Kreyòl-MT: Building MT for Latin American, Caribbean and Colonial African Creole Languages

Nathaniel R. Robinson[1] **Raj Dabre**[3] **Ammon Shurtz**[2] **Rasul Dent**[4]
Onenamiyi Onesi[5] **Claire Bizon Monroc**[4] **Loïc Grobol**[6] **Hasan Muhammad**[1]
Ashi Garg[1] **Naome A. Etori**[7] **Vijay Murari Tiyyala**[1] **Olanrewaju Samuel**[8]
Matthew Dean Stutzman[2] **Bismarck Bamfo Odoom**[1] **Sanjeev Khudanpur**[1]
Stephen D. Richardson[2] **Kenton Murray**[1]
[1]Johns Hopkins University, USA; [2]Brigham Young University, USA;
[3]National Institute of Information and Communications Technology, Japan;
[4]Inria Paris; [5]Nile University of Nigeria; [6]Université Paris Nanterre;
[7]University of Minnesota - Twin Cities, USA; [8]University of Toronto
nrobin38@jhu.edu

Abstract

A majority of language technologies are tailored for a small number of high-resource languages, while relatively many low-resource languages are neglected. One such group, Creole languages, have long been marginalized in academic study, though their speakers could benefit from machine translation (MT). These languages are predominantly used in much of Latin America, Africa and the Caribbean. We present the largest cumulative dataset to date for Creole language MT, including 14.5 M unique Creole sentences with parallel translations—11.6 M of which we release publicly, and the largest bitexts gathered to date for 41 languages—the first ever for 21. In addition, we provide MT models supporting all 41 Creole languages in 172 translation directions. Given our diverse dataset, we produce a model for Creole language MT exposed to more genre diversity then ever before, which outperforms a genre-specific Creole MT model on its own benchmark for 23 of 34 translation directions.

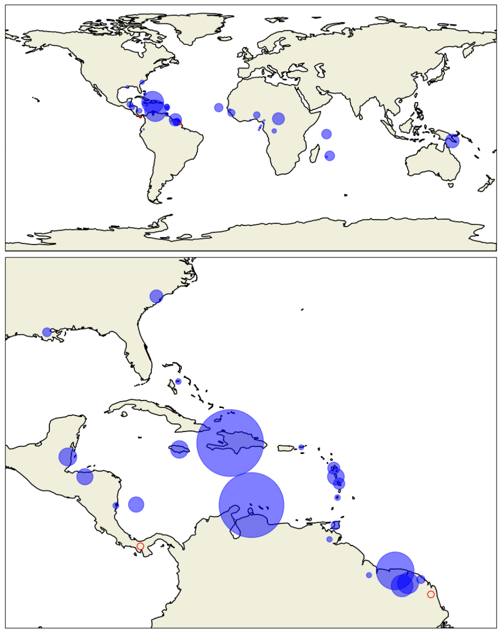

Figure 1: Dataset sizes plotted geographically, with centroids from Glottolog. Each circle's area is proportionate to the square root of the data amount for each language, to facilitate viewing.

1 Motivation

From northern Brazil to the Gulf of Mexico, spanning an area including the Caribbean and Central American west coast, lies the Creole "civilizational region" (Glissant, 2008). One of its chief characteristics: a multiplicity of Creole languages, born from contact of African language speakers with European languages in the colonial era. Low-resource Creole languages are widely spoken here and throughout the world (Rickford and McWhorter, 2017; Mufwene, 2008; Bartens, 2021; Velupillai, 2015). Historic linguistic marginalization has stymied their technological advancement: few language technologies exist for these languages despite their many speakers (Lent et al., 2023).

Better MT could greatly benefit Creole language speakers. Many live in areas where their language is in the minority. Panama and Costa Rica are home to communities of West Indian descent who have maintained Creole languages (Conniff, 1983; Herzfeld, 1980). Large Haitian-speaking communities live in the Dominican Republic (Zhong, 2023), Chile, Mexico (Audebert, 2017), Brazil (Terry and Mayes, 2019), and the Bahamas (Perry, 2023; McCartney, 2013; Knowles, 2018). Language is one

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
June 16-21, 2024

of the first obstacles to immigrants' social integration, and many report daily reliance on MT (Neto et al., 2020). The lack or low accuracy of such technologies can thus contribute to social exclusion.

Multiple Creole-speaking communities regularly fall victim to natural disasters (Heinzelman and Waters, 2010; Margesson and Taft-Morales, 2010; Rasmussen et al., 2015; Look et al., 2019). As the frequency of Atlantic hurricanes may be accelerated by global climate change (Hosseini et al., 2018), machine translation can provide useful tools to facilitate communication during international relief efforts (Lewis, 2010; Hunt et al., 2019).

Yet colonial-era stigmas dismissing Creole languages as broken or incomplete persist, and serve as justifications to advantage European languages at their expense (Alleyne, 1971; DeGraff, 2003). Association with lower economic status and limited use in official settings then inhibit data collection and Natural Language Processing (NLP) development for these languages (Lent et al., 2023). We expand on Lent et al. (2023, 2022) to unify community efforts in advancing Creole NLP.

In addition to meeting community needs, Creole language MT presents avenues of exploration for low-resource NLP. Many Creole languages have documented linguistic relationships with high-resource languages, as well as lexical and morphosyntactic proximity to each other (Rickford and McWhorter, 2017). (See § 3.) As such, they have potential for cross-lingual transfer (Lent et al., 2023), a powerful technique for low-resource NLP (Pfeiffer et al., 2020; Kim et al., 2019). This potential presents an opportunity to develop technologies for many Creole languages at once. But since state-of-the-art NLP methods rely on machine learning, this development is not possible without data. We present an expansive dataset for MT of Creole languages, as a meaningful first step towards developing their technologies. We contribute:

- The largest, most genre-diverse MT datasets ever compiled for 41 Creole languages, including the first ever for 21 Creole languages
- A public dataset of 11.6M aligned sentences and 3.4M monolingual sentences for 40 Creole languages[1]
- Public models achieving state-of-the-art performance on a published Creole language benchmark for 23 language directions

2 Background and Related Work

Despite its potential benefits, previous Creole language MT development has been scarce. Google Translate,[2] a common MT interface, only supports one Latin American Creole language (Haitian). NLLB-200 (NLLB Team et al., 2022), a state-of-the-art MT model in the number of languages it supports (204), only supports five Creole languages. The emergence of large language models like ChatGPT,[3] trained on massive unlabeled datasets, presents encouraging potential for more universal MT support. However, recent studies indicate that LLMs are unable to compete with supervised MT for low-resource languages (Robinson et al., 2023a; Zhu et al., 2023). Researchers have also speculated that LLMs' MT capabilities in high-resource languages are due to curated bitexts in their training data (Briakou et al., 2022). Hence, developing bitext corpora for low-resource languages is still important in building MT for them.

Some prior works have approached Creole language MT, like the inclusion of early Haitian-English MT in the DIPLOMAT system (Frederking et al., 1997) and the 2011 Workshop on Statistical Machine Translation (WMT) (Callison-Burch et al., 2011). Creole languages that are the focus of more recent MT research include Haitian and Jamaican (Robinson et al., 2022a, 2023b), Naija (Adelani et al., 2022a; Ogueji and Ahia, 2019), Mauritian (Dabre and Sukhoo, 2022a), Sranan Tongo (Zwennicker and Stap, 2022), and Singlish (Wang et al., 2017; Liu et al., 2022). Additional NLP research has been conducted in sentiment analysis and named entity recognition for Naija (Oyewusi et al., 2020; Muhammad et al., 2022, 2023; Adelani et al., 2021); syntactic analysis for Singlish (Wang et al., 2017), Naija (Caron et al., 2019), and Martinican (Mompelat et al., 2022); and inference for Jamaican (Armstrong et al., 2022).

These works, while valuable steps forward, have been limited in their scope of languages. The prior work most comparable to ours is the recent CreoleVal (Lent et al., 2023), which focused on building datasets for 28 Creole languages. They focused on machine comprehension, relation classification, and MT. Lent et al.'s (2023) MT dataset is a significant contribution to our work. They do not publicly

[1] Visit our repository https://github.com/JHU-CLSP/Kreyol-MT for software, models, and data download.

[2] https://translate.google.com

[3] chat.openai.com

release their MT datasets, however. While portions of the data we collected (including part of the CreoleVal set) are not publicly releasable, we release 11.6M aligned bitext sentences and 3.4M monolingual sentences in 40 languages. The genre coverage of CreoleVal's solely religious and Bible text data is another limitation shared by prior works, which can preclude general-purpose MT. For example, Robinson et al.'s (2023b) Haitian model trained on primarily religious texts achieves an impressive BLEU score of 68.0 on its same-genre test set, but a score of 14.7 on a Wikipedia-style test set. This indicates that Creole language MT models trained on specific genres may not be applicable to general domains. We provide the most diverse Creole language MT data yet, both in terms of languages and genres. (See § 3.3.)

Other multilingual MT works include some Creole languages. In addition to NLLB Team et al.'s (2022) five included Creole languages, Yuan et al.'s (2023a) multilingual Lego-MT model supports 8 Creole languages (out of 433 total). Our work is a significant expansion on this, making a meaningful contribution to the broader effort of low-resource dataset curation—including the WMT'23 shared task on the topic (Sloto et al., 2023), which our work follows by involving noisy data filtering (Minh-Cong et al., 2023) and document alignment (Steingrímsson, 2023).

3 Methodology and Dataset

The languages we include in our study are in Table 1. We retrieve each language's vitality, official recognized status by a political entity (*Off.*), and number of speakers from Ethnologue.[4] We label a language as a majority language (*Maj.*) if it is spoken by a majority of the population in one of its native countries or territories. This project includes languages from 24 Latin American and Caribbean countries and territories. (See Figure 1.)

As mentioned in § 2, Lent et al.'s (2023) CreoleVal work is most comparable to our own. But the set of languages we include differs from theirs. Their focus was on Creole languages in general, while ours is narrowed to those of the Americas, allowing us to dive deeper and include more languages with greater linguistic commonalities. The languages we include have shared patterns of historical formation. They generally have morphosyntactic commonalities with Niger-Congo languages (Kouwenberg and Lacharité, 2004; Castillo and Faraclas, 2006) and large-scale lexical overlap with Romance and Germanic languages (Valdman, 2000; Winford, 1997). For example, Haitian and Jamaican have extensively borrowed vocabulary from older forms of French and English, respectively, but they are morphosyntactically closer to Gbe, Kwa, and Igboid languages (Lefebvre, 2011; Brousseau, 2011; Seguin, 2020; Mufwene, 2002; Kouwenberg, 2008; Farquharson, 2012). Because we restrict our focus based on these linguistic traits, we also include some African Creole languages that are close phylogenetic relatives to our American focus languages, with some linguists arguing for a common ancestor (McWhorter, 2000).[5] For instance, Sierra Leonne's Krio is likely related to Maroon Creole languages like Ndyuka and Saramaccan (Bhatt and Plag, 2012). Linguists have long noted the linguistic proximity of Louisiana Creole and French Guianese Creole to Creole languages of the Indian ocean (Mauritian, Seychellois, and Réunion Creole) (Papen, 1978). Papiamento is considered a descendant of Kabuverdianu by some (Romero, 2010), and Jamaican has the same phylogenetic relatives as Ghanaian Pidgin (Amoako, 1992; Cassidy, 1966).

