianese fables sourced by the French national library
- State-of-the-art MT improvements across the genres of our newly curated datasets, per our own test sets

2 Related Work

Very little previous work has been published on machine translation (MT) of Belizean Kriol and French Guianese Creole. CreoleVal (Lent et al., 2024), a project introducing a multilingual machine translation model for 28 Creole languages, was the first published work on Belizean Kriol machine translation. Like in this work, Lent et al. focused on translation between Belizean Kriol and English. They put together a Belizean Kriol bitext with English, including 12,085 lines of training data. Using these data, they trained the CreoleM2M model,[3] which acheives 44.4 chrF (Popović, 2015) for Belizean Kriol to English translation and 46.3 chrF for English to Belizean Kriol (on their own test set). However, none of the CreoleVal data were publicly released to the research community.

The other publication to benchmark Belizean Kriol translation was Kreyòl-MT (Robinson et al., 2024), a project that produced a dataset and model for translation of 41 Creole languages, primarily of the Colonial African diaspora. The Kreyòl-MT model[4] supports translation in 172 language directions—all of which include one Creole language and one other language, usually English or French. The Kreyòl-MT model achieves 53.3 chrF on the Kreyòl-MT test set for Belizean Kriol to English translation, and 83.3 chrF for English to Belizean Kriol (on their own private test set). The same model achieves very similar scores on the CreoleVal (Lent et al., 2024) test set: 53.3 chrF into English, and 83.5 into Belizean Kriol, establishing Kreyòl-MT as state-of-the-art for Belizean Kriol translation with English. The dataset used to develop Kreyòl-MT contains 31,002 total lines of bitext for Belizean Kriol aligned with any other language, but only 229 of these lines (all aligned with English translations) were released to the research community in the Kreyòl-MT public dataset.[5]

Kreyòl-MT is also the only published work to address French Guianese Creole MT. On its own test set, Kreyòl-MT achieves 71.9 chrF translating French Guianese Creole into French and 62.4 translating French into French Guianese Creole. This was accomplished using only 292 lines of bitext for training data, indicating that these high scores may be a result of over-fitting to the training domain. This entire bitext was released publicly, along with 50 lines of bitext for testing and 50 for development/validation.

3 Task Description

The first shared task for Creole language machine translation (CreoleMT) for the Tenth Conference on Machine Translation (WMT25) (Robinson et al., 2025) was announced to further advance the state of machine translation for Creole languages. The shared task is comprised of two subtasks: one for systems and one for data. We participate only in the data subtask, which solicits new bitext data for Creole language MT.

As part of the set up for the systems subtask of the shared task, the task organizers established the public Kreyòl-MT dataset as the primary source of training data for the constrained submission track. To make the evaluation fair, they named the Kreyòl-MT **pubtrain** model—an mBART (Liu et al., 2020) initialization developed only with the publicly released Kreyòl-MT train and development sets and hosted at https://huggingface.co/jhu-clsp/kreyol-mt-pubtrain with the name **kreyol-mt-pubtrain**—as the baseline model for benchmarking. (The flagship Kreyòl-MT model was trained on some proprietary data not available for public release and hence would not be a fair baseline.) Since the datasets we curate in this work are publicly releasable, we compare their utility to that of other publicly available data and also adopt **kreyol-mt-pubtrain** as our primary base-

[3] https://huggingface.co/prajdabre/CreoleM2M

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

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

line model for MT experiments.

4 Dataset Creation

We detail our creation approaches for all datasets.

4.1 Belizean Kriol dictionary

Our Belizean Kriol-English bitext is taken entirely from aligned sentences in *Kriol-Inglish Dikshineri/English-Kriol Dictionary* (Herrera et al., 2009). The sentences were extracted from a PDF rendering of the book using the PyPDF2 library[6] and software that will be released before publication of this work.

The scraping script used regular expressions to extract all complete sentences used in examples with translations. With the assumption that sentences would alternate between Belizean Kriol and English, we used this software to label every odd sentence as "English" and every even sentence as "Belizean Kriol." As expected, this method of labeling led to a large number of alignment errors.

