Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

CreoleVal

Multilingual Multitask Benchmarks for Creoles

Lent, Heather; Tatariya, Kushal; Dabre, Raj; Chen, Yiyi; Fekete, Marcell; Ploeger, Esther; Zhou, Li; Heje, Hans Erik; Kanojia, Diptesh; Belony, Paul; Bollmann, Marcel; Grobol, Loïc; de Lhoneux, Miryam; Hershcovich, Daniel; DeGraff, Michel; Søgaard, Anders; Bjerva, Johannes

DOI (link to publication from Publisher):
10.48550/arXiv.2310.19567

Creative Commons License
CC BY-NC-SA 4.0

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Lent, H., Tatariya, K., Dabre, R., Chen, Y., Fekete, M., Ploeger, E., Zhou, L., Heje, H. E., Kanojia, D., Belony, P., Bollmann, M., Grobol, L., de Lhoneux, M., Hershcovich, D., DeGraff, M., Søgaard, A., & Bjerva, J. (2023). *CreoleVal: Multilingual Multitask Benchmarks for Creoles*. arXiv. https://doi.org/10.48550/arXiv.2310.19567

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: February 10, 2025

arXiv:2310.19567v1 [cs.CL] 30 Oct 2023

CreoleVal: Multilingual Multitask Benchmarks for Creoles

Heather Lent[1], **Kushal Tatariya**[2], **Raj Dabre**[3], **Yiyi Chen**[1], **Marcell Fekete**[1], **Esther Ploeger**[1], **Li Zhou**[4,6], **Hans Erik Heje**[1], **Diptesh Kanojia**[5], **Paul Belony**[7], **Marcel Bollmann**[8], **Loïc Grobol**[9], **Miryam de Lhoneux**[2], **Daniel Hershcovich**[4], **Michel DeGraff**[10], **Anders Søgaard**[4], **Johannes Bjerva**[1]
[1]Aalborg University, Denmark, [2]KU Leuven, Belgium,
[3]National Institute of Information and Communications Technology, Japan,
[4]University of Copenhagen, Denmark, [5]University of Surrey, UK,
[6]University of Electronic Science and Technology of China, China,
[7]Kean University, USA, [8]Linköping University, Sweden,
[9]Université Paris Nanterre, France, [10]Massachusetts Institute of Technology, USA

Abstract

Creoles represent an under-explored and marginalized group of languages, with few available resources for NLP research. While the genealogical ties between Creoles and other highly-resourced languages imply a significant potential for transfer learning, this potential is hampered due to this lack of annotated data. In this work we present CREOLEVAL, a collection of benchmark datasets spanning 8 different NLP tasks, covering up to 28 Creole languages; it is an aggregate of brand new development datasets for machine comprehension, relation classification, and machine translation for Creoles, in addition to a practical gateway to a handful of preexisting benchmarks. For each benchmark, we conduct baseline experiments in a zero-shot setting in order to further ascertain the capabilities and limitations of transfer learning for Creoles. Ultimately, the goal of CREOLEVAL is to empower research on Creoles in NLP and computational linguistics. We hope this resource will contribute to technological inclusion for Creole language users around the globe.

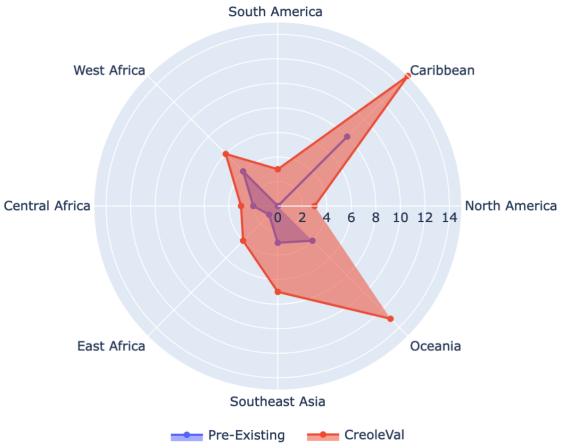

Figure 1: CREOLEVAL expands the availability of labeled data for Creoles around the globe. This chart shows the increased availability of datasets for concrete tasks, across Creoles from different regions. Before CREOLEVAL, only 11 Creoles had data for at least 1 pre-existing task, and now 28 Creoles have labeled data for at least 1 task and at most 6 tasks.

1 Introduction

Despite efforts to extend advances in Natural Language Processing (NLP) to more languages, Creoles are markedly absent from multilingual benchmarks. As such, progress towards reliable NLP for Creoles remains impeded, and consequently there is a dearth of language technologies available for the hundreds of millions of people who speak Creoles around the world. The omission of Creoles from such benchmarks can be attributed to two key factors: modality and stigmatization. The first, modality, is a notable factor as some Creoles are rarely used in writing, and thus text-based NLP is largely moot, highlighting a need for efforts in speech technology for Creoles. The latter, stigmatization, is perhaps the most salient of the two, however. As the history of many Creole languages is intricately interwoven with broader Western imperialism, colonialism, and slavery, Creole languages are often subjected to the stigmas and prejudices stemming from these historical atrocities (Alleyne, 1971; DeGraff, 2003).

On the surface, social prejudices against Creoles may seem extraneous in the context of NLP. However, the consequences of this stigmatization are palpable in preventing data collection for these languages. For example, it can be greatly challenging to collect data for a language without official status in a given country, even if it is the most widely used language by the populace; common

sources for language data like government documentation, educational materials, and local news may not be available. Moreover, even if a Creole is someone's primary language, sociolinguistic barriers[1] rooted in stigma may further prevent people from using it in various contexts, making opportunities for gathering data even more sparse. Lastly, even when financial resources are available to compensate crowd-workers, logistical challenges can significantly impede data collection efforts for Creole languages (Hu et al., 2011).