3.1 Collection methods

We divide our dataset collection methodology into two stages: *gathering* and *extraction*. 25 Creole languages were selected for an active *gathering* effort (underlined in Table 1). We excluded Haitian because its data was abundant in already identified sources. Further data was found for 17 other languages, also included in Table 1. For each of the 25 *gathering* languages, we performed the following steps. **First**, we searched research databases using query templates to track down already curated datasets. We searched "`[language name]`" on the ACL Anthology,[6] followed by "`[language name]` machine translation", "`[language name]` NLP", and "`[language name]` translation" on Google Scholar.[7] Query results from these search engines were typically prohibitively many, so we browsed top results until it became clear that the remainder

[4] https://www.ethnologue.com

[5] We also include three distinct control languages: Sango and Kituba (African Creole languages with morphosyntactic proximity to Niger-Congo languages but no lexical proximity to European languages), and Tok Pisin (an Oceanic Creole language with lexical proximity to a Germanic language—English—but no relation to Niger-Congo languages.)

[6] https://aclanthology.org

[7] https://scholar.google.com

					Speakers / k			
Language	**ISO**	**Glottocode**	**Native to...**	**Vitality**	*L1*	*L2*	**Off.**	**Maj.**
Saint Lucian Patois	**acf**	sain1246	Saint Lucia	Stable	760	-	✗	✓
Bahamian Creole	**bah**	baha1260	Bahamas	Stable	340	-	✓	✓
Berbice Dutch	**brc**	berb1259	Guyana	Extinct	0	0	✗	✗
Belizean Kriol	**bzj**	beli1260	Belize	Institutional	170	-	✗	✓
Miskito Coast Creole	**bzk**	nica1252	Nicaragua	Institutional	18	-	✗	✗
Garifuna	**cab**	gari1256	Central America	Endangered	120	-	✗	✗
Negerhollands	**dcr**	nege1244	U.S. Virgin Islands	Extinct	0	0	✗	✗
Ndyuka	**djk**	ndyu1242	Suriname, French Guiana	Stable	68	-	✗	✗
Guadeloupean Creole	**gcf**	**guad1243**	Guadeloupe	Stable	580	-	✗	✓
Martinican Creole	gcf	**mart1259**	Martinique	Stable	520	-	✗	✓
French Guianese Creole	**gcr**	guia1246	French Guiana	Stable	180	-	✗	✓
Gullah	**gul**	gull1241	South Carolina, Georgia	Endangered	250	-	✗	✗
Creolese	**gyn**	creo1235	Guyana	Stable	720	-	✓	✓
Haitian	**hat**	hait1244	Haiti	Institutional	13 000	69	✓	✓
San Andrés-Providencia	**icr**	sana1297	Colombia	Stable	12	-	✗	✗
Jamaican Patois	**jam**	jama1262	Jamaica	Stable	3100	3.7	✗	✓
Karipúna	**kmv**	kari1301	Brazil	Endangered	2.4	-	✗	✗
Louisiana Creole	**lou**	loui1240	Louisiana	Endangered	4.8	-	✗	✗
Media Lengua	**mue**	medi1245	Ecuador	Endangered	2.6	-	✗	✗
Papiamento	**pap**	papi1253	Aruba, Curaçao, Bonaire	Institutional	350	20	✓	✓
San Miguel Creole	**scf**	sanm1305	Panama	Extinct	0	0	✗	✗
Saramaccan	**srm**	sara1340	Suriname, French Guiana	Stable	35	-	✗	✗
Sranan Tongo	**srn**	sran1240	Suriname	Institutional	520	150	✗	✓
Vincentian Creole	**svc**	vinc1243	Saint Vincent	Stable	110	-	✗	✓
Trinidadian Creole	**trf**	trin1276	Trinidad	Stable	1000	-	✗	✓
Angolar	**aoa**	ango1258	São Tomé and Príncipe	Endangered	12	-	✗	✗
Saotomense	**cri**	saot1239	São Tomé and Príncipe	Endangered	56	-	✗	✗
Seychellois Creole	**crs**	sese1246	Seychelles	Institutional	88	-	✓	✓
Annobonese	**fab**	fada1250	Equatorial Guinea	Stable	6.6	-	✗	✗
Fanakalo	**fng**	fana1235	South Africa	Endangered	0	5.1	✗	✗
Pichi	**fpe**	fern1234	Equatorial Guinea	Institutional	15	190	✗	✗
Ghanaian Pidgin	**gpe**	ghan1244	Ghana	Institutional	2.0	5000	✗	✗
Kabuverdianu	**kea**	kabu1256	Cape Verde	Institutional	1200	14	✗	✓
Krio	**kri**	krio1253	Sierra Leonne	Institutional	820	7400	✗	✓
Kituba	**ktu**	kitu1246	Central Africa	Institutional	12 000	800	✓	✓
Mauritian	**mfe**	mori1278	Mauritius	Institutional	1000	6.5	✗	✓
Naija	**pcm**	nige1257	Nigeria	Institutional	4700	120 000	✗	✓
Guinea-Bissau Creole	**pov**	uppe1455	Guinea-Bissau	Institutional	340	1500	✗	✓
Principense	**pre**	prin1242	São Tomé and Príncipe	Endangered	0.2	-	✗	✗
Réunion Creole	**rcf**	reun1238	Réunion	Stable	810	-	✗	✓
Sango	**sag**	sang1328	Central African Rep.	Institutional	620	4600	✓	✓
Tok Pisin	**tpi**	tokp1240	Papua New Guinea	Institutional	130	4000	✓	✓
Cameroonian Pidgin	**wes**	came1254	Cameroon	Institutional	12 000	-	✗	✓

Table 1: Above are Creole languages of the Americas (Latin America, the Caribbean, and surrounding area). Below are Creole languages of Africa, as well as Tok Pisin. We refer to languages by their **bolded** language codes. Underlined languages are those on which we focused for data gathering, outlined in § 3. The last two columns indicate respectively whether the language has official status and whether it is a majority language in one of its native countries/territories.

Bitext					Monolingual		
	Web		*PDF*				
Prev. pub.	*aligned*	*articles*	*aligned*	*other*	*Prev. pub.*	*Web*	*PDF*
14196475	21963	216756	18614	4683	2767602	607657	27081

Table 2: Number of segments gathered from each source type/extraction method.

were no longer relevant. (Individual data gatherers used best judgment for each language.) For some languages we multiplied queries to accommodate for alternate language names. (See Table 8 in Appendix A) **Second**, for each language, we checked each of the following databases for parallel or monolingual corpora: OPUS (Tiedemann, 2012), Oscar (Ortiz Suárez et al., 2019) and LDC.[8] **Third**, we scoured the web for books with translations or additional resources. **Fourth**, we contacted researchers in the languages' speaking communities for leads to potential data sources. Though we attempted various methods of contact whenever possible, response rates were low for this step. In all we gathered 107 sources such as anthology websites promoting cultural heritage or language revitalization, educational materials, and government documents.

After completing *gathering*, we moved to *extraction*. At this stage, we divided each of the gathered resources into groups, based on the data format. The six groups were: (1) parallel data previously published as a bitext, (2) web sources with aligned parallel sentences, (3) web sources containing unsegmented articles of text with translations, (4) PDF sources with aligned parallel sentences, (5) other PDF sources, and (6) sources of monolingual data. The amount of segments for each of these groups is summarized in Table 2, with a breakdown per language in Table 7 (Appendix A).

We immediately consolidated the previously published bitexts, including single-language resources from LAFAND-MT (Adelani et al., 2022a), KreolMorisienMT (Dabre and Sukhoo, 2022a) and the Caribe Trinidadian-English dataset (Smith, 2022). We gathered monolingual data from the MADLAD-400 clean corpus (Kudugunta et al., 2024) and JamPatoisNLI (Armstrong et al., 2022). To create novel bitexts, we then used the BeautifulSoup[9] and Selenium[10] Python packages to extract text from web sources and the PyPDF2[11] package for PDF sources. Organizing bitexts from sources with aligned parallel sentences was generally straightforward. When only unsegmented articles with translations were available, we segmented and aligned text based on punctuation and then manually corrected errors. Manual correction was performed by data extractors with sufficient proficiency in the languages involved. Our diverse team has some proficient speakers in our target languages (one L1 Martinican Creole speaker, two L2 Haitian speakers, one L2 Guadeloupean Creole speaker, one L2 Louisiana Creole speaker, two L2 Naija speakers, and one L2 Ghanaian Pidgin speaker). Though this did not cover anywhere near all of our included languages, proximity between Creole languages and related languages made the task doable.[12] In general, the data extraction details varied for each source, since PDFs and websites have a wide variety of individualized styles and formats. In all, we extracted data from 39 of the 107 sources from our *gathering* stage.[13] Attributions for all data sources are in Appendix D.

As an appendage to our methodology, we incorporate data from published multilingual resources, including all bitexts from CreoleVal (Lent et al., 2023), Lego-MT (Yuan et al., 2023b), FLORES-200 dev and NLLB train data (NLLB Team et al., 2022), and AfricaNLP'23 (Robinson et al., 2023b) for any of our focus languages.[14] We retrieved Bible translations from JHU (McCarthy et al., 2020a) for 18 languages, with up to four unique translations for each. We selected the four fullest

[8] https://www.ldc.upenn.edu

[9] https://www.crummy.com/software/BeautifulSoup/

[10] https://www.selenium.dev

[11] https://https://github.com/py-pdf/pypdf

[12] For instance, one Louisiana Creole speaker and one Haitian speaker—both also proficient in English, French, and linguistics—were able to manually correct alignment for Seychellois-English bitexts, due to Seychellois' lexical overlap with Haitian, Louisiana Creole, and French.

[13] The remainder contained data in less accessible formats that we deemed too tedious to include in this work. We intend our dataset to be a living resource and hope to continually add to it as we encounter and format more data for Creole languages.

[14] We ensure not to include any data labeled for testing in our own data for training or tuning.