We corrected these alignment errors via a semi-automated approach. First we trained a decision tree classifier using `scikit-learn`[7] for language identification (LID), to distinguish Belizean Kriol and English sentences. We first used both source and target sides of the Kreyòl-MT public train set for Belizean Kriol-English MT to fit a simple word-based sentence embedder. For 100-dimensional embeddings, the embedder assembled the 100 most common words in the combined source-target corpus. Each sentence is then assigned a binary 100-dimensional vector where the element at each position k indicates the presence or absence of the kth most common word. We then used the same data (the Kreyòl-MT Belizean Kriol-English train bitext) to fit the decision tree, embedding each sentence with our embedder. We used the corresponding Kreyòl-MT validation/development set to evaluate our decision tree, which guided our decision to use 300-dimensional vectors instead of 100-dimensional, and our decision to use a decision tree rather than logistic regression or a random forest. (The decision tree fit with 100-dimensional vectors achieved the highest validation accuracy.)

Next we set up a system where one sentence collection (odd sentences) and the other (even sentences) are initially labeled arbitrarily as "English" and "Belizean Kriol," respectively. Human intervention is requested only whenever the LID classifier does not label both sentences as their expected languages, during which intervention the human can make any adjustments needed, including switching which language corresponds to which sentence set by default.

By this process we achieved a fully aligned bitext of 5,530 lines. Since the previous largest publicly available bitext for Belizean Kriol-English tranaslation was size 240 (Robinson et al., 2024), this amounts to a 2,304% (or 24-fold) increase in publicly available data for the language pair.

The human guiding the alignment process, while not a speaker of Belizean Kriol, speaks English natively and has high familiarity and professional experience with Creole languages. This rendered the alignment guidance easy, and we believe its quality to be high. Table 1 displays a perfectly random sampling of five aligned sentences from the bitext. Note that even non-speakers of Belizean Kriol can verify the validity of these alignments, due to the linguistic proximity of these languages.

4.2 French Guianese YouVersion Bible

The French Guianese Creole biblical data we extracted from the online YouVersion Bible[8] using the BeautifulSoup Python library.[9] The site includes one French Guianese Creole translation of the Gospel of John.[10] We scraped this and aligned it with the corresponding texts in a French Bible translation, the New Geneva Edition, on the same site,[11] resulting in 879 aligned sentences. Since the previous largest public translation dataset for French Guianese Creole-French translation contained 447 aligned sentences (Robinson et al., 2024), this amounts to a 197% increase in publicly available data.

Similar to Table 1, Table 2 displays a perfectly random sampling of five aligned sentences from the bitext.

4.3 French Guianese Gallica fables

We sourced our final bitext from the collection *Introduction à l'histoire de Cayenne: suivie d'un recueil de contes, fables et chansons en créole avec traduction en regard, notes et commentaires* (de Saint-Quentin and de Saint-Quentin, 1872).

[6] https://pypdf2.readthedocs.io/en/3.x
[7] https://scikit-learn.org
[8] https://www.bible.com
[9] https://tedboy.github.io/bs4_doc
[10] https://www.bible.com/bible/2963/JHN.1.GCR07
[11] https://www.bible.com/bible/106/JHN.1.NEG79

English	Belizean Kriol
By the time John was twelve, he was already tall like his dad.	Bai di taim Jan twelv, ih mi don taal laik ih pa.
I'm missing my watch from off the bureau; somebody must have stolen it.	Ah di misn mi wach aaf a mi byooro; sohnbadi mos a teef it.
The woman helped the thieves to escape.	Di uman mi help di teef dehn fi eskayp.
The man took a long time to cut the yard because his machete was dull.	Di man tek lang fi chap di yaad, kaa ih masheet mi dol.
We went to Betty's birthday party last night.	Wi mi gaahn da Beti bertday paati laas nait.

Table 1: Five randomly sampled aligned English and Belizean Kriol translations

French	French Guianese Creole
C'est ici le pain qui descend du ciel, afin que celui qui en mange ne meure point.	Mé dipen ki désann di syèl-a, sala ka manjé li péké mouri.
Le Père aime le Fils, et il a remis toutes choses entre ses mains.	Papa Bondjé kontan so Pitit, é li bay li tout pouvwè ansou tout bagaj.
Pourquoi m'interroges-tu? Interroge sur ce que je leur ai dit ceux qui m'ont entendu; voici, ceux-là savent ce que j'ai dit.	Poukisa to ka kèksyoné mo? Kèksyoné moun-yan ki kouté mo-a, sa-ya, yé byen savé sa mo di yé."
Si je n'étais pas venu et que je ne leur aie point parlé, ils n'auraient pas de péché; mais maintenant ils n'ont aucune excuse de leur péché.	Si mo pa té vini é si mo pa té palé pou yé, yé pa té ké koupab di péché. Mé anprézan, yé pa ganyen pyès èskiz pou yé péché.
Jésus leur répondit: J'ai fait une œuvre, et vous en êtes tous étonnés.	Jézi réponn yé: "Mo fè roun sèl kichoz, é zòt èstébékwé.