Stigmatization of Creoles is also an ongoing issue in the scientific domain, which further inhibits work in NLP. Indeed, prejudice against Creoles is deeply ingrained in linguistics, manifested in the common misconception that Creoles are incomplete or under-developed languages, in direct opposition to concepts like linguistic relativism and Universal Grammar (Kouwenberg and Singler, 2009; Aboh and DeGraff, 2016). The *Othering* of Creoles which has occurred in linguistics has led to a research landscape where Creole languages are typically categorized as *exceptions* amongst languages, and thus separated from other languages. Take for example the widely-used WALS database, which lists Creoles as having the language family "other"; works in NLP or computational linguistics relying on WALS to sample languages from diverse range of families as a part of their methodology typically exclude Creoles from their work.[2] And while other such resources exist to specifically cater to Creoles (e.g., APICS (Michaelis et al., 2013)), the creation of entirely *separate* resources to specifically accommodate Creoles is emblematic of their ghettoization within scientific spaces. For readers interested in the prominent debates about Creoles in linguistics, we defer to DeGraff (2005).

Inclusion of Creoles In an effort to enable NLP research on Creoles, we introduce CREOLEVAL, a set of benchmarks covering a wide variety of tasks for up to 28 Creole languages. Enabling NLP research on Creoles offers significant possibilities. First, this will enable development of language technologies for Creoles, potentially improving technological inclusion of the speakers of these languages. While increasing the number of NLP datasets for Creoles is important, an crucial note here is that as set of languages, Creoles are not a monolith. In some contexts, a Creole can be someone's mother tongue, and the sole language they speak; in other cases, Creoles can play an important role as a lingua-franca within linguistic diverse communities, and for this reason, deserve special attention of the NLP community (Bird, 2021). Due to their status as marginalized[3] languages, we highlight the importance of community involvement when designing CREOLEVAL. Inspired by recent recommendations on participatory machine learning (Sloane et al., 2022), we build on previous work by Lent et al. (2022b), and attempt to strike a balance by creating resources that can be beneficial for both Creole-speaking communities and the NLP community. Creating the technologies explicitly sought after by various Creole-speaking communities remains an open area for future work, and we believe the that the benchmarks and baselines in CREOLEVAL can be useful to this end. Second, from a scientific perspective, we argue that Creoles offer an opportunity for careful development and evaluation of transfer learning methods, e.g., leveraging similarities to a Creole's ancestor languages. For example, consider Chavacano, a language spoken in the Philippines with genealogical ties to Spanish, Tagalog, and other languages. Below is a sample sentence (Steinkrüger, 2013) in Chavacano, with an accompanying Spanish and English translation, annotated with Subject, Verb, and Object roles:

- Chavacano: "Ya-mirá$_V$ el mga ómbre$_S$ un póno de ságing$_O$."
- Spanish: "Los hombres$_S$ vieron$_V$ un árbol de plátano$_O$."
- English: "The men$_S$ saw$_V$ a banana tree$_O$."

While Chavacano shares some vocabulary with Spanish, it grammatically maintains the VSO word order of Tagalog. Hence, from a transfer learning perspective, one could expect that transfer from Spanish could be useful in terms of lexical overlap, but not syntax. As many Creoles are genealogically related to other higher-resourced

[1] In some Creole-speaking communities, the local Creole language is viewed as a "corrupted" version of the historically related European language, with names like "broken English". Thus, speakers of such varieties would not even identify their variety as a separate language.

[2] Newer typological databases such as Grambank have fortunately addressed this issue.

[3] Notably, a handful of Creoles do have official language status by law in their respective lands: Haitian Creole, Seychelles Creole, Bislama, and Sango.

languages (e.g. English, French, Spanish, Portuguese, Dutch), resource availability permits research on Creoles that can help shed light on the underlying mechanics of transfer learning. To this effect, the baselines presented in this work pertain to zero-shot transfer learning, in order to ascertain the current viability of transfer learning for Creoles. Ultimately, the goal of CREOLEVAL is to facilitate research on transfer learning, computational linguistics, as well as general linguistic research on Creole languages. By providing this resource, we hope that inclusion of Creoles in multilingual evaluations will become a default practice in NLP.

Contributions In this work, we introduce new datasets for three different NLP tasks (machine comprehension, relation classification, and machine translation) for understudied Creole languages. We expand the scope of CREOLEVAL by packaging these new datasets together with pre-existing tasks for Creoles (i.e., dependency parsing, named entity recognition, sentiment analysis, sentence matching, natural language inference, and machine translation), resulting in a publicly available repository. This repository facilitates further work on Creoles for the NLP community, as we provide a single gateway to this diverse group of languages, allowing for straight-forward data exploration, experimentation, and evaluation. The 28 Creole languages covered in CREOLEVAL are, unfortunately, unequally represented across tasks due to the difficulties of gathering and curating data. However, the addition of our new development data greatly expands upon the existing number of NLP tasks for Creoles (see Figure 1). For all the datasets comprising CREOLEVAL, we present baseline experiments with additional analysis on the efficacy of transfer learning for Creoles. Our code, data, documentation, and models are available at the public repository.[4] Where we cannot provide data for copyright reasons (i.e., Bible data), we provide detailed documentation and code to allow for reproducibility.

2 Background

Previous Work Prior works in NLP primarily focus on individual Creole languages, such as Antillean Creole (Mompelat et al., 2022), Jamaican Creole (Armstrong et al., 2022), Mauritian Creole (Dabre and Sukhoo, 2022), Nigerian Pidgin (Ogueji and Ahia, 2019; Caron et al., 2019; Oyewusi et al., 2020; Adelani et al., 2021; Muhammad et al., 2022, 2023), Singlish (Wang et al., 2017; Liu et al., 2022), and Sranan Tongo (Zwennicker and Stap, 2022).[5] A few works specifically investigate Creoles as a collection of languages, with interest in LMs (Lent et al., 2021) and transfer learning (Lent et al., 2022a). Lent et al. (2022b) further discusses some of the social aspects to consider for responsible NLP for Creoles, due to the languages' stigmatization (Alleyne, 1971; Siegel, 1999; Kouwenberg and Singler, 2009).