English and French Bibles with reasonably modern text style from the same corpus, and formed up to eight bitexts for each language. We also contribute 360K previously unreleased unique parallel sentences from the Church of Jesus Christ of Latter-day Saints for hat, pap, and tpi with English. This data comes from religious sources, including scripture, instruction, discourses, humanitarian resources, genealogy, and administrative documents. We scraped small parallel corpora in the educational domain from APICS (Michaelis et al., 2013) for 39 languages. We aggregated further published bitexts by crawling the web via Python's mtdata package, which points to data from multiple online sources. Last, we added monolingual Wikipedia dumps for jam, gcr, gpe, and srn.

3.2 Data amounts

Table 3 compares our own data size with the largest previously collected dataset for each language. Ours is the first dataset for 20 languages, and the first public dataset for 9 more. (See Table 5 for a comparison with individual prior works.)

Our datasets are, to our knowledge, the largest ever collected for each of these Creole languages, (1) because ours contain a conglomerate of the compared previously disparate sets, and (2) because of the additional data previously inaccessible to MT that we gathered in our *extraction* process. These latter novel data make up 262 k of the bitext sentences in our dataset. Figure 1 illustrates the size of each language's dataset, with circles at the coordinates for each language from Glottolog (Nordhoff and Hammarström, 2011) (zoomed in on Latin America). Filled blue markers have area corresponding to the square root of the bitext size. Hollow red markers indicate languages for which we found no bitext data (scf, kmv).

3.3 Diversity of genres

As mentioned in § 2, prior works in Creole language MT are severely restricted in terms of genre, prohibiting them from general-purpose MT. We mitigate this by including a diversity of genres. Figure 2 shows that more than half of languages (26/41) cover at least two genres.[15] The "Other/Mix" genre dominates charts for Haitian, Papiamento, and Total, because of the large NLLB train sets (sets we do not even include in our model training, see § 4). We include a final chart indicating the total with "Other/Mix" excluded, showing no majority among other genres. We acknowledge that our dataset's genre diversity can still be improved in future iterations, but we highlight the dramatic increase in diversity it offers, compared to previous work.

[15]Table 6 in Appendix A contains the same information in colorblind-friendly numerical form.

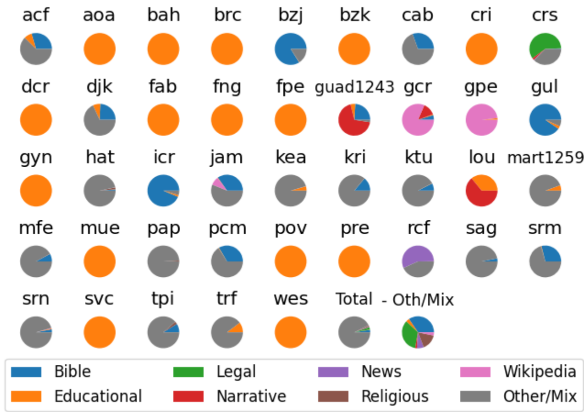

Figure 2: Genre proportions of each language's data (bitext and monolingual). We exclude the "Other/Mix" genre in the final pie to filter out large NLLB sets and show no majority among other genres.

4 Modeling Experiments

We now describe the experimental setup to show the utility of our datasets by training NMT systems. We train and release models on our full datasets, but we also train on only the data we publicly release, to show its stand-alone utility. We experiment with cleaning our train sets versus leaving them as-are, and in some experiments we fine-tune mBART (Liu et al., 2020) rather than training from scratch.

4.1 Dataset Preprocessing

Two collections: We maintain two primary data collections: one with all available data, henceforth called **all**; and a subset with all publicly releasable data, henceforth called **public**. To ensure the comparability of our models with previous work, we do not modify already available test splits. Moreover, for our experiments, we ensure that there is no overlap between these pre-existing test sets and our own data, by removing from our train/dev splits every sentence pair where either the source or the target is present in a pre-existing test set.[16]

[16]This is made necessary by the variety of sources we used but results in a loss of less than 1 % of the overall sentence pairs and allows for comparisons with the state-of-the-art that are as fair as possible.

	Max. prev.	Ours (pub. / all)		Max. prev.	Ours (pub. / all)		Max. prev.	Ours (pub. / all)
acf	15989	4406 / 23916	**gcf**	96	6467 / 6467	**mue**	-	147 / 147
aoa	-	198 / 198	**gcr**	-	1433 / 1433	**pap**	4898029	4968965 / 5363394
bah	-	327 / 327	**gpe**	-	223 / 223	**pcm**	31128	8084 / 47455
brc	-	222 / 222	**gul**	7990	266 / 8831	**pov**	-	480 / 480
bzj	23406	229 / 31002	**gyn**	-	258 / 258	**pre**	-	243 / 243
bzk	-	391 / 391	**hat**	4256455	5715227 / 6023034	**rcf**	-	285 / 285
cab	20879	- / 20879	**icr**	15702	317 / 16774	**sag**	262334	260560 / 535310
cri	-	306 / 306	**jam**	25206	434 / 28713	**srm**	42303	440 / 59053
crs	222613	3186 / 225875	**kea**	129449	132931 / 132931	**srn**	583830	6620 / 615010
dcr	-	189 / 189	**kri**	50438	185 / 66736	**svc**	-	321 / 321
djk	45361	15266 / 68833	**ktu**	7886	175 / 10737	**tpi**	424626	451758 / 925648
fab	-	204 / 204	**lou**	-	1860 / 1860	**trf**	-	1691 / 1691
fng	-	160 / 160	**mart1259**	-	5153 / 5153	**wes**	-	223 / 223
fpe	-	259 / 259	**mfe**	191909	25633 / 233320			

Table 3: Size of largest quality bitext data collected for Creole languages to date, compared with our full bitext sets and its public subset. Bitext size is measured as the number of unique Creole language sentences paired with a translation in any target language

Evaluation splits: After filtering, we prepare a train/dev/test split for each language pair of the remaining data by aggregating all sentences and splitting randomly with a fixed random seed, and a target ratio of 85 % / 5 % / 10 %, with minimum 50 and maximum 2000 sentences for the dev and test sets. We discard train sets for which less than 100 sentences are present. We exclude NLLB training sets for `hat`, `kea`, `pap`, `sag`, and `tpi` due to their overbearing size and observed poor quality, reducing our train set by 10.6M parallel sentences to a size of 450K. We still conduct zero-shot evaluations for language pairs for which training data was removed (indicated with an '*' in Figure 3).

Cleaning: Each dev and test set was cleaned according to the GILT Leaders Forum's Best Practices in Translation Memory Management.[17] We removed segments that were: empty, containing unbalanced brackets, mostly non-alphabetic characters, containing a source the same as the target, fewer than 3 words, and containing a higher number of characters than 5 standard deviations above the mean for data of that language. We normalized escaped characters and entities, white spaces, quotation marks, and the overall character set for each language. We removed any spurious characters that do not contribute semantically or syntactically in a segment and remove duplicates for all segments after cleaning to ensure there is no development and test set contamination. We toggle whether we thus clean train sets in both **public** and **all**, leading to a total of four configurations: **public**, **public-cleaned**, **all**, **all-cleaned**. Note that the dev and test sets, and hence models and their results, between **public** and **all** are not comparable, but those among cleaned and non-cleaned versions are (since the former variable implies independently split test sets, but latter variable toggles only whether *train* data were cleaned).

[17]We release cleaning software on our repository `https://github.com/JHU-CLSP/Kreyol-MT`.

4.2 Implementation and Training

We train models with YANMTT (Dabre et al., 2023), a toolkit that supports multilingual training and fine-tuning of mBART-50 (Tang et al., 2021) (minute model details in Appendix B.3). We first train from **scratch**, using all train data to fit a multilingual SentencePiece tokenizer (Kudo and Richardson, 2018) of 64 k subwords for all languages. Given our four experimental configurations, we train four such models. Next, we fine-tune many-to-many **mBART**-50 (Tang et al., 2021), repurposing the language indicator tokens for our Creole languages for simplicity. To keep a manageable computation budget for the compute-hungry mBART models, we only train them on the **clean** data configurations. We decode translations with beam size 4 and length penalty 0.4.

5 Results and Discussion

Figure 3 shows the performance of all models, tested for all available language pairs. We use chrF (Popović, 2015) because previous studies (Mathur et al., 2020; Rei et al., 2020) show that it correlates better with human judgments than BLEU (Papineni et al., 2002), and because semantic metrics such as COMET (Rei et al., 2020) and BLEURT

Figure 3: chrF scores on our newly created test sets using models trained on the **all** (top) and **public** (bottom) splits of our datasets. Given X-Y pair, → and ← represent the X to Y and Y to X translation, respectively. Zero-shot pairs are marked with an '*' sign.

(Sellam et al., 2020) lack support for low-resource languages.[18] Because many MT researchers have a more intuitive grasp of BLEU than chrF, we also include BLEU scores in Figure 5 of Appendix C.

As expected, high-resource pairs exhibited better translation quality than their lower resource counterparts, with some deviations from this trend possibly due to inevitable non-uniformity in test set genre. Some pairs like Trinidadian-English exhibited chrF up to 84, and some zero-shot pairs like Annobonese-English scored as low as 0.

Despite the generally lower performance on their languages, we highlight the potential value of our dataset's smallest bitexts. As stated in § 1, Creole languages' linguistic relationships open the possibility of powerful cross-lingual transfer. Previous studies (Ernštreits et al., 2022; Dabre et al., 2020; Arivazhagan et al., 2019; NLLB Team et al., 2022) have shown that MT models trained on numerous languages can often be adapted to translate new languages with few examples. Our results corroborate this finding; despite having only 391 parallel sentences to split for Miskito Coast Creole (bzk), our models achieved BLEU and chrF up to 43.7 and 60.4, respectively, translating into English. Vincentian Creole (svc) performed even better: up to 55.5 BLEU and 66.9 chrF into English, despite only 321 parallel sentences total. (Such scores would likely not be attainable via traditional bilingual training on such small sets (Arivazhagan et al., 2019).) Strikingly, we perceived even higher performance on zero-shot Martinican-to-English translation, with 69.9 BLEU and 84.9 chrF. We hope future studies will reveal more about our dataset's potential for low-resource cross-lingual transfer, and how it interfaces with dataset genre and diversity. This is an important exploration, not only to engineer better low-resource language systems, but as a scientific inquiry with general implications for low-resource language technologies.

Impact of corpora cleaning: In Figure 3, regardless of the use of **public** or **all** data, models trained on cleaned data (triangles and stars) typically outperform their counterparts trained on non-cleaned data (circles and diamonds), for higher resource languages. Cleaning eliminates noise and reduces variability in the especially noisy Creole datasets, which helps translation quality. However, cleaning can hurt performance by reducing the already scarce data for the lower-resource, more noisy corpora (like Pichi-English with 259 sentence pairs).

[18] Other recent low-resource MT works also prioritize chrF (Robinson et al., 2023a; Dabre and Sukhoo, 2022b).

