# 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**[4,7], **Ruth-Ann Armstrong**[5], **Abee Eijansantos**[8],
**Catriona Malau**[6], **Hans Erik Heje**[1], **Ernests Lavrinovics**[1], **Diptesh Kanojia**[9],
**Paul Belony**[10], **Marcel Bollmann**[11], **Loïc Grobol**[12], **Miryam de Lhoneux**[2],
**Daniel Hershcovich**[4], **Michel DeGraff**[13], **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]Meta, USA, [6]University of Newcastle, Australia,
[7]University of Electronic Science and Technology of China, China,
[8]Zamboanga State College of Marine Sciences and Technology, Philippines,
[9]University of Surrey, UK, [10]Kean University, USA, [11]Linköping University, Sweden,
[12]Université Paris Nanterre, France, [13]Massachusetts Institute of Technology, USA
hcle@cs.aau.dk

## 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 a number of 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 novel development datasets for reading 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, we see CREOLEVAL as an opportunity to empower research on Creoles in NLP and computational linguistics, and in general, a step towards more equitable language technology around the globe.

## 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

---

[1]In some Creole-speaking communities, the local Creole language is viewed as a ''corrupted'' language, with names like ''broken English''. Thus, speakers of Creoles might not even identify their variety as a separate language.

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, this prejudice 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 (DeGraff, 2005; Kouwenberg and Singler, 2009; Aboh and DeGraff, 2016). This *othering* of Creoles that has occurred in linguistics has led to a research landscape where Creoles are typically categorized as *exceptions* among languages, and thus separated from other languages. Take, for example, the widely used WALS database (Dryer and Haspelmath, 2013), 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 consequently exclude Creoles from their work (Rama and Kolachina, 2012; Vylomova et al., 2020; Bjerva et al., 2020; Vastl et al., 2020; Yu et al., 2021; Chronopoulou et al., 2023).[2] Beyond WALS, this pattern of exclusion is palpable across NLP, as demonstrated by the marked absence of Creoles in works investigating multilinguality through the lens of language families (Majewska et al., 2020; Jayanthi and Pratapa, 2021; Şahin, 2022; Xu et al., 2022). And while other resources exist to specifically cater to Creoles (e.g., APICS; Michaelis et al., 2013), the creation of *separate* resources to specifically accommodate Creoles is emblematic of their ghettoization within scientific spaces. In this vein, though Creoles are the singular focus of this work, our datasets, code, and models will allow others to easily incorporate Creoles into broader variety of projects, thus helping remedy the isolation of Creoles across NLP.

**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, a 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 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,

---

[2]For a critical overview of typologically diverse sampling based on language families, see Ploeger et al. (2024).

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

but not syntax. As many Creoles are genealogically related to higher-resourced 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, in line with previous works for other truly low-resource languages (Ebrahimi et al., 2022; Snæbjarnarson et al., 2023). 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 (reading comprehension, relation classification, and machine translation) for understudied Creole languages. We expand the scope of CREOLE-VAL by packaging these 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), in a public repository (see Appendix C Table 8). 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 constituting 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 a 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.
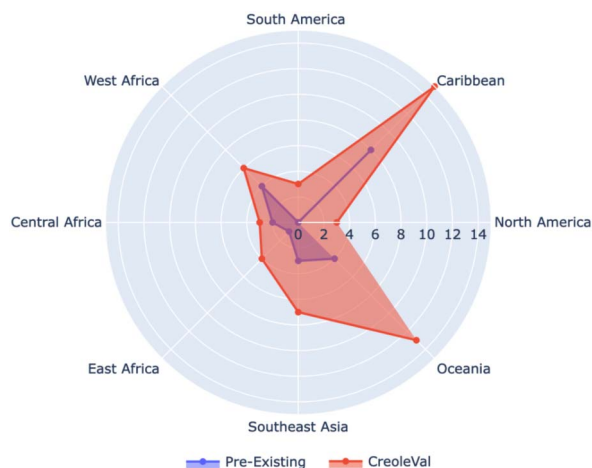
Figure 1: CREOLEVAL expands the availability of labeled data for Creoles around the globe. This chart shows the increased availability of datasets for concrete tasks, across Creoles from different regions. Before CREOLE-VAL, only 11 Creoles had data for at least 1 pre-existing task, and now 28 Creoles have labeled data for at least 1 task and at most 6 tasks.

## 2 Background

**Linguistic Context** The "Creole" label has been assigned to languages known to have arisen through contact between a linguistically diverse set of languages, as a consequence of human movement throughout history (Kouwenberg and Singler, 2009). For example, in contrast with Romance languages, which have a clear traceable history from Vulgar Latin (Alkire and Rosen, 2010), the phylogenetic origins of any given Creole language is more varied. This is because Creoles descend from a combination of languages belonging to *different* families (Michaelis et al., 2013), as illustrated by Creoles across the Caribbean (e.g., Jamaican Patois), which have close ancestral ties to both Indo-European languages (e.g., English) and African ones (e.g., Twi), as a result of European colonialism (Patrick, 2004). Due to this genealogical context, linguists have looked to Creoles to investigate the process by which new languages emerge (Bickerton, 1983; Baker, 1994; Mufwene, 1996; Lefebvre, 2001; DeGraff, 2001; Veenstra, 2008) and continue to evolve (Croft, 2000; Mufwene, 2008, 2009, 2015). Among linguists today, there is no consensus on whether Creoles constitute a separate language family (Bakker et al., 2011; Aboh, 2016), or whether the *label* of "Creole" itself is even

linguistically valid for discriminating between languages, beyond mere sociohistorical backgrounds (DeGraff, 2005; McWhorter, 2005).