Transfer Learning Transfer learning represents the ability of a model to perform reasonably well over data *outside* the scope of the original training data (Zhuang et al., 2019). In NLP, transfer learning has thus been effective for extending models from higher-resourced languages to lower-resourced ones, especially when the languages in question have similar genealogy, typology, and script (Pires et al., 2019; Wu and Dredze, 2019; Nooralahzadeh et al., 2020; Zhao et al., 2021; de Vries et al., 2021, 2022). In the context of Creoles however, some initial research suggests that transfer-learning from genealogically related languages may not be entirely straightforward. de Vries et al. (2022) investigate the most effective language pairs for transfer learning of part-of-speech (POS) tagging; while this work does not outright focus on Creoles, a notable finding is that Swedish – not English nor Portuguese – is the most useful language for transferring POS tags to Nigerian Pidgin. Moreover, in a direct investigation of transfer learning for Creoles, Lent et al. (2022a) found that LMs trained on multiple ancestor languages failed to transfer well to Creoles on limited downstream tasks. Further investigation is required to understand why both the aforementioned studies obtained seemingly counter-intuitive results. However, other work investigating the underlying mechanisms that empower transfer learning have indicated that the success of transfer learning may be less dependent on genealogical language relatedness, and more dependent on other factors like sub-word overlap (Pelloni et al., 2022).

[4] www.github.com/hclent/CreoleVal

[5] See https://creole-nlp.github.io/ for a comprehensive list of datasets for Creoles.

	Data	#Lang	#Creole	#Anc
mBERT	Wikipedia	104	0	6
XLM-R	CC100	100	0	6
mT5	CC4	101	1	6
mBART-50	custom	50	0	5

Table 1: Coverage of total **Lang**uages, **Creole**s, and their **Anc**estor languages in training data for popular multilingual LMs. For mt5, 0.33% of the training data comes from Haitian Creole. mBART-50 is trained on the same 25 languages from XLM-R and an additional 25 languages from regular mBART (Liu et al., 2020). While we do not experiment with BLOOM (Scao et al., 2023), it can be noted that 0.0002% of the Big Science Corpus contains Lingala, a Creole related to Bantu.

Multilingual Language Models Selecting a pertinent language model (LM) is typically the first step for any attempt at transfer learning. Creoles, however, are largely absent from the most commonly used multilingual LMs (see Table 1). For this work, we choose to work with mBERT (Devlin et al., 2019), XLM-Roberta (Conneau et al., 2020), mT5 (Xue et al., 2021) for natural language understanding tasks, and mBART-50 (Tang et al., 2020) for generation tasks. Despite an ostensible lack of coverage for Creoles, these models do include relevant pre-training data for some genealogically related languages.

3 Natural Language Understanding of Creoles

Tasks across natural language understanding (NLU) test a model's capacity for grasping syntax and semantics. Typical tasks, such as sentiment analysis and named entity recognition, require sizeable amounts of training data is for models to exhibit decent performance. In order to expand on the availability of NLU data for Creoles we introduce two brand new benchmark datasets for machine comprehension and relation classification, before experimenting with a set of pre-existing NLU tasks for Creoles. Our baselines are in a zero-shot transfer learning for Creoles, as this is the most typical setup for working with languages with little to no data.

3.1 Machine Comprehension

Most pre-existing NLU tasks for Creoles largely examine syntax (see Section 3.3), and there is a dearth of NLU tasks for Creoles that evaluate semantic understanding. As curating naturally occurring language data for a new task is often prohibitively expensive, dataset translation is a typical alternative, though translation can be complicated by cultural differences between the source and target audience (Hershcovich et al., 2022). In this work, we translate MCTest, a machine comprehension dataset introduced by Richardson et al. (2013), as it pertains to a semantically oriented task, and as the general domain and smaller data size make translation feasible. Machine comprehension is an NLU task where a model ischallenged to correctly answer questions contingent to a specified piece of text. The MCTest dataset is composed of short stories intended for school-aged children, each accompanied with four multiple choice questions, that require different levels of reasoning to answer (i.e., context from one or multiple sentences is needed for a human to successfully answer the question).

	mBERT	XLM-R
Haitian-direct	51.60%	39.16%
Haitian-localized	50.83%	43.33%
Mauritian	49.10%	43.33%
English	63.33%	45.00%

Table 2: Accuracy results for MCTest160 development data, when trained on the English MC160 training data.

Translation We chose to translate the MC160 development set because of the relatively general domain, and smaller size, which makes it feasible for translation (30 stories, 120 questions). We hired professional translators, to translate the English MC160 development set, into both Haitian Creole and Mauritian Creole. Although we had budget for even more translations, these were the only two Creole languages that we could find translators for. Notably, there are two different translations for Haitian Creole: a direct translation, and a localized translation. As opposed to the direct translation, the localized version is a culturally-sensitive translation, with minor changes to include names, places, and activities that are directly pertinent to a Haitian audience (Roemmele et al., 2011; Hershcovich et al., 2022). For example, the original English dataset may discuss an ice cream truck (directly translated to "*kamyon krèm*"), though ice cream is not a typical desert in Haiti; thus in the localized dataset, "ice cream truck" has been changed to "*machann*

Dataset	Sent. Enc.	bert-base-multilingual-cased				xlm-roberta-base			
	Rel. Enc.	Bb-nli	Bl-nli	Xr-100	Xr-b	Bb-nli	Bl-nli	Xr-100	Xr-b
Dev (en)		66.70±1.89	74.01±3.11	71.34±1.27	71.47±2.90	56.34±1.14	62.86±1.70	52.42±0.43	51.78±0.
bi		22.40±2.47	21.67±6.77	24.53±11.28	27.53±12.26	13.75±1.94	17.78±3.46	10.63±5.08	11.64±4.62
cbk-zam		30.00±1.32	31.66±5.63	28.78±3.65	29.42±2.08	14.77±2.16	11.78±4.22	13.37±1.17	15.93±3.90
jam		14.57±0.81	11.28±1.63	16.33±0.63	16.40±0.94	5.42±0.55	9.92±4.03	10.52±1.60	9.47±3.01
phi		14.78±1.30	9.09±2.71	12.55±1.76	12.09±2.22	11.60±4.16	12.00±0.99	10.15±4.88	7.91±2.42
tpi		16.64±6.25	22.63±1.67	22.09±5.50	20.67±5.41	15.84±4.20	17.11±1.83	13.49±2.30	12.55±0.56
AVG		19.68	19.27	20.86	**21.22**	12.28	**13.72**	11.63	11.50