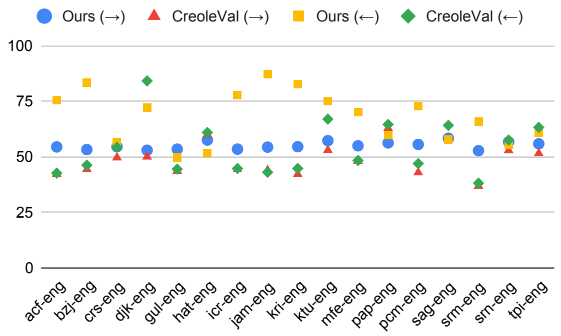

Figure 4: chrF for our model by fine-tuning **mBART** on **all cleaned** compared with CreoleVal. Given X-Y pair, → and ← represent the X to Y and Y to X translation, respectively.

Does fine-tuning help? From Figure 3, a comparison of **scratch** and **mBART** models trained on cleaned data (squares and x's) reveals that, while models trained from scratch exhibit strong results, fine-tuning on mBART is better for both **public** and **all** data use. These fine-tuned models scored up to 8 chrF better than **scratch** models. **Scratch** models converged after 500K training steps of 33k-token batches, and **mBART** fine-tuning needed 500K-600K steps of only 8.2k-token batches. While this implies that **mBART** is data efficient, **mBART** fine-tuning itself is slow due to training about 700% more parameters. Indeed, when comparing the total training time, **mBART** fine-tuning needed up to a week whereas **scratch** training needed only up to two days on the same hardware.

Comparison on existing test sets: Figure 4 compares our best models against results reported by CreoleVal (Lent et al., 2023) on relevant language pairs from the CreoleVal test sets. Since CreoleVal data is not public, we consider it apt to evaluate our best model (**mBART**) fine-tuned on **all cleaned** data. Overall, our models surpass CreoleVal models with +6.4 average chrF X→ENG and +14.1 average chrF ENG→X. Our dataset is much larger and more domain diverse than CreoleVal's; hence our improved results on most language pairs show that increasing data tends to be beneficial even if it reduces the domain specificity. We note that on a few language directions (9 of 34), CreoleVal's model still beat ours, indicating the possibility for negative interference from a more diverse train set in some instances. In Appendix C we provide some additional comparisons of bilingual versus multilingual fine-tuning and evaluations on other existing benchmarks: Lego-MT (Yuan et al., 2023a) and KreolMorisienMT (Dabre and Sukhoo, 2022a).

6 Conclusion

In this work, we compile the most comprehensive dataset to date for MT of Creole languages in the Americas. By aggregating disparate previous works and incorporating new data sources via scraping the web and PDFs, we expand MT datasets to 21 new languages and produce the largest and most genre-diverse in 20 more. We release translation models in 172 language directions, with 23/32 beating state-of-the-art benchmark performance, as well as a public dataset with 11.6M aligned bitexts and 3.4M monolingual sentences.

A large multilingual bitext like ours has potential to build the best yet or first ever MT models for many languages, something we accomplished on a surface level in this work but hope future works will continue. The data present a number of other potential uses, including: (1) training language models for applications like spelling correction (Abdulrahman and Hassani, 2022; Etoori et al., 2018; Al-Jefri and Mahmoud, 2013); (2) availing textual data for applications like speech recognition and speech translation, which can be vital to low-literacy communities (Robinson et al., 2022b; Gao et al., 2021; Rossenbach et al., 2020); (3) potential to study cross-lingual transfer between Creole languages and their phylogenetic relatives in yet unseen depth (Robinson et al., 2023b); (4) research on the effects of linguistic data augmentation for small MT datasets (focusing on Creole languages' unique linguistic position); (5) development of MT-assisted documentation tools for those languages that are endangered (Bird and Chiang, 2012); (6) the introduction of a common repository where any researchers in Creole NLP can accumulate datasets together and advance in collaboration (Lent et al., 2023); etc. We hope that this work will provide valuable translation technologies to communities that have been historically under-served and inspire community-oriented efforts to further expand work on these low-resource languages.

Acknowledgements

We thank Suzanna Sia, Chris Emezue, Heather Lent, David Mortensen, Oliver Mayeux, Michael Gisclair, Arya McCarthy, Ruth-Ann Armstrong, Carter Charles, Jeff Allen, Jamell Dacon, and Fritz-Carl Morlant for their contributions to the ideas and processes of this work.

Limitations

Across languages, writing systems change over time due to linguistic changes, such as the loss of distinctions between sounds; or metalinguistic changes, such as the desire to associate a speech community with a more prestigious one (or conversely show that the speech community is distinct). This concern is exacerbated for Creole languages, which tend to have very recent and often still developing standardization processes (Deuber and Hinrichs, 2007; Valdman, 2005; Rajah-Carrim, 2009). For several of our languages, especially Louisiana Creole and French Guianese Creole, we rely extensively on texts that were written more than 100 years ago and thus use spelling systems that have been partially or wholly superseded by new systems. In principle, it is possible to rewrite such texts with more recent conventions using language-specific scripts, but we opted to use the original orthographies for this work to keep the processing pipelines as similar as possible across languages. In the course of our data collection efforts, we identified several sources which we did not have time to process. In the future, we intend to continue adding new sources into our dataset, with a preference for those which are already publicly available.

Currently in our experiments, the newly created development and test scores are not comparable across the **public** and **all** splits due to their independent splitting processes (and hence dissimilar test sets). Our future work will focus on having development and test sets that are common across both splits, regardless of whether the training data in said splits are cleaned or not, for consistent comparisons.

Ethics Statement

Because Creole languages have frequently been the target of "othering" and marginalization (DeGraff, 2005; Lent et al., 2023), it is important to approach Creole language technologies with sensitivity. From a linguistic standpoint, the question of "What exactly makes a language Creole?" has contributed to the lack of prestige that many Creole languages currently face. For this study, we use existing literature to identify languages that have been considered Creole, but do not seek to assert a singular "Creole essence."

As with any other linguistic community, there are considerable differences in opinion concerning the desirability of MT in various Creole-speaking communities. We acknowledge that MT technologies do not inherently benefit all Creole language speakers. Many of them can already use existing MT tools in a different language, lessening the immediate benefit of MT tools in the relevant Creole. Others may be concerned about machine translation displacing human translators, or view the entire concept of MT as offensive, as it directly broaches the subject of linguistic differences between their languages and European languages (differences which are still broadly stigmatized). We also acknowledge that the intended use of some of the resources we collected may not have been MT. We do not wish to undermine the original purposes of anthology-style data but hope our work will support these endeavors.

We also acknowledge that the texts we have assembled, especially those which are older, religious and/or political, reflect many different viewpoints that may be considered dated, contested, or offensive in some cultural contexts. This is a natural part of the data collection process, but it is not an endorsement of the content of any given text. We did not seek to remove such viewpoints from our data, as they are culture-specific. Another risk of data collection is the inclusion of personally identifiable information that may pose a risk to some users. This is a particular problem with a commonly used Haitian-English bitext from WMT 2011 (Callison-Burch et al., 2011). Though it is difficult to avoid data contamination in this vein completely, we avoid including this dataset to mitigate this risk. We also acknowledge the potential for bias in our dataset, since it is not perfectly balanced in terms of genres and topics. We encourage more application-oriented work in the future to report MT results broken down by test set genre.

In our conception, the primary beneficiaries of Creole MT technologies would be monolingual Creole language speakers. Many monolingual Creole language speakers have limited literacy and would perhaps benefit from speech translation systems more than text-based systems. As such, we encourage future work in the area of speech technologies for Creole languages and hope the textual materials and models we provide in this work can be of use to that end.

Lastly, ChatGPT was used to assist software writing for this project. We acknowledge the ethical implications of LLM use are still being understood.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Roshna Abdulrahman and Hossein Hassani. 2022. A language model for spell checking of educational texts in Kurdish (Sorani). In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 189–198, Marseille, France. European Language Resources Association.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022b. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. MasakhaNER: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Majed M. Al-Jefri and Sabri A. Mahmoud. 2013. Context-sensitive arabic spell checker using context words and n-gram language models. In *2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*, pages 258–263.

Mervyn C Alleyne. 1971. Acculturation and the cultural matrix of creolization. *Pidginization and creolization of languages*, 1971:169–186.

Joe KYB Amoako. 1992. Ghanaian pidgin english: In search of synchronic, diachronic, and sociolinguistic evidence. *Unpublished PhD dissertation). University of Florida at Gainsville*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges.

Ruth-Ann Armstrong, John Hewitt, and Christopher D Manning. 2022. Jampatoisnli: A jamaican patois natural language inference dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5307–5320.

Cedric Audebert. 2017. The recent geodynamics of haitian migration in the americas: refugees or economic migrants? *Revista Brasileira de Estudos de População*, 34:55–71.

C. Baissac. 1888. *Le folk lore de l'He-Maurice (texte eréole et traduction française)*. Littératures populaires de toutes les nations. Maisonneuve et C. Leclere.

Angela Bartens. 2021. The making of languages and new literacies: San andrés-providence creole with a view on jamaican and haitian. *Lingüística y Literatura*, 42(79):237–256.

Parth Bhatt and Ingo Plag. 2012. *The structure of creole words: Segmental, syllabic and morphological aspects*, volume 505. Walter de Gruyter.

Steven Bird and David Chiang. 2012. Machine translation for language preservation. In *Proceedings of COLING 2012: Posters*, pages 125–134, Mumbai, India. The COLING 2012 Organizing Committee.

Eleftheria Briakou, Sida Wang, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. BitextEdit: Automatic bitext editing for improved low-resource machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1469–1485, Seattle, United States. Association for Computational Linguistics.

Anne-Marie Brousseau. 2011. One substrate, two creoles. In Claire Lefebvre, editor, *Creoles, their Substrates, and Language Typology*, pages 105–153. John Benjamins, Amsterdam.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the sixth workshop on statistical machine translation*, pages 22–64.

Bernard Caron, Marine Courtin, Kim Gerdes, and Sylvain Kahane. 2019. A surface-syntactic UD treebank for Naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 13–24. Association for Computational Linguistics.

Frederic G Cassidy. 1966. Multiple etymologies in jamaican creole. *American Speech*, 41(3):211–215.

Yolanda Rivera Castillo and Nicholas Faraclas. 2006. The emergence of systems of lexical and grammatical tone and stress in caribbean and west african creoles. *STUF - Language Typology and Universals*, 59(2):148–169.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49:375–395.

Raphaël Confiant. 2007. Dictionnaire créole martiniquais-français. *(No Title)*.

Michael L Conniff. 1983. Black labor on a white canal: West indians in panama, 1904-1980.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Diptesh Kanojia, Chinmay Sawant, and Eiichiro Sumita. 2023. YANMTT: Yet another neural machine translation toolkit. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 257–263, Toronto, Canada. Association for Computational Linguistics.