Table 2: Five randomly sampled aligned French and French Guianese Creole biblical translations

This digitized book contains French Guianese fables written in French Guianese Creole with French translations. The Gallica website[12] from which we sourced this text contained eight such fables, organized into a total of 48 sections. Once again using BeautifulSoup for web-scraping, we extracted the sentences and aligned the text sections automatically. Sentence alignment was then performed manually, along with a substantial amount of error correction; many of the text segments on the web-page had what appeared to be OCR-induced noise.[13] As a disclaimer, the author who performed this manual alignment and cleaning is not proficient in French Guianese Creole. However, he is proficient in both French and another French-related Creole language (Haitian), and he has both linguistic training and experience in bitext development. Sentence alignment and cleaning decisions were mostly made clear by cognates and context. The changes to the originally web-scraped document are publicly available for perusal on GitHub.[14]

Similar to Tables 1 and 2, Table 3 displays five randomly sampled aligned and cleaned sentence pairs from the fables dataset.

5 Dataset Details

Our datasets consist of one genre each, educational material in Belizean Kriol and English, and Biblical text and literature in French Guianese Creole and French. Both datasets we shuffled randomly and then split into train, dev, and test sets with an 80-10-10 split ratio. See the shared task findings (Robinson et al., 2025) for data cards of our submitted datasets.

5.1 Fulfillment of task requirements

We now detail how our datasets fulfill each of the data requirements set forth for the shared task.[15]

1. **Translations in datasets must be completely conducted by native or proficient speakers of both languages.** The translations in our Belizean Kriol-English bitext come from *Kriol-Inglish Dikshineri/English-Kriol Dictionary* (Herrera et al., 2009), a publication authorized by the Belize Ministry of Education. The French Guianese Creole-French sentences come from Bible translations and fable

[12] https://gallica.bnf.fr/ark:/12148/bpt6k82939m/texteBrut, accessed August 2025

[13] "Typos" caused by errors in an optical character recognition system that may have digitized the text from an original print source

[14] https://github.com/n8rob/creolemt_wmt25_jhu_submission/pull/1/files

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

French	French Guianese Creole
Il resta ainsi longtemps,	Li rété bon moso konsa
Si vous voulez vous laver dedans,	Si zôt oulé lavé landan,
Il en restait encore une assez grande quantité;	Bon moso té rété enko
sans pouvoir en prendre un seul pour mon souper.	san mo pa pouvé kienbé oun pou mo sonpé.
J'en retirerai un peu avec mes dents »	M'a tire moso ké mo dan »

Table 3: Five randomly sampled cleaned and aligned French and French Guianese Creole fable translations

translations organized by the Bibliothèque nationale de France. The reputability of these sources convince us that such translations were performed by competent translators.

2. **Participants must demonstrate convincingly that one language in each submitted bitext is considered a Creole language.** Both Belizean Kriol and French Guianese Creole are universally considered Creole languages as attested by their inclusion in APICS (Michaelis et al., 2013), among other sources (McWhorter, 2005; DeCamp, 1968).

3. **If submitting a test set, participants must use it to evaluate performance of an MT model and provide compelling evidence that performance aligns with conventional wisdom.** We include results to address this point, for all three of our submitted datasets, in §6.

6 Experimental Results

Here we detail experiments that minimally show the utility of our datasets, as requested by the shared task organizers. We show both SpBLEU (Papineni et al., 2002; Goyal et al., 2022) and chrF++ scores (Popović, 2015, 2017) for the **`kreyol-mt-pubtrain`** model both out of the box, and after fine-tuning on our train sets (with our dev sets used to determine time of early stopping). We trained with a learning rate of $2 * 10^{-5}$, a batch size of 4, a weight decay of 0.01, and a maximum of 10 training epochs. We implemented early stopping with a stopping patience of 2 and a stopping threshold of 0.01. We fine-tuned each model on both translation directions for the language pair in question, and with early stopping we ended up finetuning for Belizean Kriol-English translation for 13,272 training steps, and for French Guianese Creole-French translation for 4,683 training steps.