Table 3: Relation Classification performance measured by macro F1 score on English validation (dev) set and Creole test sets. AVG shows the overall performance per setup across all Creole languages. **Bold** indicates the best performance for each sentence encoder setting. Sent. Enc.: sentence encoder. Rel. Enc.: relation encoder.

fresko", a cart which sells a shaved-ice desert enjoyed in Haiti. We hope that these two different Haitian Creole datasets for machine comprehension, can also be useful in evaluating progress in cross-cultural NLP.

Results and Analysis For our benchmark experiments on the Creole MCTest160 development set, we use a simple transformer-based baseline approach, leveraging mBERT and XLMR as the basis of these models. We finetune them for 10 epochs over the English MCTest160 training set. A summary of our results are in Table 2, with full results and hyperparameter settings documented in the accompanying Github repository. mBERT outperforms XLMR, although XLMR performs better over the localized data than the direct translation for Haitian. The performance on Haitian and Mauritian is surprising, as both mBERT and XLMR have seen no Creole data. It's particularly noteworthy that mBERT results on Creoles outperforms XLM-R's English performance by far. In comparison, a random baseline on MCTest160 yields an accuracy of 25%, and Attentive Reader (Hermann et al., 2015) has an accuracy of 42% on English data.

3.2 Relation Classification

Relation classification (RC) aims to identify semantic associations between entities within a text, essential for applications like knowledge base completion (Lin et al., 2015) and question answering (Xu et al., 2016). In this work, we introduce the first manually-verified RC datasets for five Creole languages: Bislama, Chavacano, Jamaican Patois, Pitkern, and Tok Pisin.

Our dataset is sourced from Wikipedia, where we found 16 Creoles with a presence, though only 9 had readily-available Wikidumps.[6] Many Creole Wikipedia entries are short and *templatic*, likely due to machine generation. This templatic nature, however, facilitates the creation of a relation classification dataset, as it allows for easy identification of entities and relations.

To construct the dataset, we preprocess[7] Wikipedia dumps and perform automatic entity linking using OpenTapioca (Delpeuch, 2019). Sentences are then clustered based on latent templates to facilitate manual annotation. Despite not being native speakers of the Creole languages, our familiarity with their ancestor languages (English, French, Spanish) assisted annotation. While Creoles are in no way mutually intelligible with these related languages, when examining a group of sentences with the same latent template, familiarity with the ancestor languages helped in identifying entities. The process resulted in high-quality evaluation data for 5 of the 9 initially identified Creole Wikipedias. Each dataset contains 97 evaluation samples[8].

We establish a benchmark for Creole RC using a zero-shot cross-lingual transfer approach: we employ pre-trained language models (LMs) that have not been exposed to Creole data and train exclusively on English data.

Model and Training We adopt the method introduced by Chen and Li (2021), which excels in zero-shot transfer learning for RC on Wikipedia and Wikidata (Han et al., 2018). This approach projects both sentences and their associated relation descriptions into a shared embedding space, minimizing distances between them while performing classification. For training, we use the UKP dataset (Sorokin and Gurevych, 2017),

[6] bi, cbk-zam, gcr, hat, jam, pap, pih, sg, tpi

[7] https://github.com/attardi/wikiextractor

[8] For complete dataset statistics, and further discussion on the templates, see the repository.

Task	Language	Dataset	Metric	mBERT	XLM-R	mT5
UDPoS	pcm	UD_Naija-NSC (Caron et al., 2019)	Acc	0.98	**0.98**	0.98
	singlish	Singlish Treebank (Wang et al., 2017)	Acc	0.91	**0.93**	0.91
NER	pcm	MasakhaNER (Adelani et al., 2021)	Span-F1	0.89	0.89	**0.90**
	bi	WikiAnn (Pan et al., 2017)	Span-F1	**0.94**	0.90	0.72
	cbk			**0.96**	0.96	0.94
	ht			0.78	**0.84**	0.48
	pih			**0.90**	0.88	0.61
	sg			0.89	**0.93**	0.79
	tpi			**0.91**	0.89	0.75
	pap			**0.90**	0.89	0.85
Sentiment Analysis	pcm	AfriSenti (Muhammad et al., 2023)	Acc	0.66	**0.68**	0.67
	pcm	Naija VADER (Oyewusi et al., 2020)	Acc	0.71	**0.72**	0.72
NLI	jam	JamPatoisNLI (Armstrong et al., 2022)	Acc	0.74	**0.76**	0.66
Sentence Matching	cbk-eng	Tatoeba (Artetxe and Schwenk, 2019)	Acc	**15.9**	3.9	6.5
	gcf-eng			**12.8**	4.9	6.9
	hat-eng			23.9	18.5	**37.9**
	jam-eng			**19.9**	9.6	10.3
	pap-eng			**22.4**	6.1	15.9
	sag-eng			5.7	2.1	**7.3**
	tpi-eng			7.2	3.3	**7.6**

Table 4: Baseline scores for pre-existing NLU tasks for Creoles. Additional experiments, results, and analysis are included in the CreoleVal repository's documentation.

which contains 108 Properties (i.e., relations in Wikidata). In contrast, our Creole datasets feature just 13 Properties, four of which are not present in the UKP dataset. Five relations are separated for validation. We fine-tune multilingual models mBERT and XLM-R (Conneau et al., 2020) using multilingual sentence transformers (Reimers and Gurevych, 2019). The sentence encoder employs mBERT and XLM-R,[9] while the relation encoder uses one of four alternative models, denoted Bb-nli, Bl-nli, Xr-b, Xr-100[10] here, as sentence embeddings of the relation descriptions from Wikidata.