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

Raj Dabre and Aneerav Sukhoo. 2022b. KreolMorisienMT: 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.

A. 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.

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

Michel DeGraff. 2005. Linguists' most dangerous myth: The fallacy of creole exceptionalism. *Language in society*, 34(4):533–591.

Dagmar Deuber and Lars Hinrichs. 2007. Dynamics of orthographic standardization in jamaican creole and nigerian pidgin. *World Englishes*, 26:22–47.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco GuzmÃ¡n, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. XLEnt: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10424–10430, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Valts Ernštreits, Mark Fišel, Matīss Rikters, Marili Tomingas, and Tuuli Tuisk. 2022. Language resources and tools for livonian. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 13(1):13–36.

Pravallika Etoori, Manoj Chinnakotla, and Radhika Mamidi. 2018. Automatic spelling correction for resource-scarce languages using deep learning. In *Proceedings of ACL 2018, Student Research Workshop*, pages 146–152, Melbourne, Australia. Association for Computational Linguistics.

Joseph Farquharson. 2012. *The African lexis in Jamaican: Its linguistic and sociohistoircal significance*. Ph.D. thesis, University of the West Indies.

A. Fortier. 1895. *Louisiana Folk-tales: In French Dialect and English Translation*. Memoirs of the American Folk-Lore Society. American Folk-lore society.

Robert Frederking, Ralf D Brown, and Christopher Hogan. 1997. The diplomat rapid development speech mt system. In *Proceedings of Machine Translation Summit VI: Systems*, pages 261–262.

Changfeng Gao, Gaofeng Cheng, Runyan Yang, Han Zhu, Pengyuan Zhang, and Yonghong Yan. 2021. Pre-training transformer decoder for end-to-end asr model with unpaired text data. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6543–6547. IEEE.

Édouard Glissant. 2008. Creolization in the making of the americas. *Caribbean Quarterly*, 54:81 – 89.

Adrien Guillory-Chatman, Oliver Mayeux, Nathan Wendte, and Herbert Wiltz. 2020. *Ti Liv Kréyòl (Second edition)*.

Jessica Heinzelman and Carol Waters. 2010. *Crowdsourcing crisis information in disaster-affected Haiti*. JSTOR.

Glaude Herby. 2012. Creoloral (oral corpus with annotations).

Anita Herzfeld. 1980. Limon creole and panamanian creole: comparison and contrast. Mid-America Linguistics Conference.

SR Hosseini, M Scaioni, M Marani, et al. 2018. On the influence of global warming on atlantic hurricane frequency. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42(3):527–532.

Matthew Hunt, Sharon O'Brien, Patrick Cadwell, and Dónal P O'Mathúna. 2019. Ethics at the intersection of crisis translation and humanitarian innovation. *Journal of Humanitarian Affairs*, 1(3):23–32.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Abdul D Knowles. 2018. Case study: Preventing and resolving conflict between bahamian nationals and the haitian diaspora that reside in the bahamas. *International Journal of Law and Public Administration*, 1(2):65–73.

Silvia Kouwenberg. 2008. The problem of multiple substrates: The case of Jamaican Creole. In Susanne Michaelis, editor, *Roots of creole structures: Weighing the contribution of substrates and superstrates*, pages 1–27. John Benjamins.

Silvia Kouwenberg and Darlene Lacharité. 2004. Echoes of africa: Reduplication in caribbean creole and niger-congo languages. *Journal of Pidgin and Creole Languages*, 19:285–331.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2021. QED: A framework and dataset for explanations in question answering. *Transactions of the Association for Computational Linguistics*, 9:790–806.

Claire Lefebvre. 2011. Substrate features in the properties of verbs in three atlantic creoles: Haitian Creole, Saramaccan and Papiamentu. In Claire Lefebvre, editor, *Creoles, their Substrates, and Language Typology*, pages 127–154. John Benjamins, Amsterdam.

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, Hans Erik Heje, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, Anders Søgaard, and Johannes Bjerva. 2023. CreoleVal: Multilingual multitask benchmarks for creoles.

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*, Saint Raphaël, France. European Association for Machine Translation.

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

Zhengyuan Liu, Shikang Ni, Aiti Aw, and Nancy Chen. 2022. Singlish message paraphrasing: A joint task of creole translation and text normalization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3924–3936.

Cory Look, Erin Friedman, and Geneviève Godbout. 2019. The resilience of land tenure regimes during hurricane irma: How colonial legacies impact disaster response and recovery in antigua and barbuda. *Journal of Extreme Events*, 6(01):1940004.

Rhoda Margesson and Maureen Taft-Morales. 2010. Haiti earthquake: Crisis and response. Library of Congress Washington DC Congressional Research Service.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020a. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Arya D McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020b. The johns hopkins university bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892.

Juan McCartney. 2013. The rise of the haitian population: Community expands since independence. *The Nassau Guardian*, 17.

John McWhorter. 2000. *The missing Spanish creoles: Recovering the birth of plantation contact languages*. Univ of California Press.

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

Nguyen-Hoang Minh-Cong, Nguyen Van Vinh, and Nguyen Le-Minh. 2023. A fast method to filter noisy parallel data wmt2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 359–365.

Ludovic Mompelat, Daniel Dakota, and Sandra Kübler. 2022. How to parse a creole: When martinican creole meets french. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4397–4406.

Salikoko Mufwene. 2002. Socio-economic historical arguments for a gradual and heterogeneous development of patois in Jamaica. In *Biennial Meeting of Society for Caribbean Linguistics*.

Salikoko S Mufwene. 2008. Pidgins and creoles. In *The Handbook of World Englishes*, pages 313–327. Blackwell Publishing Ltd, Oxford, UK.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023. SemEval-2023 task 12: Sentiment analysis for African languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2319–2337, Toronto, Canada. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Adelani, Anuoluwapo Aremu, and Idris Abdulmumin. 2022. NaijaSenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602.

Antonio Carvalho Neto, Fernanda Versiani, Kelly Pellizari, Carolina Mota-Santos, and Gustavo Abreu. 2020. Latin american, african and asian immigrants working in brazilian organizations: facing the language barrier. *Revista Economia & Gestão*, 20(55):87–101.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *First International Workshop on Linked Science 2011-In conjunction with the International Semantic Web Conference (ISWC 2011)*.

Kelechi Ogueji and Orevaoghene Ahia. 2019. Pidginunmt: Unsupervised neural machine translation from west african pidgin to english. *arXiv preprint arXiv:1912.03444*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, and Olalekan Akinsande. 2020. Semantic enrichment of nigerian pidgin english for contextual sentiment classification. *arXiv preprint arXiv:2003.12450*.

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.

A. Parépou and M. Fauquenoy. 1987. *Atipa: (roman guyanais) Paris, A. Ghio, 1885*. Textes, études et documents. Editions l'Harmattan.

Charmane M Perry. 2023. 'real bahamians' and 'paper bahamians': Haitians as perpetual foreigners. *Latin American and Caribbean Ethnic Studies*, 18(1):122–140.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Ronald Pinas, Lucien Donk, Hertoch Linger, Arnie Lo-Ning-Hing, Tieneke MacBean, Celita Zebeda-Bendt, Chiquita Pawironadi-Nunez, and Dorothy Wong Loi Sing. 2007. *Wortubuku fu Sranan Tongo*.

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.

Aaliya Rajah-Carrim. 2009. Use and Standardisation of Mauritian Creole in Electronically Mediated Communication1. *Journal of Computer-Mediated Communication*, 14(3):484–508.

Andrew Rasmussen, Eddy Eustache, Giuseppe Raviola, Bonnie Kaiser, David J Grelotti, and Gary S Belkin. 2015. Development and validation of a haitian creole screening instrument for depression. *Transcultural psychiatry*, 52(1):33–57.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

John R Rickford and John McWhorter. 2017. Language contact and language generation: Pidgins and creoles. *The handbook of sociolinguistics*, pages 238–256.

Nathaniel Robinson, Cameron Hogan, Nancy Fulda, and David R. Mortensen. 2022a. Data-adaptive transfer learning for translation: A case study in Haitian and jamaican. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 35–42, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, Swetha Gangu, David R Mortensen, and Shinji Watanabe. 2022b. When is tts augmentation through a pivot language useful? In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2022, pages 3538–3542.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023a. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Nathaniel Robinson, Matthew D. Stutzman, Stephen D. Richardson, and David R. Mortensen. 2023b. African substrates rather than european lexifiers to augment african-diaspora creole translation. In *4th Workshop on African Natural Language Processing*.

Ulisdete Rodrigues. 2007. Fonologia do caboverdiano : das variedades insulares à unidade nacional.

Simon Romero. 2010. A language thrives in its caribbean home. *New York Times*, 4.

Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2020. Generating synthetic audio data for attention-based speech recognition systems. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7069–7073.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Luisa Seguin. 2020. Transparency and language contact: The case of haitian creole, french, and fongbe. *Journal of Pidgin and Creole Languages*, 35(2):218–252.

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, Online. Association for Computational Linguistics.

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the WMT 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 95–102.

Keston Smith. 2022. Trinidad english creole to english dataset.

Steinþór Steingrímsson. 2023. A sentence alignment approach to document alignment and multi-faceted filtering for curating parallel sentence pairs from web-crawled data. In *Proceedings of the Eighth Conference on Machine Translation*, pages 366–374.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Kyllah Terry and A Mayes. 2019. New haitian migration patterns end in displacement. *UCLA Latin American Institute, April*, 17.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Albert Valdman. 2000. L'évolution du lexique dans les créoles à base lexicale française. *L'Information Grammaticale*, 85:53–60.

Albert Valdman. 2005. Vers la standardisation du créole haïtien. *Revue Francaise De Linguistique Appliquee*, 10:39–52.

Viveka Velupillai. 2015. *Pidgins, creoles and mixed languages*. Creole Language Library. John Benjamins Publishing, Amsterdam, Netherlands.

Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal dependencies parsing for colloquial singaporean english. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1732–1744.

Donald Winford. 1997. Re-examining caribbean english creole continua. *World Englishes*, 16(2):233–279.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2023a. Lego-mt: Learning detachable models for massively multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11518–11533.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2023b. Lego-MT: Learning detachable models for massively multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11518–11533, Toronto, Canada. Association for Computational Linguistics.

Joseph Zhong. 2023. Haiti and the dominican republic's long road to economic growth divergence.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

Just Zwennicker and David Stap. 2022. Towards a general purpose machine translation system for sranantongo. *arXiv preprint arXiv:2212.06383*.

A Additional Data Information

Table 5 contains bitext sizes for individual previous publications for each language, compared to our own datasets (summarized in Table 3). Table 6 provides numerical values for our dataset's genre composition (corresponding to Figure 2). Table 7 gives language-by-language details of the types of data we extracted in our collection methodology (summarized in Table 2). Table 8 shows alternate names for languages with more than one used in data gathering.

