JamPatoisNLI: A Jamaican Patois Natural Language Inference Dataset

Ruth-Ann Armstrong **John Hewitt** **Christopher Manning**
Department of Computer Science
Stanford University
{ruthanna,johnhew,manning}@cs.stanford.edu

Abstract

JamPatoisNLI provides the first dataset for natural language inference in a creole language, Jamaican Patois. Many of the most-spoken low-resource languages are creoles. These languages commonly have a lexicon derived from a major world language and a distinctive grammar reflecting the languages of the original speakers and the process of language birth by creolization. This gives them a distinctive place in exploring the effectiveness of transfer from large monolingual or multilingual pretrained models. While our work, along with previous work, shows that transfer from these models to low-resource languages that are unrelated to languages in their training set is not very effective, we would expect stronger results from transfer to creoles. Indeed, our experiments show considerably better results from few-shot learning of JamPatoisNLI than for such unrelated languages, and help us begin to understand how the unique relationship between creoles and their high-resource base languages affect cross-lingual transfer. JamPatoisNLI, which consists of naturally-occurring premises and expert-written hypotheses, is a step towards steering research into a traditionally underserved language and a useful benchmark for understanding cross-lingual NLP.

1 Introduction

The extensive progress that has been made in NLP research in recent years has largely been constrained to around 20 of the 7000 languages spoken around the world (Maguéresse et al., 2020). Creole languages, which emerge as a result of contact between speakers of different vernaculars, are even further underexplored (Lent et al., 2022b).

This work contributes to addressing this gap. We present JamPatoisNLI, the first natural language inference dataset in Jamaican Patois, which is an English-based creole spoken in the Caribbean. Additionally, to our knowledge, no other natural language inference corpus exists for any other creole language.

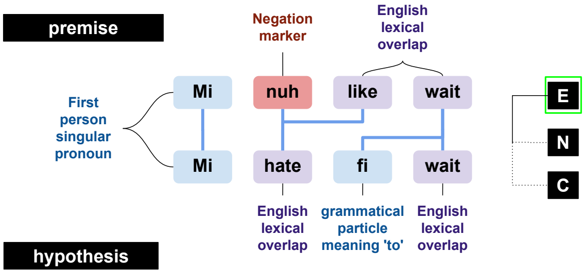

Figure 1: Linguistic features relevant for textual entailment classification for Jamaican Patois and lexical overlap with English.

Jamaican Patois is one of over 100 creole languages spoken by millions of inhabitants of different regions across the world, including Africa, the Caribbean, the Americas, islands in the Indian Ocean and the Pacific Ocean (including Australia and the Philippines) and South Asia (Romaine, 2017; Bakker and Daval-Markussen, 2013). Though there has been a recent spike in interest in work on low-resource languages in the NLP community (Kuriyozov et al., 2022; Kumar et al., 2022; Ebrahimi et al., 2021; Inuwa-Dutse, 2021; Hasan et al., 2020; Agić and Vulić, 2019; Chowdhury et al., 2018; Kumar et al., 2019; Das et al., 2017; Adewumi, 2022), creoles in particular are extremely under-explored in spite of the prevalence of their usage globally (Lent et al., 2022b). Working more with this class of languages is an important step in ensuring that the benefits of NLP technology are more equitably distributed globally.

Additionally, the class of creole languages is a uniquely interesting point of study within the space of multilingual NLP. Though creoles like Jamaican Patois have distinct morphosyntactic features, they often share significant lexical overlap with the high-resource base languages from which they are derived. This makes it possible to study cross-lingual transfer between high-resource and

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5336 - 5349
December 7-11, 2022 ©2022 Association for Computational Linguistics

low-resource languages that are distinct, but share similar lexicons. In particular, JamPatoisNLI provides a benchmark for NLP researchers working to understand cross-lingual transfer to languages outside the training data of large pretrained multilingual models. Creole languages like Jamaican Patois have the unique property of being outside the pretraining data of these models, yet highly related to their base languages, which are present in the datasets used to train the models.[1]

JamPatoisNLI was constructed using both naturally occurring and newly constructed utterances of Jamaican Patois rather than through translation. This mitigates the problem of skewed cross-lingual transfer results which arises when the test dataset consists of translated examples but the training dataset does not (Artetxe et al., 2020). This also enhances the *ecological validity* (de Vries et al., 2020) of the dataset, as it is grounded in real world usage of the language and is thus a more relevant, realistic benchmark. These two features mean that work done with the dataset will be particularly useful for moving towards developing technologies for speakers of the language.