Results and Analysis Table 3 illustrates the performance of RC in each setting. We observe notably worse performance in Creole languages compared to English. This highlights the particular challenge of leveraging pretrained LM's for zero-shot cross-lingual transfer for RC for Creoles, due to the lack of representation of Creoles in the LM training data. In addition, the choice of the sentence encoder is a primary determinant of performance of Creole RC. When using mBERT as the sentence encoder, the performance of Creole RC tends to be slightly better than XLM-R. Under the same sentence encoder, different relation encoders exhibit slight variations in performance.

3.3 Prior NLU Benchmarks

In addition to the datasets that we introduce, there are a handful of pre-existing, labeled datasets for Creole languages in the space of NLU. In order to facilitate concentrated efforts on Creole NLP, we have gathered these tasks and packaged the baseline experiments for them with the CreoleVal repository. For each of these prior benchmarks, we provide code to run baseline experiments with three multilingual LMs (mBERT, XLM-R and mT5). We compare performance on the test set for each task and LM in Table 4.

4 Natural Language Generation of Creoles

Unlike NLU, where the model aims to predict an accurate label, natural language generation (NLG) is arguably a more challenging task as models should generate output that is *adequate* as well as *fluent*. A lack of data – both in terms of size and domain – further complicates NLG for Creole languages. In this paper, we introduce 2 new machine translation (MT) datasets for Creoles. The first covers 26 Creoles with text drawn from the religious domain, and the second is a small, but very high quality, Hatian Creole dataset in the educational domain. We also conduct experiments and evaluate performance on a pre-existing MT dataset for Mauritian Creole.

[9] Respectively, `bert-base-multilingual-cased`, `xlm-roberta-base`.

[10] Respectively,`bert-base-nli-mean-tokens`, `bert-large-nli-mean-tokens`, `xlm-r-bert-base-nli-mean-tokens`, `xlm-r-100langs-bert-base-nli-mean-tokens`.

4.1 CreoleM2M MT

As the world's most translated text, the Bible is a typical starting point for gathering language data in a low-resource scenario. While Bible data has a number of limitations (e.g., fixed domain, archaic language, and translationese (Mielke et al., 2019)), notable benefits include its size and parallelism with other languages, which lends itself aptly to MT. We gathered parallel corpora for 26 Creole Bibles from Mayer and Cysouw (2014),[11] along with additional texts from the JW300 corpus (Agić and Vulić, 2019). In total, our parallel MT corpus contains 3.4M sentences and 71.3M and 56.3M Creole and English words, respectively, making it the largest Creole parallel corpus to date. Furthermore, we split from the Bible part of the corpus, 1,000 and 2,000 sentences for each Creole and English and use them for development and testing, respectively. Note that the development and test sets are N-way parallel (N=27: 26 Creoles and English). We ensured that there is no overlap between the training, development, and test data.

4.1.1 Experiments

We fine-tune mBART-50-MT (Tang et al., 2020) and also train models mBART from scratch, over the parallel Bible text.

Vocabulary For models trained from scratch, we use the training data and create a shared tokenizer of 64,000 subwords for all 26 Creoles and English using *sentencepiece* (Kudo and Richardson, 2018). Due to the large number of languages, we only train bilingual models and leave multilingual models for future work. While we could have created separate vocabularies for bilingual models, a shared tokenizer will be helpful in ensuring consistency with future planned multilingual model experiments. For the fine-tuned models, we use the mBART-50 tokenizer containing 250,000 subwords. Although this tokenizer's vocabulary was not explicitly trained on Creoles, we expect the subwords from related parent languages to be sufficient.

Training We trained our models using the YANMTT toolkit[12] (Dabre and Sumita, 2021), which supports training models from scratch as well as by fine-tuning mBART models. In this paper, we train models from scratch as well as by fine-tuning the mBART-50-MT model[13] as done earlier by Dabre and Sukhoo (2022). The training utilizes Adam optimizer (Kingma and Ba, 2014), and trains till convergence. We evaluate the training performance on the development set using BLEU score as a metric after every 1,000 training steps. The training process determines convergence when BLEU scores do not improve for 20 consecutive evaluations.[14]

Decoding We perform decoding using beam search with a beam of size 4 and a length penalty of 0.8. Due to the large number of language pairs, we do not tune these parameters for each language pair.

Results and Analysis Figure 2 shows the performance in terms of chrF and BLEU scores for Creole to English and English to Creole translation for the test set of the CreoleM2M benchmark. For models trained from scratch, it is clear that the performance is correlated with the size of the parallel corpus. Therefore, fine-tuning the mBART-50-MT model leads to significant improvements in translation quality by up to 19.2 BLEU and 17.3 chrF for Creole to English translation and up to 16.9 BLEU and 13.5 chrF for English to Creole translation. We noted that both BLEU and chrF scores are correlated[15] with each other. It is important to note that fine-tuning is not always a good idea for the Creoles with more training data available. In most larger-resourced settings, we observed a reasonable drop in translation quality, indicating that the fine-tuned model converges too quickly; and is unable to learn well from the training data.

4.2 MIT-Haiti MT

While Bible translations can provide initial data for training MT systems, this domain is markedly limited, highlighting a need for MT datasets for Creoles originating from other, more generalizable domains. To this end, we introduce the BANK

[11] To access the raw Bible corpora, one must request the authors due to copyright issues.

[12] https://github.com/prajdabre/yanmtt

[13] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[14] Note that we anneal the learning rate by half when the BLEU scores don't improve for 10 consecutive evaluations and then again by half if the scores don't improve for 15 consecutive evaluations. Therefore, after cutting the learning rate by half (each time) for the final convergence decision, we wait for 20 consecutive evaluations to declare model convergence.