**Previous Work** Prior work in NLP primarily focuses on individual Creole languages, such as Antillean Creole (Mompelat et al., 2022), Chavacano (Eijansantos et al., 2022), Jamaican Patois (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 studies specifically investigate Creoles as a collection of languages, with interest in language models (LMs) (Lent et al., 2021) and transfer learning (Lent et al., 2022a). Lent et al. (2022b) further discusses some of the social considerations for responsible NLP for Creoles, due to the languages' stigmatization and vulnerability (Alleyne, 1971; Siegel, 1999; Kouwenberg and Singler, 2009). We expand upon this existing body of research on Creoles by contributing high-quality evaluation data across a variety of tasks, ensuring that future work in Creole NLP has increased opportunities for measuring progress. While benchmarking constitutes only a small part of quality assurance for any model in practice, the creation of benchmarks also serves as an invitation to the broader research community to engage with new tasks and languages, as evidenced by the success of datasets like MasakhaNER (Adelani et al., 2021) and shared tasks (Mager et al., 2021; Ebrahimi et al., 2023; Muhammad et al., 2023; Pal et al., 2023) at bringing more languages into the mainstream of NLP research. As such, the CREOLEVAL evaluation benchmarks can similarly encourage increased involvement of Creoles in research, with the end result of faster progress towards better language technologies for Creole language speakers.

**Transfer Learning** Transfer learning is the process by which a model is trained to make use of knowledge learned in the context of one task or language, with the aim of generalizing to other tasks or languages *outside* the scope of the original training data (Zhuang et al., 2019). Over the years, many different techniques have been proposed for achieving cross-lingual transfer, such as learning alignments between words (Yarowsky et al., 2001; Padó and Lapata, 2014; Agić et al., 2016; Dou and Neubig, 2021) and word vectors (Klementiev et al., 2012; Grave et al., 2018; Kementchedjhieva et al., 2019), so knowledge from one language can be lent to another on the basis of inferred common ground. Another popular approach relies on unsupervised pre-training of LMs over large corpora, in order to establish a strong but generalized baseline of knowledge (Raffel et al., 2019). In this setting, transfer learning has been effective for extending models trained over 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 allow for transfer learning have indicated that its success in this setting may be less dependent on genealogical language relatedness, and more dependent on other factors like sub-word overlap (Pelloni et al., 2022).

**Multilingual Language Models** Selecting a pertinent 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 a lack of

---

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

| | Data | #Lang | #Creole | #Anc |
|---|---|---|---|---|
| **mBERT** | Wikipedia | 104 | 1 | 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. mBERT training data includes Haitian Creole. 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.

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 for models to exhibit decent performance. In order to expand on the availability of NLU data for Creoles we introduce two novel benchmark datasets for reading 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 setting for Creoles, as this is the most typical setup for working with languages with little to no data (Ebrahimi et al., 2022).

### 3.1 Reading 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 reading 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. Reading comprehension is an NLU task where a model is challenged 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).

**Translation** We chose to translate the MCTest160 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 professional 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). For example, the original English dataset may discuss an ice cream truck (directly translated to ''*kamyon krèm*''), though ice cream is not a typical dessert in Haiti; thus in the localized dataset, ''ice cream truck'' has been changed to ''*machann fresko*'', a cart which sells a shaved-ice desert enjoyed in Haiti. The addition of these two different Haitian Creole datasets for reading comprehension additionally paves the way for evaluating progress in cross-cultural NLP (Hershcovich et al., 2022).

**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 fine-tune them for 10 epochs over the English MCTest160 training set.

|                   | 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.

A summary of our results is in Table 2, with full results and hyperparameter settings documented in the accompanying Github repository. mBERT out-performs XLMR, although XLMR performs better over the localized data than the direct translation for Haitian. The performance on Haitian can likely be attributed to the fact that mBERT has been pre-trained on Haitian, while XLMR has not. Meanwhile, the performance on Mauritian is surprising as neither models have seen this language. It's particularly noteworthy that mBERT results on Creoles outperforms XLMR'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 four Creole languages: Bislama, Chavacano, Jamaican Patois, 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] While Wikipedia is a common source for gathering data, poor quality of articles is an outstanding issue known to plague Wikipedias of lower-resourced languages (Kreutzer et al., 2022). As such, the creation of our RC datasets involves speakers of the Creoles to ensure quality, and preserve the domain, allowing for integration of Creoles into the broader spectrum of RC projects (Sorokin and Gurevych, 2017; Köksal and Özgür, 2020; Nag et al., 2021; Chen and Li, 2021; Chen et al., 2022).

To construct the dataset, we first preprocess[7] Wikipedia dumps and perform automatic entity linking using OpenTapioca (Delpeuch, 2019). Unsurprisingly, we observe that many Creole Wikipedia entries are short and *templatic*, possibly due to machine generation. This templatic nature, however, facilitates the annotation process for RC, as it allows for straightforward identification of entities and relations by the authors of this work who have linguistic training. For example, consider the following examples from the Tok Pisin Wikipedia:

- Talin i kapitol bilong Estonia.
- Vilnius i kapitol bilong Lituwenia.
- Busares i kapitol bilong Romenia.
- Budapest i kapitol bilong Hangri.
- Stockholm i kapitol bilong Suwidan.

From these samples above, we can infer a latent template of ''[[CITY]] is the capital of [[COUNTRY]]'', with the entities having the relation ''capital of'' (P1376 in Wikidata). Thus, to facilitate manual annotation of relations, and corrections of the automatic entity tagging, we automatically cluster sentences based on the latent templates. Thereafter, sentences with annotated triples not found in Wikidata are discarded.

After the annotation process, speakers of the pertinent Creole languages assessed the quality of the samples, and furthermore provided spelling and grammar corrections, where deemed necessary. This quality assurance process was complemented by a linguistic expert who cross-referenced the datasets with linguistic grammars to identify possible errors. The process resulted in high-quality evaluation data for 4 of the 9 initially identified Creole Wikipedias, with each dataset contains 97 evaluation samples.[8]

We establish a benchmark for Creole RC using a zero-shot cross-lingual transfer approach: We employ LMs that have not been exposed to the Creoles 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