B Experimental Details

We describe some additional details related to training and evaluation.

B.1 mBART-50 Token Repurposing

mBART-50 has 51 special tokens corresponding to exactly 51 languages in our experiments, of which 41 are Creoles. Therefore, we simply repurpose these tokens where we fix the tokens for English, French, Spanish and Portuguese to the corresponding tokens in the mBART-50 tokenizer and then randomly assigned other tokens to Creoles. We found that this only affects initial training, as the model has to re-learn that the token is used to translate into another language.

B.2 Training and Convergence

We train all models until convergence[19] on the validation sets, evaluated every 5000 steps, and up to a maximum of 500 000 steps. For the models trained from scratch, this took three days on a single node of 8 32 GiB V100 GPUs. (Fine-tuning mBART took much longer, as noted in § 5.) See Table 4 for the sizes of our models.

B.3 Hyperparameters

The relevant training hyperparameters are given in Table 4.

Optimizer	AdamW
Learning Rate	1e–3 (3e–4)
Weight Decay	1e–5
#encoder/decoder layers	6 (12)
hidden size	512 (1024)
FFN size	2048 (4096)
#parameters	77 M (611 M)
data sampling temperature	2.0[20]
batch size	32 768[21] (8192)
dropout	0.1 (0.1)
label smoothing	0.1

Table 4: Hyperparameter settings for models trained from scratch. Values in parentheses indicate those used for fine-tuning mBART-50.

B.4 Additional Models Trained

We train a simpler version of *m2m-mBART* focusing only on Haitian–English bidirectional translation to determine whether its better to focus on fewer language pairs during fine-tuning. The key changes in hyperparameters is learning rate (3e–5) and dropout (0.3).

C Additional Results

C.1 BLEU Scores For Our Test Sets

Figure 5 shows the BLEU scores on our newly proposed test sets. These are analogous to the chrF scores presented in Figure 3.

C.2 Bilingual vs Multilingual mBART fine-tuning

Figure 6 shows the results of fine-tuning mBART-50 only on Haitian–English vs on multilingual data, for the **public** split mentioned in Section 4.1. We can see that bilingual models are inferior to multilingual models. The same trend exists for the **all** split.

C.3 LegoMT and KreolMorisienMT Results

Figure 7 shows results of our models on the KreolMorisienMT (Dabre and Sukhoo, 2022a) test sets, and Figure 8 shows results of the same models on the Lego-MT (Yuan et al., 2023a) test sets. Yuan et al. (2023a) did not report results on their own test splits; hence we do not compare our models' performance directly with theirs. Once again, our fine-tuned models give the best performance.

[19] We use annealing to declare convergence: we first wait until the model does not show BLEU improvements for 10 consecutive evaluations. We then decrease the learning rate by half and wait until the model does not show BLEU improvements for 15 evaluations. Finally, we decrease the learning rate by half again and if the model does not show improvements for 20 evaluations then we declare convergence.

[20] To slightly oversample the smaller corpora. We noticed in our peliminary experiments that the standard temperature of 5.0 caused the higher resource pairs to suffer.

[21] This is the total batch size in tokens over 8 32 GiB V100 GPUs.

	CreoleVal	JHU	Lego-MT	FLORES	AfricaNLP	NLLB	Ours	Ours
Public?	✗	✗	✓	✓	✗	✓	✓	✗ ✓
acf	7864	15989	-	-	-	-	4406	23916
aoa	-	-	-	-	-	-	198	198
bah	-	-	-	-	-	-	327	327
brc	-	-	-	-	-	-	222	222
bzj	14911	23406	-	-	-	-	229	31002
bzk	-	-	-	-	-	-	391	391
cab	-	20879	-	-	-	-	-	20879
cri	-	-	-	-	-	-	306	306
crs	222613	5055	-	-	-	-	3186	225875
dcr	-	-	-	-	-	-	189	189
djk	45361	23748	7868	-	-	-	15266	68833
fab	-	-	-	-	-	-	204	204
fng	-	-	-	-	-	-	160	160
fpe	-	-	-	-	-	-	259	259
gcf	-	-	96	-	-	-	6467	6467
gcr	-	-	-	-	-	-	1433	1433
gpe	-	-	-	-	-	-	223	223
gul	7870	7990	-	-	-	-	266	8831
gyn	-	-	-	-	-	-	258	258
hat	210593	72354	477048	2006	179435	4256455	5715227	6023034
icr	7799	15702	-	-	-	-	317	16774
jam	7988	25206	26	-	5118	-	434	28713
kea	-	-	-	2009	-	129449	132931	132931
kri	50438	23740	-	-	-	-	185	66736
ktu	7886	5055	-	-	-	-	175	10737
lou	-	-	-	-	-	-	1860	1860
mart1259	-	-	-	-	-	-	5153	5153
mfe	191909	23625	399	-	-	-	25633	233320
mue	-	-	-	-	-	-	147	147
pap	397354	5018	269	2009	-	4898029	4968965	5363394
pcm	31128	15905	-	-	-	-	8084	47455
pov	-	-	-	-	-	-	480	480
pre	-	-	-	-	-	-	243	243
rcf	-	-	-	-	-	-	285	285
sag	262334	16952	9	2009	-	235749	260560	535310
srm	42303	23531	-	-	-	-	440	59053
srn	583830	24569	-	-	-	-	6620	615010
svc	-	-	-	-	-	-	321	321
tpi	398341	81595	70	2009	-	424626	451758	925648
trf	-	-	-	-	-	-	1691	1691

Table 5: Size of total bitext data collected for Creole languages to date, compared with our full combined bitext sets. Bitext size is measured as the number of unique Creole language segments paired with a translation in any target language

Lang	Bible	Educational	Legal	Narrative	News	Religious	Wikipedia	Other/Mix
acf	15989	4406	0	0	0	0	0	33778
aoa	0	198	0	0	0	0	0	0
bah	0	327	0	0	0	0	0	0
brc	0	222	0	0	0	0	0	0
bzj	54933	229	0	0	0	0	0	10213
bzk	0	391	0	0	0	0	0	0
cab	20879	0	0	0	0	0	0	47471
cri	0	306	0	0	0	0	0	0
crs	5055	273	443948	15719	4141	0	0	279331
dcr	0	189	0	0	0	0	0	0
djk	23815	7398	0	0	0	0	0	66491
fab	0	204	0	0	0	0	0	0
fng	0	160	0	0	0	0	0	0
fpe	0	259	0	0	0	0	0	0
gcf	1559	304	0	4446	0	0	0	158
gcr	879	159	0	2388	0	0	15141	0
gpe	0	223	0	0	0	0	12425	0
gul	7990	262	0	0	0	0	0	579
gyn	0	258	0	0	0	0	0	0
hat	71958	9359	0	4	0	115583	5452	5828317
icr	15702	317	0	0	0	0	0	755
jam	18420	233	0	0	0	0	4588	29647
kea	0	6108	0	0	229	0	0	132314
kri	16113	185	0	0	0	0	0	99454
ktu	5055	175	0	0	0	0	0	62584
lou	0	668	0	1192	0	0	0	0
mart1259	0	283	0	0	0	0	0	4870
mfe	23624	258	0	274	0	0	0	277331
mue	0	147	0	0	0	0	0	0
pap	5018	2996	0	0	0	65573	92	7430445
pcm	15905	253	0	0	0	0	0	31297
pov	0	480	0	0	0	0	0	0
pre	0	243	0	0	0	0	0	0
rcf	0	285	0	0	83659	0	0	63975
sag	16952	192	0	0	0	0	0	518166
srm	23531	440	0	0	0	0	0	57013
srn	24567	6607	0	0	0	0	4600	766364
svc	0	321	0	0	0	0	0	0
tpi	81178	62	0	0	0	25018	7	819383
trf	0	174	0	0	0	0	0	1517
wes	0	223	0	0	0	0	0	0
Total	449122	45777	443948	24023	88029	206174	42305	16561453

Table 6: Amount of sentences for each language in each genre.

	Bitext				Monolingual			
	Web		*PDF*					
Lang	*Prev. pub.*	*aligned*	*articles*	*aligned*	*other*	*Prev. pub.*	*Web*	*PDF*
acf	19510	0	0	4406	0	30257	0	0
aoa	0	198	0	0	0	0	0	0
bah	0	327	0	0	0	0	0	0
brc	0	222	0	0	0	0	0	0
bzj	30773	229	0	0	0	0	34373	0
bzk	0	391	0	0	0	0	0	0
cab	20879	0	0	0	0	47471	0	0
cri	0	306	0	0	0	0	0	0
crs	222690	273	0	0	2912	61696	445177	15719
dcr	0	189	0	0	0	0	0	0
djk	61435	7398	0	0	0	28871	0	0
fab	0	204	0	0	0	0	0	0
fng	0	160	0	0	0	0	0	0
fpe	0	259	0	0	0	0	0	0
gcf	64	4844	1559	0	0	0	0	0
gcr	0	159	880	0	394	0	15140	1994
gpe	0	223	0	0	0	0	12425	0
gul	8569	262	0	0	0	0	0	0
gyn	0	258	0	0	0	0	0	0
hat	5900275	165	122594	0	0	0	7639	0
icr	16457	317	0	0	0	0	0	0
jam	28311	233	169	0	0	19756	4419	0
kea	131454	484	860	0	133	0	229	5491
kri	66551	185	0	0	0	49016	0	0
ktu	10562	175	0	0	0	57077	0	0
lou	0	440	0	228	1192	0	0	0
mart1259	0	231	0	4870	52	0	0	0
mfe	232788	258	0	274	0	68167	0	0
mue	0	147	0	0	0	0	0	0
pap	5294750	146	65665	2833	0	2140730	0	0
pcm	47202	253	0	0	0	0	0	0
pov	0	480	0	0	0	0	0	0
pre	0	243	0	0	0	0	0	0
rcf	0	285	0	0	0	60098	83659	3877
sag	535118	192	0	0	0	0	0	0
srm	58613	440	0	0	0	21931	0	0
srn	608397	606	4	6003	0	182532	4596	0
svc	0	321	0	0	0	0	0	0
tpi	900560	63	25025	0	0	0	0	0
trf	1517	174	0	0	0	0	0	0
wes	0	223	0	0	0	0	0	0
Total	14196475	21963	216756	18614	4683	2767602	607657	27081

Table 7: Number of segments gathered from each source type/extraction method for each language