[15] We calculated a Pearson correlation score of 0.98.

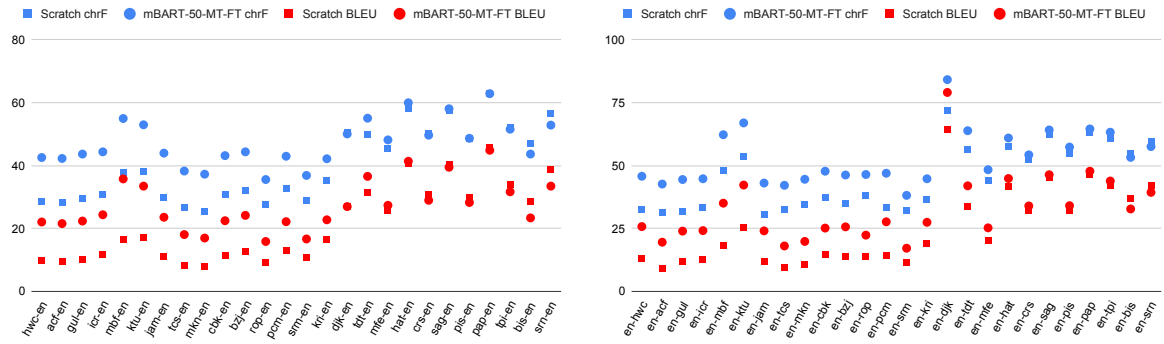

Figure 2: chrF (blue color) BLEU (red color) scores obtained using baseline models (scratch; square points) and fine-tuned models (mBART-50-MT-FT; circle points) on the Bible corpus for XX-En (left) and En-XX (right) language pairs, where XX represents Creole languages. The language pairs are ordered from left to right in increasing sizes of parallel corpora from 4,366 pairs to 583,746 pairs

DONE MIT-AYITI, or in English, the MIT-Haiti Corpus: a manually-verified, high-quality collection of parallel Haitian Creole sentences with English, French, and Spanish translations. This data comes from Platform MIT-Haiti[16], a learning platform with educational material for students in Haitian Creole. We scrape the entire website, including the web text and PDFs. The parallel sentences for this MT corpus come from 60 multilingual stories (the PDFs and their converted plain text transcriptions); these stories were each manually cleaned and corrected (i.e., in cases where the PDF-reader made mistakes in transcribing, these were manually corrected), aligned, and verified by a subset of the authors, who have qualifications in both linguistics and NLP. For the remaining monolingual Haitian text without direct parallel translations, we manually clean and verify these sentences with the same process, and release a small set of monolingual examples (∼8200 utterances), which could potentially be useful for few-shot continued pre-training of a language model. Although this dataset is relatively small, we would like to stress that it is high quality, as it comes directly from a community that actively fosters education and writing in Haitian Creole.

OPUS for MIT-Haiti To establish the baseline performance on the MIT-Haiti Corpus, we leverage pre-trained OPUS-MIT models (Tiedemann and Thottingal, 2020). In Table 5, we show the performance of pre-trained OPUS-MT models on the MIT-Haiti benchmarks. These models were previously benchmarked on the Tatoeba and/or JW300 corpus, which are limited in complexity and domain, respectively. By extending this to the MIT-Haiti Corpus, we can gain an insight into the performance of these models on more diverse usage of Haitian Creole. We translate from Spanish, French and English into Haitian Creole, because this translation direction has the potential to be useful for (monolingual) speakers of Haitian Creole, as it provides increased information access. Notably, the scores on the MIT-Haiti benchmarks are considerably lower than those on previous benchmarks. For instance, the English to Haitian Creole model scores 45.2 BLEU and 59.2 chrF on the Tatoeba test set[17], while it retrieves only 14.7 BLEU and 35.8 chrF on the MIT-Haiti Corpus. This suggests that previous benchmarks are likely to be overly optimistic.

CreoleM2M for MIT-Haiti Table 5 contains the results for the fine-tuned CreoleM2M models on the MIT-Haiti Corpus. We can see that the BLEU and chrF scores are 18.6/38.1 and 22.0/43.9 for Haitian Creole to English and English to Haitian Creole, respectively. Despite the domain differences between CreoleM2M's training data (religion) and the MIT-Haiti benchmarks (education), a brief manual inspection revealed that the translation quality is not particularly bad, however the generated translations tend to contain spurious religious content. Extensive human evaluation of these translations will help in better understanding of the limitations of our CreoleM2M models in a cross-domain setting.

[16] https://mit-ayiti.net/

[17] https://huggingface.co/Helsinki-NLP/opus-mt-en-ht

model	source	target	# lines	BLEU	chrF
OPUS	es	ht	102	12.1	32.9
	fr	ht	1,503	11.8	33.5
	en	ht	1,559	14.7	35.8
CreoleM2M	en	ht	1,559	22.0	43.9
	ht	en		18.6	38.1

Table 5: Performance of OPUS models (opus-mt-en-ht, opus-mt-es-ht, opus-mt-fr-ht) on our MIT-Haiti Corpus benchmarks, as well as the results of decoding the MIT-Haiti benchmarks using the fine-tuned CreoleM2M Haitian Creole models.

4.3 Prior NLG Benchmarks

KreolMorisienMT (Dabre and Sukhoo, 2022) is a dataset for machine translation of Mauritian Creole (i.e., Kreol Morisien) to and from English and French. The dataset spans multiple domains spanning the Bible, children's stories, commonly used expressions and some books. We refer the reader to Dabre and Sukhoo (2022) for further details. In this paper, we focus only on translation to/from English. We combine the training data from the Kreol Morisien part of the CreoleM2M dataset with KreolMorisienMT's training data and then train MT models to show the impact of our newly mined data. We filter out those sentences from CreoleM2M, which are present in the development and test sets of KreolMorisienMT, for clean evaluation. This gives us 188,820 sentence pairs, which is almost an order of magnitude larger than the 21,810 sentence pairs in KreolMorisienMT. As a baseline, we only train models with the CreoleM2M data containing 167,010 sentence pairs after removing the development and test set sentences of KreolMorisienMT.