---

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

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

[8]For a complete discussion on dataset creation, latent templates, and manual review processes, see Appendix A

| 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) | | 59.63±3.48 | 76.15±1.59 | 63.47±1.75 | 62.15 ±1.65 | 46.76±2.58 | 50.58±2.08 | 49.11±2.51 | 49.04±1.49 |
| bi | | 28.01±2.42 | 25.61±3.92 | 27.66±5.45 | 25.96±3.80 | 18.81±4.04 | 9.62±0.78 | 19.42±4.51 | 14.79±1.77 |
| cbk-zam | | 20.06±5.88 | 20.85±6.03 | 17.67±6.68 | 17.39±6.45 | 27.08±6.86 | 18.48±6.83 | 18.50±2.77 | 20.32±2.73 |
| jam | | 26.97±5.87 | 15.65±5.00 | 20.07±5.93 | 23.98±7.24 | 10.62±1.27 | 9.42±5.71 | 9.06±1.70 | 10.22±0.92 |
| tpi | | 23.57±4.17 | 22.90±2.97 | 22.86±8.13 | 21.42±5.96 | 9.36±3.77 | 11.64±5.54 | 8.31±8.07 | 8.48±4.78 |
| AVG | | **24.65** | 21.25 | 22.06 | 22.19 | **16.47** | 12.29 | 13.82 | 13.45 |

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.

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), 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 mBERT and XLM-R (Conneau et al., 2020) models 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 as Bb-nli, Bl-nli, Xr-b, and Xr-100[10] here, as sentence embeddings of the relation descriptions from Wikidata.

**Results and Analysis** Table 3 shows the performance of RC in each setting. We observe worse performance in the Creoles than English. This highlights the particular challenge of leveraging pretrained LMs 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 XLMR. Under the same sentence encoder, different relation encoders exhibit slight variations in performance. We speculate that mBERT may perform better due to its pre-training over Wikipedia,

in contrast to XLMR, which is pre-trained over a wider variety of domains. Previous works on multilingual factuality also observe mBERT outperforming XLMR (Jiang et al., 2020; Fierro and Søgaard, 2022).

### 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). In contrast with the brand new datasets presented in CREOLEVAL, the majority of prior benchmarks allow for supervised learning. Thus, in order to ascertain the expected performance for these tasks given the data available, we train and evaluate fully supervised models, where training data exists (UDPoS, NER, and sentiment analysis). For JamPatoisNLI (Armstrong et al., 2022), we reproduce the authors' results by following the reported methodology: First we fine-tune on English MNLI (Williams et al., 2018), before doing few-shot learning on 250 samples of Jamaican Patois. The sentence-matching Tatoeba task (Artetxe and Schwenk, 2019) is the only without dedicated training or few-shot data, and thus is the only task where we evaluate the zero-shot performance of the pertinent LMs. The performance on the test set for each task and LM in Table 4. Unsurprisingly, performance is best when training data is available, though few-shot learning shows promising results in the case of JamPatoisNLI. However, previous work has noted that a high token overlap is needed to successfully achieve

---

[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`.

| Task | Language | Dataset | Metric | mBERT | XLM-R | mT5 |
|------|----------|---------|--------|-------|-------|-----|
| **UDPoS** (supervised) | 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 |
| | pcm | MasakhaNER (Adelani et al., 2021) | Span-F1 | 0.89 | 0.89 | **0.90** |
| **NER** (supervised) | bis | | | **0.94** | 0.90 | 0.72 |
| | cbk-zam | | | **0.96** | 0.96 | 0.94 |
| | hat | | | 0.78 | **0.84** | 0.48 |
| | pih | WikiAnn (Pan et al., 2017) | Span-F1 | **0.90** | 0.88 | 0.61 |
| | sag | | | 0.89 | **0.93** | 0.79 |
| | tpi | | | **0.91** | 0.89 | 0.75 |
| | pap | | | **0.90** | 0.89 | 0.85 |
| **SA** (supervised) | 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** (few-shot) | jam | JamPatoisNLI (Armstrong et al., 2022) | Acc | 0.74 | **0.76** | 0.66 |
| **Sentence Matching** (zero-shot) | cbk-eng | | | **15.9** | 3.9 | 6.5 |
| | gcf-eng | | | **12.8** | 4.9 | 6.9 |
| | hat-eng | | | 23.9 | 18.5 | **37.9** |
| | jam-eng | Tatoeba (Artetxe and Schwenk, 2019) | Acc | **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: dependency parsing (UDPoS), named entity recognition (NER), sentiment analysis (SA), natural language inference (NLI), and sentence matching. Additional experiments, results, and analysis are included in the CreoleVal repository's documentation.

cross-lingual transfer for languages *unseen* by a pre-trained LM (Winata et al., 2022). As spelling conventions for many Creoles have greatly diverged from those of ancestor languages (e.g., ''Pwofesè'' in Haitian Creole to ''Professeure'' in French), subword token overlap between Creoles and related languages will likely be low, and therefore few-shot learning may not help in such scenarios. As additional samples for few-shot learning are not available for most Creoles, there is an outstanding need for improved zero-shot performance via transfer learning, until further data can be curated.

# 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.

## 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 1,000 and 2,000 sentences for each Creole and English Bible 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. See Appendix B for exact details on dataset sizes.

---

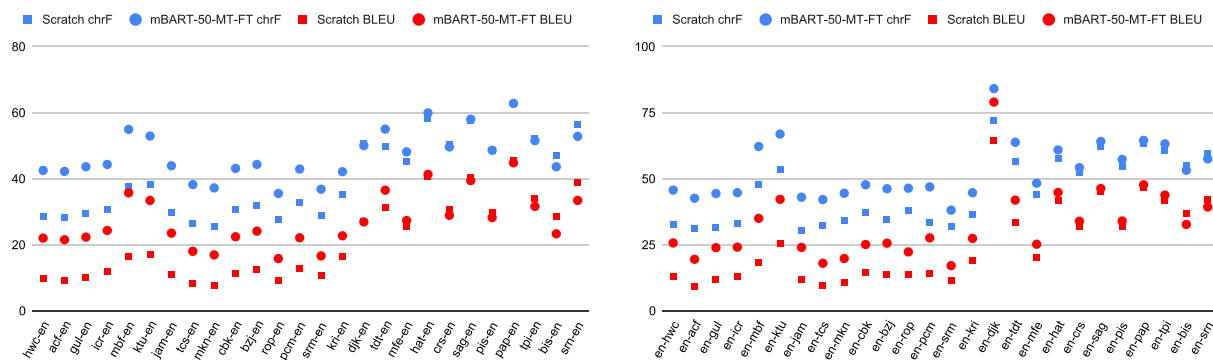[11]To access the raw Bible corpora, one must request the authors due to copyright issues.