We run studies on JamPatoisNLI transferring from monolingual English BERT, multilingual BERT, monolingual English RoBERTa and multilingual XLM-RoBERTa, finetuned on the Multi-NLI dataset, in zero-shot and few-shot settings. We find that monolingual English RoBERTa (76.50%) and multilingual XLM-RoBERTa (75.17%) achieve similar accuracies when we use the entire few-shot JamPatoisNLI training dataset with 250 examples for further finetuning. We also find that the monolingual English BERT model (66.17 %) and the multilingual BERT model (65.33 %), achieve similar accuracies when we use the entire few-shot JamPatoisNLI training dataset. In our experiments, the RoBERTa-based models strongly outperform the BERT-based models. Additionally, we find that few-shot performance on JamPatoisNLI increases much faster (with respect to the number of few-shot training examples) than on languages in AmericasNLI, which have no strong connection to a high-resource language (Ebrahimi et al., 2021). Lastly, we run qualitative experiments which leverage the relatedness between Jamaican Patois and English to understand which differences between the languages boost or inhibit the effectiveness of cross-lingual transfer.

We hope that JamPatoisNLI prompts long-term research into building NLP tools that consider the particular difficulties and opportunities of NLP for Jamaican Patois and creole languages in general.

2 Related Work

Natural Language Inference Datasets. Natural language inference (NLI), or recognizing textual entailment, is a standard benchmark task for natural language understanding (Consortium et al., 1996; Dagan et al., 2005; Storks et al., 2019).

The input to the task is a pair of sentences: the premise and the hypothesis. The goal is to output a label – entailment, neutral or contradiction – to describe the relationship between the pair. Various approaches have been used to create NLI corpora. The Stanford NLI (SNLI) (Bowman et al., 2015), Multi-NLI (MNLI) (Williams et al., 2018) and Adversarial NLI (ANLI) (Williams et al., 2020) English datasets, esXNLI Spanish dataset (Artetxe et al., 2020) Original Chinese Natural Language Inference (OCNLI) dataset (Hu et al., 2020) and code-mixed Hindi-English dataset (Khanuja et al., 2020) all consist of a mixture of pre-existing sentences and crowdsourced sentences. In the Japanese Realistic Textual Entailment Corpus, a collection of pre-existing sentences are filtered and paired using machine learning methods then manually annotated with labels (Yanaka and Mineshima, 2021).

Other NLI corpora have been made using translation techniques. The Natural Language Inference in Turkish (NLI-TR) dataset (Budur et al., 2020) was created using Amazon Translate on SNLI and MNLI. The Cross-Lingual NLI (XNLI) Corpus (Conneau et al., 2018) was created by collecting and crowd-sourcing 750 examples then hiring human translators to translate the sentences into 15 languages. Extensions of this dataset to low-resource languages such as AmericasNLI (Ebrahimi et al., 2021) and IndicXNLI (Aggarwal et al., 2022) have been created using human and machine translation methods. However, subsequent research has found that translation-based approaches to creating datasets can introduce subtle artifacts which can lead to skewed accuracies for cross-lingual transfer methods (Artetxe et al., 2020). JamPatoisNLI mitigates this problem by using original rather than translated examples.

In spite of the examples given above, generally,

[1] In large web scrapes, there likely is some Jamaican Patois language in the resulting text, but it is not, e.g., one of the languages with a Wikipedia large enough to be included in Multilingual BERT.

there is a relative dearth of datasets and research into methods for low-resource languages across NLI and other tasks. Low-resource languages can be defined as those which are 'less studied, resource scarce, less computerized, less privileged, less commonly taught or low density' (Magueresse et al., 2020).