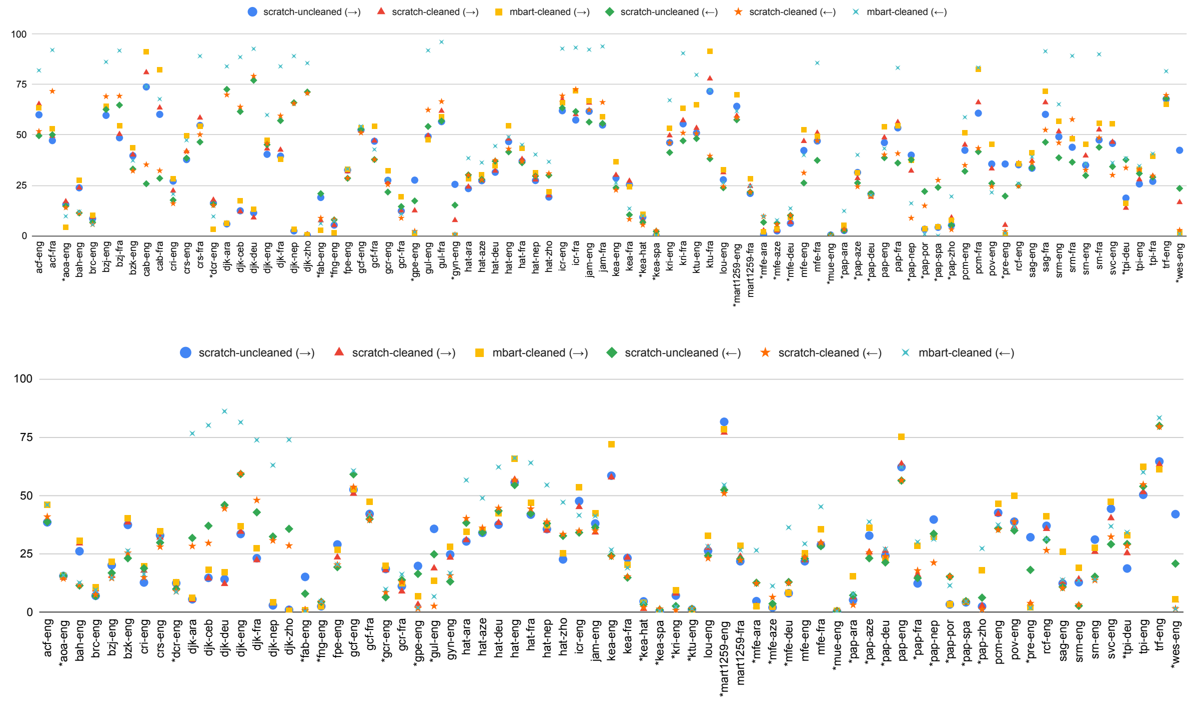

Figure 5: BLEU scores on our newly created test sets using models trained on the **all** (top) and **public** (bottom) splits of our datasets. Given X-Y pair, → and ← represent the X to Y and Y to X translation, respectively. Zero-shot pairs are marked with an '*' sign.

ISO	Name 1	Name 2	Name 3	Name 4
gcf	French Antillean	Guadeloupean Creole	Martinican	-
gcr	French Guianese	Kriyòl Gwiyannen	Kriyòl Lagwiyann	Gwiyannen
djk	Ndyuka	Eastern Maroon Creole	Aukan	Nengee
kmv	Karipuna	Amapá Creole	Uaçá Creole	-
bzk	Miskito Coast Creole	Nicaraguan Creole English	-	-
pcm	Naija	Nigerian Pidgin	-	-
kea	Cape Verdean	Kabuverdianu	-	-
mfe	Mauritian Creole	Morisyen	-	-
crs	Seychellois	Seselwa	Kreol Sesel	-

Table 8: Alternate names for languages with more than one common name. The ISO-639 code and all names used for searches in our data collection process are listed.

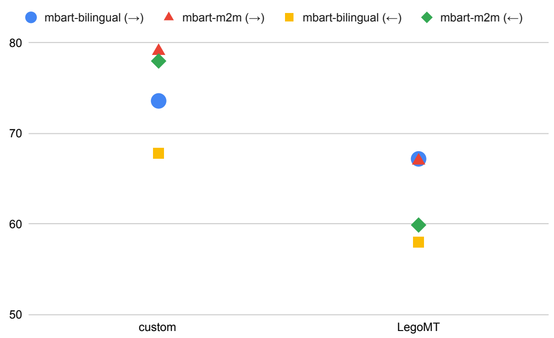

Figure 6: Comparing Haitian to English (→) and English to Haitian (←) translation quality for bilingual and multilingual fine-tuning for our custom test set as well as the LegoMT test set.

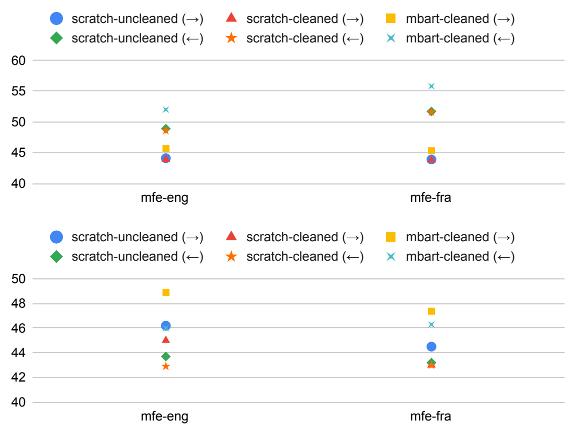

Figure 7: Results on the KreolMorisienMT test sets using the models trained on **all** (top) and **public** (bottom) data.

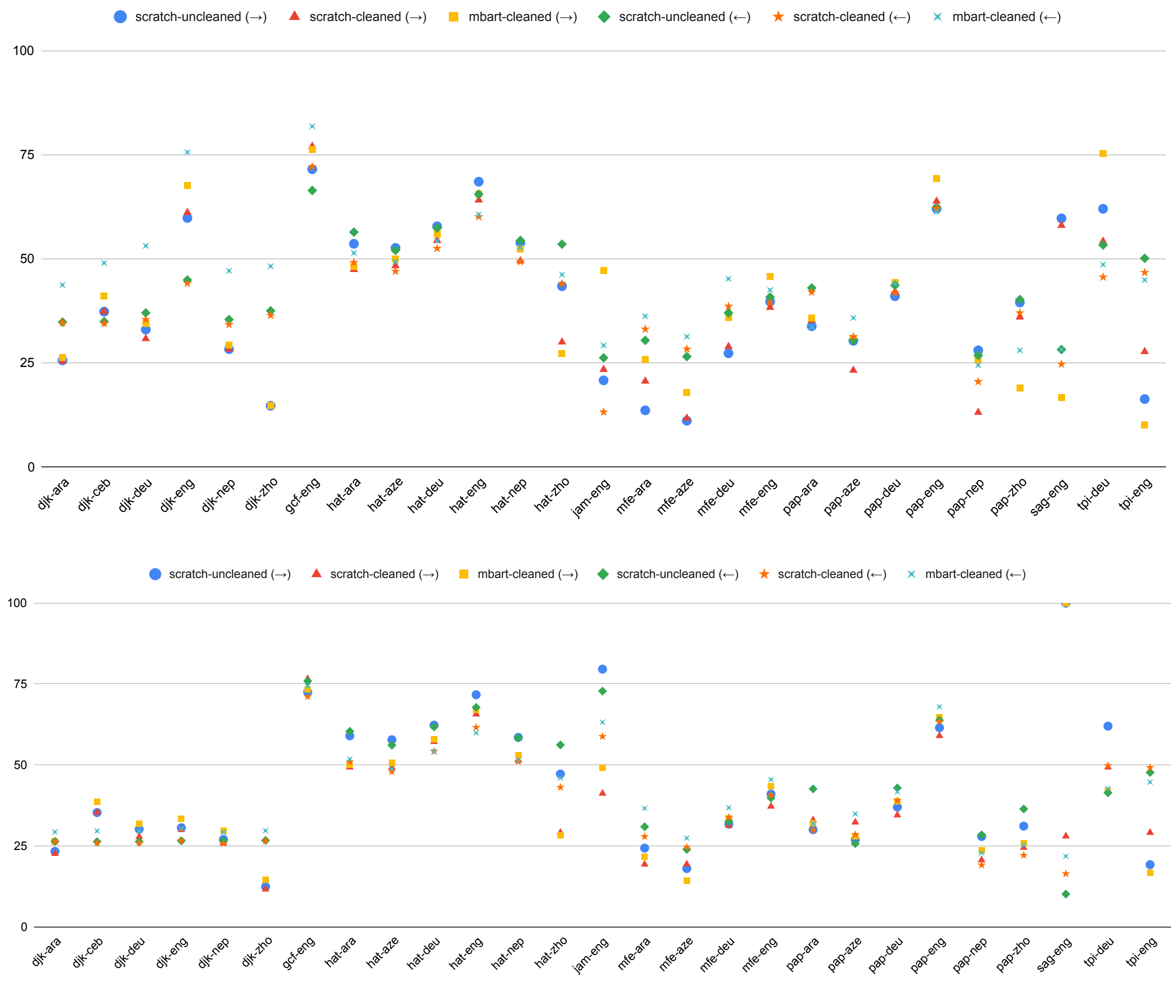

Figure 8: Results on the LegoMT test sets using the models trained on **all** (top) and **public** (bottom) data.

D Attributions

We provide exact attributions for all our data sources here. We list them with string identifiers that we used to distinguish them in our own data organization. See our repository https://github.com/JHU-CLSP/Kreyol-MT for data downloading instructions.

D.1 Resources for Bitexts

APiCS - Atlas of Pidgin and Creole Language Structures (Michaelis et al., 2013)

- **Languages**: *dcr, icr, bzj, pov, fng, trf, rcf, fpe, kri, pap, gcf, gpe, tpi, crs, ktu, pre, fab, bah, srm, gyn, djk, brc, sag, aoa, pcm, svc, mart1259, bzk, cri, gcr, kea, wes, hat, lou, srn, jam, mue, mfe, gul*
- **Links**: https://apics-online.info/

AfricaNLP-2023 - Aligned sentenced pairs from (Robinson et al., 2023b)

- **Languages**: *hat, jam*
- **Links**: https://openreview.net/forum?id=YKUv4sSOom

bible_uedin - bible-uedin-v1, Parallel corpus created from translations of the Bible (Christodouloupoulos and Steedman, 2015)

- **Languages**: *djk*
- **Links**: https://opus.nlpl.eu/bible-uedin/corpus/version/bible-uedin

Bidze2019 - Seychelles Government Budget For the Fiscal Year 2019, Office of the President of The Republic of Seychelles

- **Languages**: *crs*
- **Links**: https://www.statehouse.gov.sc/downloads?page=2

Bidze2021 - Seychelles Government Budget For the Fiscal Year 2021, Office of the President of The Republic of Seychelles

- **Languages**: *crs*
- **Links**: https://www.statehouse.gov.sc/downloads?page=1

boston-food-forest - Boston Food Forest Coalition flyers translations

- **Languages**: *kea*
- **Links**: https://www.bostonfoodforest.org/languages

CJCLDS - Online library of The Church of Jesus Christ of Latter-day Saints

- **Languages**: *pap, hat, tpi*
- **Links**: https://www.churchofjesuschrist.org/study?lang=pap (Link to full LDC dataset available on our repository: https://github.com/JHU-CLSP/Kreyol-MT.)