Figure 2: chrF (blue) BLEU (red) 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.

### 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. Here, we train models from scratch as well as by fine-tuning the mBART-50-MT model[13] following Dabre and Sukhoo (2022). The training utilizes the Adam optimizer (Kingma and Ba, 2014), and trains until convergence. We evaluate the training performance on the development set using BLEU score

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, performance appears 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. We 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 drop in translation quality, indicating that the fine-tuned model converges too quickly, and is unable to learn well from the training data.

---

[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.

## 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 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 ($\sim$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-MT 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

| 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.

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.

## 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

---

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

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

| 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.

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. 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's (2022) 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's (2022) best results.[19]

**Other** We exclude **PidginUNMT** (Ogueji and Ahia, 2019), as this unlabeled dataset pertains to unsupervised machine translation, and thus cannot be used as gold-standard evaluation data. We also exclude **WMT11** (Callison-Burch et al., 2011), as it 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 are part of a unified platform, despite previously having limited or no NLP datasets. 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

---

[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.

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 conventional 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 CREOLE-VAL 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**  While 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; Imankulova et al., 2019) have shown that even a small amount of in-domain corpus may be sufficient for adaptation to other domains.

**Mixture of Data Quality**  In this work, we put forth and experiment with a combination of higher and lower quality data, the latter coming from the

religious domain. Works in NLP have long relied on religious texts for truly low-resource languages, which often have no other available data (Agić et al., 2015, 2016). However, the use of such data comes with concerns over data quality, as such texts are often written by foreign missionaries, they cannot be considered strictly representative of the language as used by native speakers (Nida, 1945). While the inclusion of religious data is still a common necessity in the realm low-resource NLP, the addition of our higher quality data for Creoles ensures that future works will have a wider variety of resources to evaluate their systems, than previously available. Moreover, when sourcing data from domains like Wikipedia, we involve speakers and cross-reference linguistic grammars, leading us to exclude several languages due to quality issues, such as Pitkern.

**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 historic lack of 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).

**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 (Kargaran et al., 2023).

**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

## 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, RAA, AE, CM, MF, HEH, EL, 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;

---

[21]`https://commoncrawl.org/`.

[22]`https://www.jwel.mit.edu/`.

**Writing** HL, KT, RD, YC, MF, EP, LZ, DK, MB, MDL, DH, MDG, JB.

# References

Enoch O. Aboh. 2016. Creole distinctiveness: A dead end. *Journal of Pidgin and Creole Languages*, 31(2):400–418. https://doi.org/10.1075/jpcl.31.2.07abo

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 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. https://doi.org/10.1162/tacl_a_00416

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China. Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-2044

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312. https://doi.org/10.1162/tacl_a_00100

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. https://doi.org/10.18653/v1/P19-1310

Ti Alkire and Carol Rosen. 2010. *Romance Languages: A Historical Introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9780511845192

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. https://doi.org/10.18653/v1/2022.findings-emnlp.389

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. https://doi.org/10.1162/tacl_a_00288

Beryl Loftman Bailey. 1966. *Jamaican Creole Syntax*. Cambridge University Press.

Philip Baker. 1994. Creativity in creole genesis. *Creolization and Language Change*, pages 65–84. https://doi.org/10.1515/9783111339801.65

Peter Bakker, Aymeric Daval-Markussen, Mikael Parkvall, and Ingo Plag. 2011. Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole Languages*, 26(1):5–42. https://doi.org/10.1075/jpcl.26.1.02bak

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.

Derek Bickerton. 1983. Creole languages. *Scientific American*, 249(1):116–123. https://doi.org/10.1038/scientificamerican0783-116

Steven Bird. 2021. ''LT4All!? Rethinking the agenda'' keynote.

Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, Edoardo Maria Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. SIGTYP 2020 shared task: Prediction of typological features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 1–11, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.sigtyp-1.1

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. https://doi.org/10.18653/v1/W19-7803

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. https://doi.org/10.18653/v1/2021.naacl-main.272

Yuxuan Chen, David Harbecke, and Leonhard Hennig. 2022. Multilingual relation classification via efficient and effective prompting. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1059–1075, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.69

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-family adapters for low-resource multilingual neural machine translation. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 59–72, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.loresmt-1.5

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. https://doi.org/10.18653/v1/2020.acl-main.747

William Croft. 2000. *Explaining Language Change: An Evolutionary Approach*. Explaining Language Change: An Evolutionary Approach. Longman.

Terry Crowley. 2004. *Bislama Reference Grammar*. University of Hawaii Press.

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. 2001. On the origin of creoles: A cartesian critique of neo-darwinian linguistics. *Linguistic Typology*, 5(2/3):213–310.

Michel DeGraff. 2003. Against creole exceptionalism. *Language*, 79(2):391–410. `https://doi.org/10.1353/lan.2003.0114`

Michel DeGraff. 2005. Linguists' most dangerous myth: The fallacy of creole exceptionalism. *Language in Society*, 34(4):533–591. `https://doi.org/10.1017/S0047404505050207`

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.

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. `https://doi.org/10.18653/v1/2021.findings-acl.433`

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. `https://doi.org/10.18653/v1/2022.acl-long.529`

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.eacl-main.181`

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo. `https://doi.org/10.5281/zenodo.7385533`

Stephanie Durrleman. 2008. *The Syntax of Jamaican Creole*.

Martin Eberl. 2019. *Innovation and Grammaticalization in the Emergence of Tok Pisin*. Ph.D. thesis, LMU.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.acl-long.435`

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language*

*Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.americasnlp-1.23

Abee Eijansantos, Jeric Ventoza, Rochelle Irene Lucas, and Ericson Alieto. 2022. Zamboanga Chavacano verbal aspects: Superstrate and substrate influences in morphosyntactic behavior. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 723–732, Manila, Philippines. Association for Computational Linguistics.