	eval	score	OOB	FT
bzj→eng	*Dict.*	SpBLEU	17.0	**52.1**
		chrF++	37.9	**66.8**
eng→bzj	*Dict.*	SpBLEU	8.46	**48.1**
		chrF++	24.8	**58.1**
gcr→fra	*Bible*	SpBLEU	18.1	**35.3**
		chrF++	39.1	**51.7**
	Fable	SpBLEU	36.6	**41.3**
		chrF++	51.5	**54.7**
fra→gcr	*Bible*	SpBLEU	6.86	**25.4**
		chrF++	20.5	**34.8**
	Fable	SpBLEU	34.0	**37.9**
		chrF++	44.8	**48.6**

Table 4: MT scores for **`kreyol-mt-pubtrain`** out-of-the-box (OOB) and fine-tuned on our novel train sets (FT). Both SpBLEU and chrF++ scores are computed on our proposed test sets. Better results are **bold**.

Table 4 displays our MT performance scores. Fine-tuning on our training data significantly improves performance on our test sets in every circumstance. Some improvements are large, but their magnitude depends on test set language and genre. Out of the Creole language and into the Creole language, respectively, chrF++ improves by 28.9 and 33.3 on the Belizean dictionary test set, by 12.6 and 14.3 on the Guianese Bible test set, and by 3.2 and 3.8 on the Guianese fables test set. Note that Table 4 uses ISO 639-3 codes to abbreviate language names.[16]

7 Conclusion

We introduce three new datasets for Creole language translation of two language pairs:

1. **A bitext with 5,530 aligned translations of the educational genre in Belizean Kriol and English**, resulting in a 2,304% increase

[16] `bzj` = Belizean Kriol; `eng` = English; `gcr` = French Guianese Creole; `fra` = French

of publicly available bitext data for the language pair, and improvements of 28.9 chrF++ into-English and 33.3 chrF++ out-of-English over the state-of-the-art MT model trained on publicly available Kreyòl-MT data on our in-genre test set

2. **Two bitexts with combined 1,671 aligned translations in French Guianese Creole and French: 879 biblical sentences and 792 lines of Guianese folktales**, resulting in a 370% increase of publicly available bitext data for the language pair; improvements of 12.6 chrF++ into-French and 14.3 chrF++ out-of-French over the state-of-the-art MT model trained with publicly available data on our Biblical test set, and like improvements of 3.2 into-French and 3.8 out-of-French on our fables test set

We hope these incremental contributions may forward research in MT for Creole languages.

Limitations

Note that due to genre constraints, the utility of these datasets may be restricted to certain domains. For instance, the French Guianese Creole-French bitext we curated may be useful for assistance in further Bible translation, but may have limited utility beyond that.

Additionally, our compiling of the contributed datasets was constrained by time. We hope in future shared tasks to contribute even more data as we have the time to collect it. We also note that even in conducting manual cleaning and alignment of the datsets we submit here, cleaning and alignment errors can still be found in some places of the datasets. Though it may not be possible for these low-resource datasets to be completely noise-free, we hope future shared tasks will afford us the time to make extra passes and clean more thoroughly.

References

Alfred de Saint-Quentin and Auguste de Saint-Quentin. 1872. *Introduction à l'histoire de Cayenne: suivie d'un recueil de contes, fables et chansons en créole avec traduction en regard, notes et commentaires*. J. Marchand.

David DeCamp. 1968. The field of creole language studies. *Latin American Research Review*, 3(3):25–46.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Harald Hammarström, Sebastian Nordhoff, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. Glottolog. Online database.

Yvette Herrera, Myrna Manzanares, Silvaana Udz, Cynthia Crosbie, and Ken Decker. 2009. Kriol-inglish dikshineri/english-kriol dictionary. *Belize Kriol Project.–Belmopan, Belize*, 465.

Isabelle Léglise. 2013. *Multilinguisme, variation, contact. Des pratiques langagières sur le terrain à l'analyse de corpus hétérogènes*. Accreditation to supervise research, Institut National des Langues et Civilisations Orientales- INALCO PARIS - LANGUES O'.

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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

John H McWhorter. 2005. *Defining creole*. Oxford University Press.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013. *APiCS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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.

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.

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*.