Creole Languages in NLP. Creole languages are typically low-resource. These languages arise through the process of *creolization* of another class of languages called pidgins. Pidgins emerge as a result of contact between two or more groups of speakers which do not have a common language. A pidgin evolves to become a creole when it becomes the native language of the children of its speakers (Muysken et al., 1995).[2]

Within the NLP community, a few datasets for different tasks have been created for creoles using a variety of methods. NaijaSenti is a Twitter human-annotated sentiment analysis dataset which is partly comprised of 14,000 tweets in Nigerian-Pidgin or Naija, which is an English-based creole (Muhammad et al., 2022). The authors find that code-switching between these languages and English is a common feature in the dataset. They explore language adaptive finetuning and zero-shot cross lingual transfer from multilingual pretrained models, and achieve promising results. Cross-lingual Choice of Plausible Alternatives (XCOPA) (Ponti et al., 2020) is a multilingual dataset for causal common sense reasoning in 11 languages, one of which is Haitian Creole, that was created by translating English COPA. The authors find that across the languages in the dataset, translation based-approaches outperform methods which employ multilingual pretraining and finetuning. A part-of-speech tagging and dependency parsing corpus for Colloquial Singaporean English (Singlish), an English-based creole, has also been created (Wang et al., 2017) and further expanded (Wang et al., 2019) using the Universal Dependencies (Nivre et al., 2020) scheme. The dataset was created by crawling pages on online Singaporean forums.

Other work has also explored using machine learning methods for identifying and generating creole text. Chang et al. (2022) use contrastive learning to finetune BART (Lewis et al., 2019) so that the model produces novel dialogue texts in Naija and Yaounde (both English-based creoles). Soto (2020) uses a FastText (Joulin et al., 2016) based supervised classifier to identify instances of sentences in Guadeloupean Creole within a multilingual dataset.

The use of machine learning models on creole languages has also been investigated. Lent et al. (2021) find that standard language models work better than distributionally robust ones on creoles, which shows that these languages are relatively stable. Lent et al. (2022a) show that ancestor-to-creole transfer is non-trivial.

3 Jamaican Patois

3.1 Description of the Language

Jamaican Patois (or Jamaican Creole) is an English-based creole spoken by over 3 million inhabitants on the island and by Jamaicans across the diaspora globally (Mair, 2003). Jamaican Patois resulted from contact between enslaved Africans brought to the island in the 17th century and British colonists. Because it is a hybrid of the languages spoken by the two groups of people that came in contact, it exists on a continuum that ranges from more dissimilar to less dissimilar to English (Davidson and Schwartz, 1995). The terms for the classes in the continuum are the acrolect (variations which are closest to English), the basilect (variations which are furthest from English) and the mesolect (variations which are in between) (Patrick, 2019)

Examples of each are shown in Table 1.

Class	Example
Basilect	Me a nyam di bickle weh dem gi mi.
Mesolect	Me a eat di food weh dem gi mi.
Acrolect	I'm eating the food that they gave me.

Table 1: Different translations of 'I'm eating the food that they gave me' in Jamaican Patois. The basilectal extreme of the continuum consists of words that are nearly exclusively non-English. On the acrolectal extreme of the spectrum (or Jamaican Standard English), the example is identical to English.

3.2 Relevant Linguistic Features

Unstandardized Orthography. Jamaican Patois is primarily a spoken language. Though there have been efforts to develop a formal writing system for the language, none that have been developed are widely used by speakers of Patois.

Instead, speakers use spelling patterns that reflect how words in Patois are pronounced. This is

[2] We discuss the process of creolization for Jamaican Patois further in Section 3.

illustrated in Table 2. In the table, *'I want'* is spelt both *'Me wah'* and *'Mi waa'*: though the phrases yield similar pronunciations, different spellings are used.

Jamaican Patois	English
Me wah bawl.	I want to cry.
Mi waa cook.	I want to cook.

Table 2: Example of varied spelling of Patois words present in the dataset.

Vocabulary Overlap with English. Since Jamaican Patois is English-based, there is a high degree of overlap between the vocabularies used by the two languages, in spite of differences in spelling, tense and structure.

We present an example of this in the quote below. Strictly non-English vocabulary (including words such as 'a' that have different meanings in English) which are highlighted in bold, account for less than one-third of the words in the sentence.

> `It look like more tourist start come since` **`dem`** `loosen up` **`di`** `restrictions` **`dem`**`.` **`Mi`** `frighten` **`fi`** `see how` **`di`** `beach full` **`wen`** **`mi`** `go` **`a`** `Negril` **`weh`** `day.`

Therefore, JamPatoisNLI will be useful for evaluating the efficacy of methods for linguistic transfer in scenarios where there is a high degree of overlap between the source and target language.