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.240

Edouard Grave, Armand Joulin, and Quentin Berthet. 2018. Unsupervised alignment of embeddings with wasserstein procrustes. *CoRR*, abs/1805.11222.

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. https://doi.org/10.18653/v1/D18-1514

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. https://doi.org/10.18653/v1/2022.acl-long.482

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.

Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 128–139, Dublin, Ireland. European Association for Machine Translation.

Sai Muralidhar Jayanthi and Adithya Pratapa. 2021. A study of morphological robustness of neural machine translation. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 49–59, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.sigmorphon-1.6

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.479

Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*,

pages 6155–6218, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.410

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. 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 3336–3341, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1328

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

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India. The COLING 2012 Organizing Committee.

Abdullatif Köksal and Arzucan Özgür. 2020. The RELX dataset and matching the multilingual blanks for cross-lingual relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 340–350, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.32

Silvia Kouwenberg and John Victor Singler. 2009. *The Handbook of Pidgin and Creole Studies*. John Wiley & Sons. https://doi.org/10.1002/9781444305982

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72. https://doi.org/10.1162/tacl_a_00447

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. https://doi.org/10.18653/v1/D18-2012

Claire Lefebvre. 2001. Relexification in creole genesis and its effects on the development of the creole. *Creolization and Contact*, pages 9–42. https://doi.org/10.1075/cll.23.02lef

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. https://doi.org/10.18653/v1/2021.conll-1.5

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. https://doi.org/10.18653/v1/2022.insights-1.9

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 completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29. https://doi.org/10.1609/aaai.v29i1.9491

John M. Lipski and Maurizio Santoro. 2007. Zamboangueño creole spanish. *Comparative Creole Syntax*, pages 373–398.

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. https://doi.org/10.1162/tacl_a_00343

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. https://doi.org/10.18653/v1/2023.acl-long.268

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.americasnlp-1.23

Olga Majewska, Ivan Vulić, Diana McCarthy, and Anna Korhonen. 2020. Manual clustering and spatial arrangement of verbs for multilingual evaluation and typology analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4810–4824, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.423

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).

John H. McWhorter. 2005. *Defining Creole*. Oxford University Press. https://doi.org/10.1093/oso/9780195166699.001.0001

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. https://doi.org/10.18653/v1/P19-1491

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.

Salikoko S. Mufwene. 1996. The founder principle in creole genesis. *Diachronica*, 13(1):83–134. https://doi.org/10.1075/dia.13.1.05muf

Salikoko S. Mufwene. 2008. *What Do Creoles and Pidgins Tell Us About the Evolution of Language?* Equinox. https://doi.org/10.1002/9781444302851.ch54

Salikoko S. Mufwene, Hong Kong, China. 2009. The evolution of language: Hints from creoles and pidgins. *Language Evolution and the Brain*, pages 1–33.

Salikoko S. Mufwene. 2015. The emergence of creoles and language change. In *The Routledge Handbook of Linguistic Anthropology*, pages 348–365, Routledge.

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 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. https://doi.org/10.48550/ARXIV.2201.08277

Peter Mühlhäusler. 2020. *Pitkern-Norf'k: The Language of Pitcairn Island and Norfolk Island*, volume 17. Walter de Gruyter GmbH & Co KG. https://doi.org/10.1515/9781501501418

Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2021. A data bootstrapping recipe for low-resource multilingual relation classification. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 575–587, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.conll-1.45

Eugene Nida. 1945. Linguistics and ethnology in translation-problems. *Word*, 1(2):194–208. https://doi.org/10.1080/00437956.1945.11659254

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. https://doi.org/10.18653/v1/2020.emnlp-main.368

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.

Sebastian Padó and Mirella Lapata. 2014. Cross-lingual annotation projection for semantic roles. *CoRR*, abs/1401.5694.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.56

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.

Peter L. Patrick. 2004. Jamaican creole: Morphology and syntax. *A Handbook of Varieties of English*, 2:407–438.

Peter L. Patrick. 2014. Jamaican creole. *Languages and Dialects in the US: Focus on Diversity and Linguistics*, pages 126–136.

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. https://doi.org/10.18653/v1/2022.emnlp-main.503

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. https://doi.org/10.18653/v1/P19-1493

Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. What is 'typological diversity' in NLP?

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Taraka Rama and Prasanth Kolachina. 2012. How good are typological distances for determining genealogical relationships among languages? In *Proceedings of COLING 2012: Posters*, pages 975–984, Mumbai, India. The COLING 2012 Organizing Committee.

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. https://doi.org/10.18653/v1/D19-1410

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 Bourfoune, 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 Periñén, 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. https://doi.org/10.2307/3587883

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. https://doi.org/10.1145/3551624.3555285

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

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. https://doi.org/10.18653/v1/D17-1188

Patrick O. Steinkrüger. 2013. Zamboanga chabacano 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.

Martin Vastl, Daniel Zeman, and Rudolf Rosa. 2020. Predicting typological features in WALS using language embeddings and conditional probabilities: ÚFAL submission to the SIGTYP 2020 shared task. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pages 29–35, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.sigtyp-1.4

Tonjes Veenstra. 2008. Creole genesis: The impact of the language bioprogram hypothesis. *The Handbook of Pidgin and Creole Studies*, pages 219–241. https://doi.org/10.1002/9781444305982.ch9

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.sigmorphon-1.1

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. `https://doi.org/10.18653/v1/P17-1159`

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, Association for Computational Linguistics. `https://doi.org/10.18653/v1/N18-1101`

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791, Online only. 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. `https://doi.org/10.18653/v1/D19-1077`

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*.

Ningyu Xu, Tao Gui, Ruotian Ma, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2022. Cross-linguistic syntactic difference in multilingual BERT: How good is it and how does it affect transfer? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8073–8092, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*. `https://doi.org/10.3115/1072133.1072187`

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.

Dian Yu, Taiqi He, and Kenji Sagae. 2021. Language embeddings for typology and cross-lingual transfer learning. 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 7210–7225, Online. 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.