CREOLORAL - Martinican and Guadeloupean oral corpus with annotations (Herby, 2012)

- **Languages**: *gcf*
- **Links**: https://cocoon.huma-num.fr/exist/crdo/search2.xql?lang=fr&language=http%3A%2F%2Flexvo.org%2Fid%2Fiso639-3%2Fgcf

Confiant-Dictionary - Dictionnaire Créole Martiniquais - Français, Raphaël Confiant (Confiant, 2007)

- **Languages**: *mart1259*
- **Links**: https://www.potomitan.info/dictionnaire/

CreoleVal (Lent et al., 2023)

- **Languages**: *djk, kri, icr, pap, hat, bzj, sag, ktu, acf, srn, pcm, tpi, jam, crs, mfe, gul, srm*
- **Links**: https://arxiv.org/abs/2310.19567

dicoNengee - Dictionnaire Nengee - Français - English

- **Languages**: *djk*
- **Links**: https://corporan.huma-num.fr/Lexiques/dicoNengee.html

FLORES-200 (NLLB Team et al., 2022)

- **Languages**: *kea, pap, hat, sag, tpi*
- **Links**: https://github.com/facebookresearch/flores/blob/main/flores200

folklore - Excerpts from *Le folklore de l'Ile-Maurice (texte créole et traduction française)* (Baissac, 1888)

- **Languages**: *mfe*
- **Links**: https://archive.org/details/lefolkloredelile00bais/page/98/mode/2up

fortier - Excerpts from *Louisiana Folk-tales: In French Dialect and English Translation* (Fortier, 1895)

- **Languages**: *lou*
- **Links**: https://archive.org/details/b24865424/page/n11/mode/2up

GoiloText - Papiamentu Textbook, E.R. Goilo

- **Languages**: *pap*
- **Links**: https://archive.org/details/PapiamentuTextbook/mode/2up

JHU - The Johns Hopkins University Bible Corpus (McCarthy et al., 2020b)

- **Languages**: *djk, kri, icr, pap, hat, bzj, sag, cab, ktu, acf, srn, pcm, tpi, jam, crs, mfe, gul, srm*
- **Links**: https://aclanthology.org/2020.lrec-1.352/

kapes - Corrections of the "Certificat d'aptitude au professorat de l'enseignement du second degré" (CAPES) exam for Martinican and Guadeloupean creole

- **Languages**: *gcf, mart1259*
- **Links**: https://kapeskreyol.potomitan.info/

KreolMorisienMT (Dabre and Sukhoo, 2022a)

- **Languages**: *mfe*
- **Links**: https://aclanthology.org/2022.findings-aacl.3.pdf

LAFANDMT (Adelani et al., 2022b)

- **Languages**: *pcm*
- **Links**: https://github.com/masakhane-io/lafand-mt

LegoMT (Yuan et al., 2023b)

- **Languages**: *djk, pap, gcf, hat, sag, tpi, jam, mfe*
- **Links**: https://aclanthology.org/2023.findings-acl.731/

mindelo - Online dictionary

- **Languages**: *kea*
- **Links**: http://www.mindelo.info/_dico.php

MIT-Haiti, MIT-Haiti Initiative (Lent et al., 2023)

- **Languages**: *hat*
- **Links**: https://haiti.mit.edu/hat/resous/

MiBelNouvel - Translation of the Gospel of John in Guadeloupean Creole

- **Languages**: *gcf*
- **Links**: https://mibelnouvel.wordpress.com/

MultiCCAligned (Tiedemann, 2012; El-Kishky et al., 2020)

- **Languages**: *hat*
- **Links**: https://opus.nlpl.eu/MultiCCAligned.php

NLLB NLLB-v1 (NLLB Team et al., 2022; Schwenk et al., 2021)

- **Languages**: *hat, tpi, sag*
- **Links**: https://opus.nlpl.eu/NLLB/corpus/version/NLLB, https://huggingface.co/datasets/allenai/nllb

PwovebKreyol - Proverbes & expressions créoles

- **Languages**: *gcf*
- **Links**: http://pwoveb.kreyol.free.fr/proverbes.php

QCRI - (Abdelali et al., 2014)

- **Languages**: *pap*
- **Links**: https://opus.nlpl.eu/QED/corpus/version/QED

QED - (Lamm et al., 2021)

- **Languages**: *mfe, hat, pap, sag*
- **Links**: https://aclanthology.org/2021.tacl-1.48/

quentin - Excerpts from *Introduction à l'histoire de Cayenne ; suivie d'un Recueil de contes, fables et chansons en créole* (de Saint-Quentin, 1872)

- **Languages**: *gcr*
- **Links**: https://gallica.bnf.fr/ark:/12148/bpt6k82939m.r=creole%20guyanais%20quentin?rk=21459;2

SIL-Suriname - Languages of Suriname, SIL

- **Languages**: *djk, srm*
- **Links**: https://suriname-languages.sil.org/Aukan/Aukan.html, https://suriname-languages.sil.org/Saramaccan/Saramaccan.html

Saint_Lucia_Ministry_of_Ed - Kwéyòl Dictionary, Ministry of Education Government of Saint Lucia

- **Languages**: *acf*
- **Links**: http://www.saintluciancreole.dbfrank.net/dictionary/KweyolDictionary.pdf

TEC-English - Trinidad English Creole to English Dataset, University of the West Indies (Smith, 2022)

- **Languages**: *trf*
- **Links**: https://data.mendeley.com/datasets/n4259kw9y7/1

TED2020 - TED and TED-X transcripts (Reimers and Gurevych, 2020)

- **Languages**: *hat*
- **Links**: https://opus.nlpl.eu/TED2020.php

Tatoeba - Tatoeba database

- **Languages**: *pap, gcf, sag, srn, tpi, jam, mfe, hat*
- **Links**: https://tatoeba.org/en/downloads

TiLiv - Ti Liv Kreyol, A learner's guide to Louisiana creole (Guillory-Chatman et al., 2020)

- **Languages**: *lou*
- **Links**: https://dn790005.ca.archive.org/0/items/ti-liv-kreyol-second-edition/Ti%20Liv%20Kreyol%20Second%20Edition.pdf

Ubuntu - Ubuntu Translations

- **Languages**: *hat, pap*
- **Links**: https://object.pouta.csc.fi/OPUS-Ubuntu/v14.10/moses/en-ht.txt.zip, https://object.pouta.csc.fi/OPUS-Ubuntu/v14.10/moses/en_AU-ht.txt.zip, https://object.pouta.csc.fi/OPUS-Ubuntu/v14.10/moses/en_CA-ht.txt.zip, https://object.pouta.csc.fi/OPUS-Ubuntu/v14.10/moses/en_GB-ht.txt.zip, https://object.pouta.csc.fi/OPUS-Ubuntu/v14.10/moses/en-pap.txt.zip, https://object.pouta.csc.fi/OPUS-Ubuntu/v14.10/moses/en_AU-pap.txt.zip, https://object.pouta.csc.fi/OPUS-Ubuntu/v14.10/moses/en_CA-pap.txt.zip, https://object.pouta.csc.fi/OPUS-Ubuntu/v14.10/moses/en_GB-pap.txt.zip

Wikimedia

- **Languages**: *pap, hat, tpi, jam, srn, gcr*

- **Links**: https://en.wikipedia.org/

Wikipedia

- **Languages**: *gcf*
- **Links**: https://fr.wikipedia.org/wiki/Cr%C3%A9ole_martiniquais

Wortubuku - Wortubuku fu Sranan Tongo, Sranan Tongo - English Dictionary (Pinas et al., 2007)

- **Languages**: *srn*
- **Links**: https://www.sil.org/resources/archives/1538

XLent (El-Kishky et al., 2021)

- **Languages**: *hat*
- **Links**: https://aclanthology.org/2021.emnlp-main.814/

YouVersion–Bible - Life.Church online Bible

- **Languages**: *gcr*
- **Links**: https://www.bible.com/bible/2963/JHN.INTRO1.GCR07

D.2 Resources for Monolingual Corpora

Anacao - Articles from the *A Nação* newspaper

- **Languages**: *kea*
- **Links**: https://www.anacao.cv/,

atipa - Excerpts from the book *Atipa* (Parépou and Fauquenoy, 1987)

- **Languages**: *gcr*
- **Links**: https://www.google.com.ng/books/edition/_/F7bA4J4D6T4C?hl=en&kptab=overview

Belizean - Life.Church online Bible

- **Languages**: *bzj*
- **Links**: https://www.bible.com/bible/409/MAT.1.BZJ

Creolica - Online corpus of Seychellois Créole

- **Languages**: *rcf, crs*
- **Links**: https://creolica.net/Corpus-de-creole-seychellois, https://creolica.net/Corpus-de-creole-reunionnais

Fonologia - Extracted interviews from Rodrigues (2007)

- **Languages**: *kea*
- **Links**: https://core.ac.uk/download/pdf/33531609.pdf

graelo - Wikipedia dumps

- **Languages**: *gcr*
- **Links**: https://huggingface.co/datasets/graelo/wikipedia

JamPatoisNLI - (Armstrong et al., 2022)

- **Languages**: *jam*
- **Links**: https://arxiv.org/abs/2212.03419

KreolMorisienMT - (Dabre and Sukhoo, 2022b)

- **Languages**: *mfe*
- **Links**: https://aclanthology.org/2022.findings-aacl.3.pdf

MADLAD-400 - (Kudugunta et al., 2024)

- **Languages**: *mfe, pap, crs, kri, srm, jam, srn, djk, ktu, acf, rcf, cab, bzj*
- **Links**: https://arxiv.org/abs/2309.04662

MIT-Haiti - Learning Resources from the MIT-Haiti initiative (Lent et al., 2023)

- **Languages**: *hat*
- **Links**: https://haiti.mit.edu/hat/resous/

National - Official minutes of the Sittings of the House, The National Assembly of Seychelles

- **Languages**: *crs*
- **Links**: https://www.nationalassembly.sc/verbatim

Seychelles - Articles from the *NATION* newspaper, National Information Services Agency (NISA)

- **Languages**: *crs*
- **Links**: http://nation.sc/

temoignages - Articles from the *Témoignages* newpaper

- **Languages**: *rcf*
- **Links**: https://www.temoignages.re/chroniques/ote/

Wikidumps

- **Languages**: *jam, gpe, gcr, srn*
- **Links**: https://huggingface.co/datasets/graelo/wikipedia/viewer

Wikimedia

- **Languages**: *srn*
- **Links**: https://archive.org/details/srnwiki-20180101