Negation. Common markers of negation used in Jamaican Patois and their English equivalents which feature in the dataset are presented in Table 3. Examples of these markers in the dataset are presented in Table 17 in the Appendix.

Negation markers are important linguistic features in the context of NLI datasets, as their presence and interaction with other sentence components are highly relevant to the determination of the right classification for a given textual entailment example (Gururangan et al., 2018).

Jamaican Patois	English
nuh	not/don't/doesn't
cyaa/cyaan	can't
neva	never

Table 3: Markers of negation in Jamaican Patois.

4 Constructing JamPatoisNLI

For each example in the dataset, we pulled the premise from a pre-existing text source. Then, a label was randomly selected and a corresponding hypothesis was written by the first author, who speaks and writes Jamaican Patois fluently. Our methodology mirrors that of both MNLI (Williams et al., 2018) and ANLI (Williams et al., 2020).

JamPatoisNLI consists of 650 examples split across training, development and validation. Statistics for the corpus are shown in Table 5. A limited availability of native speakers to construct and annotate a large number of examples is a current problem in low-resource NLP (Magueresse et al., 2020). However, for the purposes of our experiments, the sizes of the training, validation and testing sets are sufficient for exploring few-shot finetuning techniques and obtaining useful signals about the effectiveness of different methods.

4.1 Premise Collection

Since Jamaican Patois is primarily a spoken language, there is a limited number of textual sources of Patois that are readily available online. However, Patois speakers regularly use the language for communication on social media, and in literature. These are the sources that were used for the premises in the dataset. Around 97% of examples are drawn from Twitter and the remaining examples are drawn from a cultural website, jamaicans.com, and from literature by Jamaican poets, Dr. Louise Bennett-Coverley and Shelley Sykes-Coley. The number of examples per source is outlined in Table 13 in the Appendix.

This method of construction also makes the dataset less prone to effects from translation artifacts which can skew the effectiveness of different cross-lingual transfer techniques. Artetxe et al. (2020) find that when the test dataset is made using translated examples, there is a slight overestimation of the cross-lingual transfer gap as well as the efficacy of the `TRANSLATE-TRAIN`[3] technique, and an underestimation of the efficacy of the `TRANSLATE-TEST`[4] technique. None of these effects are present when the test dataset is composed of original examples which were not created through translation. Additionally, because the

[3] The `TRANSLATE-TRAIN` technique involves translating the training dataset to the target language.

[4] The `TRANSLATE-TEST` technique involves translating the testing dataset to the source language.

Premise	Label	Hypothesis
I decided that Christmas haffi ketch me inna good mood!	**entailment** E E	Me determined fi happy wen Christmas come!
A dem fi get the money	**contradiction** C C	Dem nuh deserve di money
mi must make chicken alfredo when mi go home doe	**neutral** N N	mi love fi eat chicken alfredo
Raisin a get soak in a red label wine fi make cake	**neutral** C N	Mi granny nuh normally mek har cake dem wid raisin
I was in juicy beef and yuh know say mi stress out til mi phone drop	**entailment** E E	Mi phone drop wen mi did deh inna juicy beef

Table 4: Random sample selected from the 100 double annotated examples in the corpus, with their gold labels and validation labels (abbreviated E, N, C) by each of the annotators.

Statistic	Ent.	Neu.	Con.	Total
#Train	84	83	83	**250**
#Dev	66	67	67	**200**
#Test	67	66	67	**200**
Avg. Premise Length	12.2	13.6	11.8	**12.5**
Avg. Hypothesis Length	10.3	11.9	10.7	**11.0**
#Distinct Words	1210	1401	1187	**2612**

Table 5: Statistics across the 650 examples in the dataset, by class and in aggregate.

Metric	Accuracy	Counts
Fleiss K	88.99%	100
% Accuracy	89.00%	100
Neutral % Accuracy	75.76%	33
Entailment % Accuracy	100.00%	34
Contradiction % Accuracy	90.91%	33

Table 6: Inter-annotator agreement. We count a classification as accurate if both annotators agreed with the original annotations in the dataset.

premises of JamPatoisNLI are drawn from natural occurrences of Jamaican Patois written by various speakers of the language, the dataset better reflects the natural writing patterns of speakers than those created using machine or human translation techniques.