    `https://doi.org/10.18653/v1/2021
    .starsem-1.22`

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. `https://doi.org/10.1109
    /JPROC.2020.3004555`

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

Gözde Gül Şahin. 2022. To Augment or Not
    to Augment? A Comparative Study on Text
    Augmentation Techniques for Low-Resource
    NLP. *Computational Linguistics*, 48(1):5–42.
    `https://doi.org/10.1162/coli_a_00425`

## A   Relation Classification

Here, we thoroughly describe our steps to cre-
ate the relation classification datasets, from data
collection, to annotation and verification. This
discussion is intended to provide details for exact
replication of the work described in the paper,
for creating these datasets. For an overview, our
methodology consisted of the following steps:

1. Collecting and cleaning data from Wikipedia
   dumps, and performing automatic entity
   linking.

2. Clustering sentences which belong to the
   same latent template (i.e., the sentences ex-
   press the same relation, as evidenced by an
   exact or near-exact overlap in the text, with
   the only differences being the entities; more
   details are provided in Appendix A.2).

3. Manually verifying and correcting any
   mistakes from the automatic entity-linking.

4. Manually annotating the relation expressed
   in the sets of utterances (as grouped by the
   latent templates) and its associated Property
   in Wikidata.

5. Validating that the annotated triples indeed
   exist in Wikidata; sentences where the triples
   did not exist in Wikidata (due to gaps in the
   knowledge base) were thrown out.

6. Manually checking the correcting the an-
   notated sentences to ensure that the samples
   truly reflect real-world usage of the language.

   (a) A manual verification of each dataset
       was performed by a speaker of each
       Creole. Each sentence was assessed,
       and speakers made corrections to the
       grammar or spelling, as they saw
       fit. Whenever possible, an additional
       speaker was asked to double-check these
       changes.

   (b) Complementing the above step, a man-
       ual verification of the datasets is
       conducted using published linguistic
       grammars for the relevant language, to
       help identify potential issues in the data.

   (c) A final re-verification of the entity tag-
       ging and property labels was conducted,
       to ensure that any corrected sentences
       were still properly annotated.

For steps 1–4, we produced datasets for:
Bislama, Chavacano, Haitian Creole, Jamaican
Patois, and Pitkern, and Tok Pisin. However at
step 5, the triples for Haitian Creole were not val-
idated by the Wikidata and thus this dataset was
discarded. Here, simple triples like (apple, is_a,
fruit) were missing from the knowledge graph.
Additionally at step 6, the Pitkern samples failed
to conform with the description of the language
detailed in the grammar, and was also excluded
from this work. Ultimately, this resulted in high-
quality relation classification evaluation data for
4 of the 9 Creole Wikipedias we started with:
Bislama, Chavacano, Jamaican Patois, and Tok
Pisin.

### A.1   Data Collection and Annotation

We first clean the data and perform automatic en-
tity linking and filtering, in order to facilitate the
process of manual annotation. First, we preprocess
the Wikipedia dumps by removing unnecessary
HTML with Beautifulsoup and tokenization with
Spacy. We then automatically label entities and
link them to Wikidata, a process known as en-
tity linking, first by linking tokens with existing
Wikipedia hyperlinks within the text, and then
attempt to label any remaining entities without
hyperlinks by leveraging OpenTapioca. Before
any manual annotation over these examples, we

then attempt to automatically group sentences by latent templates, so that sentences can be annotated in groups, allowing us to identify and annotate the correct relationship between the entities, as expressed in the sentences (see ''Latent Templates'', below). To this end, we perform automatic clustering over the sentences using first fuzzy string matching with partial token sort ratio, and thereafter affinity propagation, in hopes that utterances sharing templatic spans of text will be clustered together. The result is a large set of clusters, each containing a number of utterances that are at least somewhat similar. In order to refine these clusters further, we first rank the clusters by the longest common string therein, and we then discard clusters below a certain threshold of similarity, as we can assume the sentences do not belong to the same latent template. Finally, with the highest-scoring clusters of entity-linked sentences, the authors perform a manual annotation of entities and relations.

## A.2 Latent Templates

In Section 3.2, we mention the latent templates that the sentences belong to, and how these templates enable more confident manual annotation. To clarify this, we will show some examples of latent templates, and how we map this to Wikidata Properties (i.e., relations) and entities. Note that samples were clustered by latent templates *before* validation and correction by the Creole language speakers, so the provided examples below do not represent the finalized dataset. Consider the following entity-tagged sentences in Bislama:

- Mongolia i kaontri long Esia.

- Fiji hem i wan kaontri long Pasifik.

- Jemani i kaontri long Yurop.

- Bukina Faso i kaontri long Afrika.

- Kanada i wan kaontri blong Not Amerika.

When we look at these sentences as a group (i.e., a cluster), we can see there is a latent template of [[ABC]] (hem) i (wan) kaontri (b)long [[XYZ]]. All sentences in the cluster belong to this latent template, albeit with some minor variations, which are later inspected and assessed in detail during the validation stage by a speaker of Bislama,

and additional with a cross-reference against a linguistics grammar documenting the language.

Moving on, for the **entities** themselves, we can identify the Wikidata Qcode in 2 ways:

1. The entities (e.g., Mongolia, Pasifik) were already hyperlinked in the Wikipedia article, which means we have a URL, from which we can get the gold entity Q-code.

2. The entities are Named Entities with spelling clearly influenced by English, and we can make an educated guess about the meaning.

Thus from the template and entities, we can now consider the **relation** between the entities:

(Mongolia is to Asia) as (Fiji is to Pacific) as (Germany is to Europe) as (Canada is to North America) and (Burkina Faso is to Africa).

For all of these entity pairs, to a human annotator, it is clear that the relationship is [[COUNTRY]] is in [[CONTINENT]]. Thus we can annotate the Wikidata Property as P30: ''continent of which the subject is a part''.