For the KreolMorisienMT test set, since it is standalone, we focus on standalone bilingual models and hence create a filtered version of the Kreol Morisien part[18] of CreoleM2M's training data. We use this to train separate tokenizers of 16,000 subwords for Kreol Morisien and English. One tokenizer is with this filtered version alone, and one is with a combination of the filtered version and the training data of KreolMorisienMT.

Table 6 contains results for the test set of KreolMorisienMT. We compare our models trained from scratch and fine-tuning against those of Dabre and Sukhoo (2022). The most important thing to note is that our scratch models are overwhelmingly better than corresponding models by Dabre and Sukhoo (2022). In fact, we see gains of up to 9.4 BLEU[19]. On the other hand, the filtered CreoleM2M data when used for fine-tuning, despite its size, does not lead to a model that surpasses Dabre and Sukhoo (2022)'s corresponding model that is fine-tuned on a much smaller KreolMorisienMT training dataset. However, by combining both the filtered CreoleM2M and KreolMorisienMT training datasets, we finally surpass Dabre and Sukhoo (2022)'s best results.

Other We exclude **PidginUNMT** (Ogueji and Ahia, 2019), as this unlabeled dataset pertains unsupervised machine translation. We also exclude **WMT11** (Callison-Burch et al., 2011), as this dataset was created to help victims of the 2010 earthquake in Haiti, and thus contains sensitive data.

5 Discussion and Recommendations

Implications for Transfer Learning The introduction of CREOLEVAL marks a significant step forward in bridging the technological divide for Creole languages, in the context of NLP. Prior to this work, the scarcity of resources for Creoles made progression of NLP tailored for Creole speakers close to impossible. Now, as shown in Figure 1, 28 Creole languages, which previously had limited or no NLP datasets, are now part of a unified platform. This platform enables researchers and developers to easily include Creoles in pre-existing pipelines, introducing a novel and unique low-resource scenario to NLP. Given the genealogical ties of many Creoles to (typically) higher-resourced languages[20], we expect this to allow for nuanced experimentation in transfer learning. In particular, the complex picture of Creoles, including both horizontal and vertical transfer between diverse languages, may offer the key to developing transfer learning techniques which are tuned to encapsulate specific pieces of cross-linguistic knowledge. While vocabulary might be transferred from a parent language, syntactic and semantic structures may diverge, challenging con-

[18] As mentioned in Section 4.3, we filter to remove the KreolMorisienMT test set sentences from CreoleM2M's training data.

[19] Dabre and Sukhoo (2022) do not give chrF scores in their paper and do not release their translations, making it impossible for us to compare chrF scores

[20] Some Creoles have strong genealogical ties to lower-resourced languages, such as the Niger-Congo Creoles Lingala, Kikongo-Kituba, Fanakalo, which are related to Bantu languages, and Sango, which is related to Ngbandi.

Data	Model	BLEU		chrF	
		mfe-eng	eng-mfe	mfe-eng	eng-mfe
Dabre and Sukhoo (2022)	Scratch	11.1	11.5	-	-
Dabre and Sukhoo (2022)	mBART-50-MT-FT	24.9	22.8	-	-
CreoleM2M	Scratch	16.1	11.5	38.0	37.1
CreoleM2M+KreolMorisienMT	Scratch	20.5	16.9	42.8	41.1
CreoleM2M	mBART-50-MT-FT	22.1	18.9	44.6	44.4
CreoleM2M+KreolMorisienMT	mBART-50-MT-FT	25.7	24.7	47.8	48.2

Table 6: Results on the KreolMorisienMT test sets by using CreoleM2M training data, in addition with the training data in KreolMorisienMT.

ventional transfer learning methods. Indeed, previous work has shown the difficulties of straightforward transfer learning techniques from ancestor languages (Lent et al., 2022a). We suggest that the success of transfer learning in this new domain relies on in-depth understanding of the structural and contextual intricacies of each individual Creole language, rather than a simplistic reliance on their parent languages. Moreover, we believe that work to this end has the potential to improve transfer learning methodology, as it will help researchers gain a broader understanding of the capabilities and limitations of transfer learning. Finally, beyond strict transfer learning, we also expect cultural adaptation to be a significant challenge for the future, for which CREOLEVAL provides a benchmark.

Further Resource Development While CREOLEVAL opens for straightforward inclusion of a set of Creole languages in NLP pipelines, we are still limited to textual data. While this is an important contribution which may lead to a more even playing field in terms of language technologies, it is not enough to focus on this modality. Considering the fact that many Creoles are exclusively *spoken* languages indicates that a focus on speech resource development is an important next step.

Recommendations For future work on Creole languages, be it in the context of experimentation on CREOLEVAL, or on further resource development, we recommend the following:

1. Engage with language communities. When languages are limited in resources, it is critical that any new additional resources are allocated to efforts that will benefit the communities using the language in question (Bird, 2021). For Creoles, a concrete starting point is to reach out to experts, as discussed by Lent et al. (2022b).
2. Keep in mind contextual factors such as domain and culture. Direct translations in narrow domains are likely to introduce cultural biases, which may render language technology less relevant to potential end-users (Hershcovich et al., 2022). When it is not possible to gather naturally occurring language data, we echo similar recommendations by others for culturally sensitive translations (Roemmele et al., 2011).

6 Conclusion

In this work, we have addressed the absence of Creole languages from contemporary NLP research by introducing benchmarks and baselines for a total of 28 Creole languages. We argue that this omission in previous work has hindered the progress of NLP technologies tailored to Creole-speaking populations, in addition to preventing research communities from exploring the unique linguistic situations of this diverse group of languages. With the introduction of CREOLEVAL, we have made a significant step towards bridging the gap between Creole languages and other low-resource languages in NLP. We hope that the public release of our datasets and trained models will serve as an invitation to further research in this relatively unexplored domain, and expect that NLP and computational linguistics research stand to gain significantly from embracing the linguistic and cultural diversity embodied in this group of languages.