4.2 Hypothesis Construction

The set of hypotheses in the corpus is comprised of novel sentences constructed by our first author, who is a native speaker of Jamaican Patois. For each premise, a corresponding hypothesis was written so that the pair's classification would be either `entailment`, `neutral` or `contradiction`. The criteria used for assignment of pairs to each class is shown in Figure 4 in the Appendix.

The constructed hypothesis in each example mimics the diverse spelling conventions and writing patterns used in the corresponding pre-existing premise. As such, the non-standardized nature of Jamaican Patois is reflected in both the collected and constructed sentences in the dataset.

In order to maximize the linguistic diversity of examples in the dataset, each premise was used to generate a single hypothesis (rather than three hypotheses generated per premise, which was done for MNLI (Williams et al., 2018)).

4.3 Label Validation

A random sample of 100 sentence pairs evenly distributed across the three classes was double annotated by fluent speakers of Jamaican Patois. We recruited volunteer annotators by reaching out to friends and colleagues. The labelling criteria given to the annotators were the same as those used to generate the hypotheses, and are outlined in Appendix Figure 4. In Table 6, we present statistics for inter-annotator agreement for these examples. The Fleiss Kappa accuracy for the dataset was 88.99% while the percentage accuracy was 89.00%.

5 Experiments and Results

Across our experiments, our goals are to:

1. Provide benchmarks for JamPatoisNLI thus determining the difficulty of the dataset and effectiveness of cross-lingual transfer.

2. Compare the effectiveness of cross-lingual transfer on JamPatoisNLI (a language that is *related* to language(s) present in the training corpus of each of the pretrained models we examine), to cross-lingual transfer on AmericasNLI (which contains languages that are

unrelated to any language(s) present in the training corpus of each pretrained model).

3. Leverage the nature of Jamaican Patois as a creole to further understand cross-lingual transfer.

The experiments that we conduct are done in the zero-shot and few-shot settings.

5.1 General Setup

In our experiments, we use English BERT, multilingual BERT (Devlin et al., 2018), English RoBERTa (Liu et al., 2019) and XLM-RoBERTa (Conneau et al., 2019a) as our base pretrained models. We use a two-layer perceptron with ReLU activations for the classification head, and first finetune on the MNLI training dataset. We use cased and uncased versions of each BERT-based pretrained model, and experiment with frozen and unfrozen versions,[5] for a total of eight types of BERT-based models. For our RoBERTa-based models, we also experiment with frozen and unfrozen versions for a total of four types of RoBERTa-based models. Throughout our experiments with the twelve model types, we make comparisons among the BERT-based models and the RoBERTa-based models separately.

To select the twelve MNLI finetuned models that we use for our few-shot experiments, we conduct a hyperparameter search over dropouts in the range [0.2, 0.5], batch sizes in the range [8, 32], learning rates in the range [1e-05, 1e-06] and epoch counts in the range [2, 10] and pick those that achieved reasonable accuracies on the MNLI development dataset (above 86% for unfrozen models and above 62% for frozen models).

Among the twelve selected models finetuned on MNLI, we evaluate the zero-shot and few-shot performance on each of our target datasets to determine which model types produce the highest accuracy. To compare the types of models, we fix the hyperparameters to the values in Table 16 in the Appendix, and average over three experiments with different seeds. Then, from among the eight finetuned BERT-based models, we pick the type that achieved the highest scores for the maximum number of few-shot training examples for each our validation datasets (JamPatoisNLI and AmericasNLI). We also do the same for the four finetuned RoBERTa-based models.

Hyperparameter	Best Model on JamPatoisNLI	Best Model on AmericasNLI
Finetune epoch ct.	5	5
Finetune batch size	16	16
Finetune learning rate	1e-05	1e-05
Finetune dropout	0.3	0.3
Few shot # of iter.	200	100
Few shot batch size	16	8
Few shot learning rate	5e-05	1e-05
Few shot dropout	0.25	0.25

Table 7: Final hyperparameters for best BERT-based model on JamPatoisNLI (`bert-uncased-unfrozen`) and AmericasNLI (`mbert-cased-unfrozen`).