Finally, we can automatically verify our triples (entity1, Property, entity2) against the Wikidata knowledge graph. We remove any sentences where the triple was not in the knowledge graph. This unfortunately removes correct data points, where there is simply a gap in the knowledge graph; for example, the Haitian dataset was removed for this reason, as Wikidata missed simple cases like (apple, is_a, fruit). But importantly, it also is a sanity measure of our annotation method performed by the authors, which at times required educated guesswork about the meaning of an entity, as non-native speakers, when the entity was not already hyper-linked. Presumably, if we incorrectly annotated an entity, the triple will not exist in the knowledge graph, and thus be removed. Imagine that we had incorrectly annotated [[Kanada]] (from the sentence [[Kanada]] i wan kaontri blong [[Not Amerika]].) to be the language *Kannada* (Q33673)), rather than the country *Canada* (Q16). The triple (Kannada_language, ''continent of which the subject is a part'', North America) would certainly not exist in Wikidata, and thus the entire annotated example would be removed. Yet (Canada, ''continent of which the subject is a part'', North America) is indeed in the knowledge base, so we can be confident in our annotation. Again, having samples listed together

in groups by latent templates also makes us more certain of the meaning.

Here are some more examples of latent templates in the data, and the expressed relations:

**Chavacano**
Latent template: [[PERSON]] is a [[SINGER]].
Property P106: ''occupation of a person''
Examples:

- Billie Eilish es un cantante

- Sopho Khalvashi es un cantante

- Juanes es un cantante de Colombia de pop.

- Nina Sublatti (Sulaberidze) es un cantante

- Nini Shermadini es un cantante

**Jamaican Patois**
Latent template: [[CITY]] is the capital of [[COUNTRY]]
Property P1376: ''capital of''
Examples:

- Sofiya a di kiapital fi Bulgieria.

- Broslz a di kiapital fi Beljiom.

- Ruom a di kyapital fi Itali.

- Masko a di kyapital fi Rosha.

- Atenz a di kyapital fi Griis.

### A.3 Validation and Corrections

The samples were corrected by six speakers and then further validated by one speaker in order to reflect diverse spelling conventions (Kreutzer et al., 2022). In conjunction with the validation performed by speakers, we also check published linguistic grammars for these languages, to ensure that our published datasets constitute the upmost quality.

**Validation and Corrections by Speakers** For Bislama, Chavacano, Jamaican Patois, and Tok Pisin, we collaborated with at least one speaker of the language to validate and correct the annotated samples. Here, our speakers are either semi-native speakers (i.e., they grew up using the language), or professional linguists who live in the pertinent community and speak the language on a daily basis. Indeed, as many Creoles exist as a lingua-franca in multilingual communities, there are not always ''native speakers'', in the sense that the Creole will be their mother tongue (Lent et al., 2022b). We provide details and discussion on the validation and corrections made for each language below:

- **Bislama**: The samples were corrected by one speaker. Overall, the speaker found that some sentences were completely correct, fluent Bislama, with minor spelling errors. Almost all sentences were understandable, but many contained specific grammatical errors or contained many spelling errors. Only a few sentences were completely wrong, and corrected accordingly, to capture the meaning of the annotated triple. The major grammatical errors involved missing prepositions, incorrect usage of articles, or incorrect verb tense.

- **Chavacano**: The samples were corrected by one speaker, and further validated by a second. Here, the sentences in Wikipedia were determined not to be Chavacano, but rather an approximation of Spanish. As the intended meaning of the utterances was still clear, the speaker produced new utterances in Chavacano, to correctly capture the intended meaning with the tagged entities and labeled relation.

- **Jamaican**: The samples were corrected by one speaker, and further validated by six others. The spelling and grammar of the Wikipedia sentences was found to be greatly divergent from real-world Jamaican, and thus not representative of the language. Specifically the orthography did not match what is used by Jamaican speakers, and there were a number of grammatical constructions that would not be used by native speakers. To remedy this, the speaker produced new utterances in Jamaican, to correctly capture the intended meaning with the tagged entities and labeled relation.

- **Tok Pisin**: The samples were validated by two speakers, who noted that while the data is correct, it is distinctly representative of the urban variety of the language (*Tok Pisin bilong taun*), which can vary greatly from the rural variety (*Tok Pisin bilong ples*). Thus

| Pair | Creole | Ancestor(s) | #Lines | #Words-Source | #Words-Target |
|------|--------|-------------|--------|---------------|---------------|
| **hwc-eng** | Hawaiian Pidgin | English | 4,366 | 144,281 | 102,794 |
| **acf-eng** | Saint Lucian Creole | French | 4,889 | 135,006 | 115,176 |
| **gul-eng** | Gullah | English | 4,889 | 153,823 | 115,176 |
| **icr-eng** | San Andrés–Providencia Creole | English | 4,889 | 151,372 | 115,176 |
| **mbf-eng** | Malay Baba | Malay | 4,889 | 107,234 | 115,176 |
| **ktu-eng** | Kituba | Kikongo | 4,889 | 103,577 | 115,176 |
| **jam-eng** | Jamaican Creole | English | 5,012 | 206,692 | 168,134 |
| **tcs-eng** | Torres Strait Creole | English | 6,350 | 198,593 | 152,642 |
| **mkn-eng** | Kupang | Malay | 6,422 | 214,390 | 153,596 |
| **cbk-eng** | Chavacano Creole | Spanish | 7,071 | 182,859 | 127,090 |
| **bzj-eng** | Belizean | English | 12,085 | 262,496 | 218,526 |
| **rop-eng** | Australian Kriol | English | 27,617 | 832,308 | 703,888 |
| **pcm-eng** | Nigerian Pidgin | English | 28,267 | 523,916 | 459,266 |
| **srm-eng** | Saramaccan Language | English, Portuguese | 39,640 | 973,176 | 627,273 |
| **kri-eng** | Sierra Leonean Creole | English | 47,673 | 1,039,743 | 760,699 |
| **djk-eng** | Aukan | English | 58,108 | 1,487,156 | 1,015,311 |
| **tdt-eng** | Tetun Dili | Portuguese | 118,461 | 2,209,118 | 1,923,333 |
| **mfe-eng** | Mauritian Creole | French | 189,877 | 3,549,493 | 3,014,530 |
| **hat-eng** | Haitian Creole | French | 208,772 | 4,132,691 | 3,322,288 |
| **crs-eng** | Seychellois Creole | French | 220,861 | 3,984,410 | 3,750,620 |
| **sag-eng** | Sango | Ngabandi, French | 260,853 | 6,089,066 | 4,246,373 |
| **pis-eng** | Pijin | English | 277,378 | 4,783,222 | 4,458,132 |
| **pap-eng** | Papiamento | Spanish | 396,092 | 7,297,575 | 6,384,282 |
| **tpi-eng** | Tok Pisin | English | 399,486 | 8,365,958 | 6,334,237 |
| **bis-eng** | Bislama | English | 488,393 | 10,751,097 | 7,903,431 |
| **srn-eng** | Sranan Tongo | English | 583,746 | 13,450,377 | 9,911,997 |
| **Total** | – | – | 3,410,975 | 71,329,629 | 56,314,322 |