Limitations

Although we are the first to create NLU and NLG benchmarks for up to 28 Creoles, we note the following limitations.

Limited domain diversity Although, we were able to collect reasonably large parallel corpora

for Creole MT, the data itself belongs to the religious domain and thus might not be extremely useful in a general purpose MT setting. Controversially, the Bible and other religious texts may be considered colonialist by some communities, as these texts may be used to "*provoke a culture change in these communities*" (Mager et al., 2023). However, works in domain adaptation (Chu et al., 2017) have shown that even a small amount of in-domain corpus may be sufficient for adapting our models to other domains.

Lack of reliable monolingual corpora sources Unlike resource-rich languages like English, French, and Hindi, finding monolingual corpora for Creoles is extremely difficult. One reason for this is the relatively recent interest in research on Creoles in NLP. The lack of monolingual corpora also inhibits the development of LLMs for Creoles, however even a tiny amount may be helpful for expanding existing LLMs, as shown by Yong et al. (2023).

No language identification tools A possible reason for the difficulty in obtaining Creole corpora from the web is that there are extremely limited language identification (LID) (Baldwin and Lui, 2010) tools for Creoles, and thus identifying Creole content in CommonCrawl[21] is also very difficult. Developing LID tools for Creoles will be an important future work.

Modality Many Creoles are spoken and not written, therefore text-based NLP might not be suited for them. This motivates branching out into speech-to-text (automatic speech recognition, speech translation) and speech-to-speech (translation) research.

Acknowledgments

HL, YC, MF, EP, HEH, and JB are funded by the Carlsberg Foundation, under the *Semper Ardens: Accelerate* programme (project nr. CF21-0454). KT and MDL, the computational resources and services used were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government - department EWI. MIT-Haiti is, in the main, internally funded by grants from Jameel World Education Lab[22] (for MDG). Some experiments were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Chalmers partially funded by the Swedish Research Council through grant agreement no. 2022-06725 (for MB). The translations of the MCTest dataset were funded by he European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199 (for HL) .

Contributions

We use CRediT (Contributor Roles Taxonomy https://credit.niso.org) to note the different roles undertaken by the authors:
Conceptualization AS, JB, HL;
Data curation HL, RD, YC, MF, HEH, PB;
Formal Analysis HL, KT, RD, YC, MF, EP, LZ, DK, MB, LG;
Funding acquisition JB, HL;
Investigation HL, RD, MDL, DH, MDG, AS, JB;
Methodology & Software HL, KT, RD, YC, MF, EP, LZ, HEH, DK, MB,LG;
Project administration HL;
Resources AS, JB, RD, MB, LG, MDG;
Validation HL, KT, RD, YC, LZ, MB;
Writing HL, KT, RD, YC, MF, EP, LZ, DK, MB, MDL, DH, MDG, JB.

References

Enoch Oladé Aboh and Michel DeGraff. 2016. A null theory of creole formation based on universal grammar. *The Oxford Handbook of Universal Grammar*.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa

[21] https://commoncrawl.org/
[22] https://www.jwel.mit.edu/

Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

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

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

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237, Los Angeles, California. Association for Computational Linguistics.

Steven Bird. 2021. LT4All!? rethinking the agenda.

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, Edinburgh, Scotland. Association for Computational Linguistics.

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, Paris, France. Association for Computational Linguistics.

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3470–3479, Online. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Raj Dabre and Aneerav Sukhoo. 2022. 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.

Raj Dabre and Eiichiro Sumita. 2021. YANMTT:

yet another neural machine translation toolkit. *CoRR*, abs/2108.11126.

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.

Antonin Delpeuch. 2019. Opentapioca: Lightweight entity linking for wikidata. *arXiv preprint arXiv:1904.09131*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

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

Chang Hu, Philip Resnik, Yakov Kronrod, Vladimir Eidelman, Olivia Buzek, and Benjamin B. Bederson. 2011. The value of monolingual crowdsourcing in a real-world translation scenario: Simulation using Haitian Creole emergency SMS messages. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 399–404, Edinburgh, Scotland. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

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.

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

Heather Lent, Emanuele Bugliarello, and Anders Søgaard. 2022a. Ancestor-to-creole transfer is not a walk in the park. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 68–74, Dublin, Ireland. Association for Computational Linguistics.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022b. 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.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph comple-

tion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

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.

Zhengyuan Liu, Shikang Ni, Ai Ti Aw, and Nancy F. 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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. 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)*. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

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

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, and Olalekan Akinsande. 2020. Semantic enrichment of nigerian pidgin english for contextual sentiment classification. *ArXiv*, abs/2003.12450.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Olga Pelloni, Anastassia Shaitarova, and Tanja Samardzic. 2022. Subword evenness (SuE)

as a predictor of cross-lingual transfer to low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7428–7445, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bour-

foune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jeff Siegel. 1999. Stigmatized and standardized varieties in the classroom: Interference or separation? *Tesol Quarterly*, 33(4):701–728.

Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation is not a design fix for machine learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA. Association for Computing Machinery.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789, Copenhagen, Denmark. Association for Computational Linguistics.

Patrick O. Steinkrüger. 2013. Zamboanga chaba-

cano structure dataset. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors, *Atlas of Pidgin and Creole Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Wietse de Vries, Martijn Bartelds, Malvina Nissim, and Martijn Wieling. 2021. Adapting monolingual models: Data can be scarce when language similarity is high. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4901–4907, Online. Association for Computational Linguistics.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on freebase via relation extraction and textual evidence. *arXiv preprint arXiv:1603.00957*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2019. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109:43–76.

Just Zwennicker and David Stap. 2022. Towards a general purpose machine translation system for sranantongo.