Hyperparameter	Best Model on JamPatoisNLI	Best Model on AmericasNLI
Finetune epoch ct.	3	5
Finetune batch size	32	16
Finetune learning rate	1e-05	1e-05
Finetune dropout	0.2	0.3
Few shot # of iter.	200	100
Few shot batch size	16	16
Few shot learning rate	1e-05	1e-05
Few shot dropout	0.25	0.25

Table 8: Final hyperparameters for best RoBERTa-based model on JamPatoisNLI (`roberta-unfrozen`) and AmericasNLI (`xlm-unfrozen`).

After we select the best out of the model types among the models finetuned on MNLI and further finetuned on the target fewshot datasets, we perform a final hyperparameter sweep. Tables 7 and 8 show the final set of hyperparameters that we arrived at after we conducted our sweep for the best models on the JamPatoisNLI and AmericasNLI validation sets among our BERT-based models and RoBERTa-based models.

In our few-shot finetuning setup, we select one example from each class for each "shot". For instance, using this convention, two-shot finetuning involves finetuning using six examples in total: two from each of the three NLI classes. Additionally, during few-shot finetuning, we keep all layers of the base model unfrozen.

5.2 Benchmarks for JamPatoisNLI

Setup. For JamPatoisNLI, the best BERT-based model type was the unfrozen uncased English BERT model (`bert-uncased-unfrozen`) based on accuracies on the validation set. Using the hyperparameters in Table 7, we also make comparisons to a hypothesis only baseline

[5] In our frozen model, all parameters of the pretrained base models are fixed during finetuning so that only the NLI classification head is updated, while for our unfrozen models, all model parameters are allowed to update.

# of Fewshot Class Triples	Maj. Base.	Hyp. Only Base. (bert-uncased-unfrozen)	bert-uncased-unfrozen	mbert-uncased-unfrozen	roberta-unfrozen	xlm-unfrozen
0	33.50	38.50	56.00	50.00	67.50	56.00
1	33.50	38.17	54.50	52.17	68.17	57.50
2	33.50	37.17	56.83	53.33	69.17	58.17
4	33.50	37.00	51.00	52.33	66.83	57.67
8	33.50	35.83	52.17	51.17	68.83	57.50
16	33.50	38.83	56.17	53.50	70.17	58.83
32	33.50	38.50	61.17	63.83	73.00	70.00
64	33.50	46.33	64.50	65.17	76.33	72.50
83	33.50	43.33	66.17	65.33	76.50	75.17

Table 9: Zero-shot and few-shot accuracies for different models evaluated on JamPatoisNLI averaged over three experiments with different seeds. The best models were chosen based on results for the validation set.

(`bert-uncased-unfrozen`), as well as the best multilingual BERT-based model on JamPatoisNLI, which was the unfrozen uncased multilingual BERT model (`mbert-uncased-unfrozen`).

The best RoBERTa-based model type was the unfrozen English RoBERTa model (`roberta-unfrozen`). We also include results for the best multilingual RoBERTa-based model on the dataset, which was the unfrozen XLM-RoBERTa model (`xlm-unfrozen`). The hyperparameters that we used are listed in Table 8.

Results. Our results on the test set are presented in Table 9. We found that with the maximum number training of examples, `bert-uncased-unfrozen` and `mbert-uncased-unfrozen` had relatively similar accuracies when all few-shot examples were used (66.17% and 65.33% respectively). We also found that `roberta-unfrozen` and `xlm-unfrozen` achieve similar accuracies on the full fewshot dataset (76.50% and 75.17%) respectively.

The two RoBERTa-based models significantly outperformed the two BERT-based models – in fact, the zero-shot accuracy on the `roberta-unfrozen` model (67.50%) outperforms both BERT based models when they are finetuned on the full few-shot dataset.

For our best model (`xlm-unfrozen`), the standard deviation in percentage accuracy for the maximum number of few-shot examples across ten experiments was 0.75% when evaluated on the validation set and 1.43% when evaluated on the test set.