Table 7: Statistics of the training set of the CreoleM2M dataset.

| Task | Dataset | Language (ISO-638-3) | Metric | License | Domain | Total Sent. | Total words |
|------|---------|----------------------|--------|---------|--------|-------------|-------------|
| MC | CreoleVal MC | hat-dir, hat-loc, mfe | Acc | Microsoft License | Education | 3894 | 32068 |
| RC | CreoleVal RC | bis, cbk, jam, tpi | F1 | CC0 | WikiDump | 785 | 4106 |
| MT | CreoleVal Religious MT | bzj, bis, cbk, gul, hat, hwc, jam, ktu, kri, mkn, mbf, mfe, djk, pcm, pap, pis, acf, icr, sag, srm, crs, srn, tdt, tpi, tcs | Bleu, chrF | Copyrighted | Religion | 64394 | 811741 |
| MT | CreoleVal MIT-Haiti | hat | Bleu, chrF | CC 4.0 | Education | 3164 | 36281 |
| Pretraining data | CreoleVal MIT-Haiti | hat | N/A | CC 4.0 | Education | 8281 | 116444 |
| UDPoS | Singlish Treebank◇ (Wang et al., 2017) | singlish | Acc | MIT | Web Scrape | 1200 | 10989 |
| | UD_Naija-NSC◇ (Caron et al., 2019) | pcm | Acc | CC 4.0 | Dialog | 9621 | 150000 |
| NER | MasakhaNER◇ (Adelani et al., 2021) | pcm | Span-F1 | Apache 2.0 | BBC News | 3000 | 76063 |
| | WikiAnn⋆ (Pan et al., 2017) | bis cbk hat, pih, sgg, tpi, pap | Span-F1 | Unspecified | WikiDump | 5877 | 74867 |
| SA | AfriSenti◇ (Muhammad et al., 2023) | pcm | Acc | CC BY 4.0 | Twitter | 10559 | 235679 |
| | Naija VADER⋆ (Oyewusi et al., 2020) | pcm | Acc | Unspecified | Twitter | 9576 | 101057 |
| NLI | JamPatoisNLI◇ (Armstrong et al., 2022) | jam | Acc | Unspecified | Twitter, web | 650 | 2612 |
| SM | Tatoeba⋆ (Artetxe and Schwenk, 2019) | cbk, gcf, hat, jam, pap, sag, tpi | Acc | CC-BY 2.0 | General web | 49192 | 319719 |
| MT | KreolMorisienMT◇ (Dabre and Sukhoo, 2022) | mfe | Bleu, chrF | MIT License | Varied | 6628 | 23554 |
| | | | | | New: | 80518 | 1000640 |
| | | | | | Total: | 176821 | 1995180 |

Table 8: Overview of the datasets included in CREOLEVAL. Newly introduced datasets are prefixed with ‘‘CreoleVal’’; ⋆ indicates modified and further denoised datasets based on previous works; ◇ indicates inclusion within our benchmark where we provide download and experiment scripts as part of our Github repository, but do not re-package the data itself. Task abbreviations: MC (machine reading comprehension); RC (relation classification); MT (machine translation); UDPoS (universal dependencies part-of-speech tagging); NER (named entity recognition); SA (sentiment analysis); NLI (natural language inference); SM (sentence matching). Note that for SM task, the language format is XXX-eng. For WikiAnn, NaijaVADER and JamPatoisNLI datasets, the licenses were not explicitly stated in corresponding repositories.

for future work, collecting and annotating samples that capture a wider spectrum of Tok Pisin will be key for expanding language technology to this language.

After all manual corrections were made, we conduct an additional round of manual validation, to ensure that the entity tagging and relation labels were still correct.

One common thread across all languages involved spelling, as many Creoles do not have strictly observed orthography. For example, for lesser-known named entities, there is likely to be great variation across speakers, in whether they default to English spelling, or rather attempt to represent the word according to their pronunciation. This issue highlights an area of future work, for extending Creole language datasets to capture a wider variety of voices and approaches to spelling. To this point, some speakers chose to add limitation variation across their corrections of the data. For example, in the Bislama dataset, there can be found variation in constructions combing the third person-singular pronoun and the predicate marker *i*.

Finally, while we did not have funds to pay the speakers for their assistance in this work, the speakers were invited to join the project as co-authors of this work, or otherwise be thanked by name in the Acknowledgments, per their preference. We believe no speakers were harmed in this process, and we are deeply grateful for their collaboration in this work.

**Validation through Linguistic Grammars** Full documentation of our grammar check has been submitted as supplementary material alongside this manuscript, for inspection by the reviewers. As we cite directly from published books, copyright prevents us from making our grammar check public. For Bislama we referred to Crowley (2004), for Chavacano we referred to Lipski and Santoro (2007), and for Jamaican Patois we primarily referred to Patrick (2014), but also referenced others (Durrleman, 2008; Patrick, 2004; Bailey, 1966). For Pitkern we referred to Mühlhäusler (2020), and finally for Tok Pisin we referred to Eberl (2019). Among all of these languages, Pitkern was the only case where the Wikipedia data failed to meet the description of language, and was thus removed.

## B  Machine Translation: Creole M2M

### B.1  Dataset Statistics

Table 7 shows the statistics of the training set of the CreoleM2M dataset, spanning 26 Creoles originating from one or more of 8 parent (ancestor) languages. We give the number of lines, and number of words on the source (Creole) and target (English) sides.

## C  Overview