5.3 Comparisons with AmericasNLI

Setup. A natural comparison point for JamPatoisNLI is AmericasNLI (Ebrahimi et al., 2021) as it is also a low-resource NLI dataset. However, unlike Jamaican Patois, the languages in the corpus are not closely related to any high-resource languages for which there are large pretrained language models or large natural language inference training datasets. In particular, the languages in AmericasNLI do not belong to the same family as any of the languages in the two most commonly used multilingual pretrained language models – multilingual BERT (Devlin et al., 2018) and XLM-R (Conneau et al., 2019b). JamPatoisNLI is *unseen* from the perspective of existing pretrained monolingual or multilingual models but *related* to the source language(s) involved in transfer learning, whereas AmericasNLI is both *unseen* and *unrelated*.

For our experiments, we use five of the languages in the AmericasNLI dataset, and create a randomly selected 250-200-200 train-dev-test split from among the examples in the original development dataset for each language (shown in Table 14 in the Appendix) to mirror the number of examples present in each of the splits in JamPatoisNLI.

For the AmericasNLI languages, the best BERT-based model type based on results on the validation set was the unfrozen cased multilingual BERT model (`mbert-cased-unfrozen`). The best RoBERTa-based model type was the unfrozen XLM-RoBERTa model (`xlm-unfrozen`).

Results. We present the results of our experiments on the test set in Table 10. We found that there was a significant gap in accuracies on JamPatoisNLI and AmericasNLI. Across all experiments, both zero-shot and few-shot accuracies for the JamPatoisNLI dataset exceeded those for the AmericasNLI dataset. The best JamPatoisNLI model achieved a zero-shot accuracy of 67.50% while the best AmericasNLI model achieved a zero-shot

	Avg. AmericasNLI Accuracy		Patois Accuracy	
Num.	**mbert-cased-unfrozen**	**xlm-unfrozen**	**bert-uncased-unfrozen**	**roberta-unfrozen**
0	42.00	39.60	56.00	67.50
1	41.83	39.17	54.50	68.17
2	42.67	39.50	56.83	69.17
4	42.67	40.03	51.00	66.83
8	42.70	39.93	52.17	68.83
16	43.63	42.77	56.17	70.17
32	46.40	46.07	61.17	73.00
64	48.87	47.40	64.50	76.33
83	49.23	48.83	66.17	76.50

Table 10: Test set accuracies for best BERT-based and RoBERTa-based models on the JamPatoisNLI dataset (`bert-uncased-unfrozen`, `roberta-unfrozen`) and on the AmericasNLI dataset (`mbert-cased-unfrozen`, `xlm-unfrozen`). Experiments are averaged over three seeds and the best models were chosen based on results for the validation set.

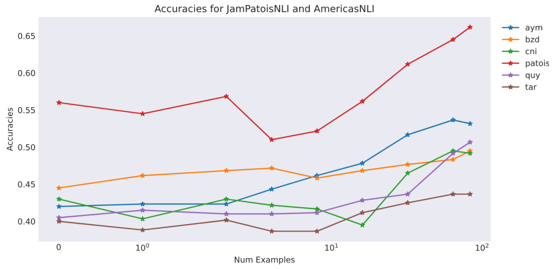

Figure 2: Plots for the best AmericasNLI model (`mbert-cased-unfrozen`) on each language, and the best JamPatoisNLI model (`bert-uncased-unfrozen`). Experiments are averaged over three seeds and the best models were chosen based on results for the val. set.

accuracy of 42.00% (both compared to a 33.50% majority baseline).

This shows that the language relatedness between Jamaican Patois and English significantly boosts the effectiveness of cross-lingual transfer learning even in the zero-shot case. For the few-shot setting, the highest accuracy achieved on the JamPatoisNLI dataset was 76.50%. The highest average accuracy achieved on the AmericasNLI dataset was 49.23%.

The plots comparing the best JamPatoisNLI model to the best AmericasNLI model on each of the respective datasets for BERT-based models and RoBERTa-based models are shown in Figures 2 and 3. For the BERT-based models, we see that cross-lingual transfer augmented by few-shot learning is quite effective for JamPatoisNLI, whereas the gains for AmericasNLI languages are rather modest. Tabulated results for these experiments can be found in Appendix Tables 18 and 19.

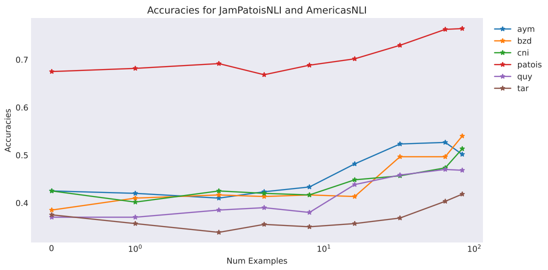

Figure 3: Plots for the best AmericasNLI model (`xlm-unfrozen`) on each language, and the best JamPatoisNLI model (`roberta-unfrozen`). Experiments are averaged over three seeds and the best models were chosen based on results for the val. set.

5.4 Experiments with Transitioning from Jamaican Patois to English

Setup. A key characteristic of Jamaican Patois is that it exists on a spectrum that ranges from highly dissimilar to English (the basilect), to highly similar to English (the acrolect). We experiment with 83-shot classification (the full set of examples in our few-shot training dataset) on an augmented test dataset derived from pairs that were incorrectly classified by at least two of the three models in our original few-shot experiments. To construct this dataset, we picked a single example for each type of misclassification with respect to the three NLI labels, for a total of 6 examples from the original dataset (which mostly fell on various points on the mesolectal range of the creole spectrum). We then wrote English translations for each of these examples (which would fall on the acrolectal end of the creole spectrum) and hand-wrote intermediate translations between them that are all valid Jamaican Patois to qualitatively study whether (and for what changes) along the path the label becomes correct. We conduct few-shot finetuning using our original training set for three models with different seeds using the parameters for the best BERT-based JamPatoisNLI model (`bert-uncased-unfrozen`), listed in Table 7.

Results. We present a qualitative example of this experiment in Table 11. Here, changing the verb from Jamaican Patois to English caused the models to switch to the correct classification. The three models switched to the correct prediction for a

Change	Premise	Hypothesis	Tgt.	M1	M2	M3
-	Any day **mi** master LumaFusion, mi lef my work.	As soon as **mi** good wid LumaFusion, mi a quit mi job	E	C	C	C
Pronoun: mi → I	Any day **I** master LumaFusion, mi **lef** my work	As soon as **I'm** good wid LumaFusion, mi **a quit** my job	E	C	C	C
Verb: lef/quit → leaving/quitting	Any day I master LumaFusion, **mi leaving** my work	As soon as I'm good wid LumaFusion, **mi quitting** this job	E	**E**	**E**	**E**
Pronoun: mi → I	**Any day** I master LumaFusion, **I'm** leaving my job	As soon as I'm good **wid** LumaFusion, **I'm** quitting my job.	E	**E**	**E**	**E**
Determiner/Preposition: Any day/wid → The day that/with	**The day that I master LumaFusion, I'm leaving my job.**	**As soon as I'm good with LumaFusion, I'm quitting my job.**	**E**	E	E	E

Table 11: Sample from Jamaican Patois to English transition dataset. The final example is in English, and we present predictions made by three models finetuned with our Patois few-shot training dataset using the parameters for the best JamPatoisNLI model in Table 7.

.

change prior to the full translation of the Jamaican Patois example to English for all but one of the originally misclassified examples in our experiments.

6 Discussion

We see that the relatedness between Jamaican Patois and English strongly contributes to the effectiveness of cross-lingual transfer in both zero-shot and few-shot settings. Additionally, although natural language inference is a higher order reasoning task, our models achieved relatively high accuracy on the JamPatoisNLI dataset by learning the task from MNLI examples in English.

A natural question that arises based on these results, is whether vocabulary overlap is the primary factor that led to the boost in effectiveness of transfer learning in these experiments, or whether a higher order notion of similarity is a larger factor. Comparing zero-shot and few-shot accuracies for other languages that are closely related to English but do not share the same degree of vocabulary overlap as an English-based creole (such as German) might be an interesting line of future research.

Interestingly, though Jamaican Patois developed as a result of contact between speakers of English and speakers of West African languages (some of which are present in multilingual BERT's and XLM-RoBERTa's training corpus), the multilingual models were not more effective base pretrained language models than the monolingual models. Another possible direction for future research might be to determine whether there are methods that allow for more effective leveraging of the multilingual characteristic of the models during finetuning for creole target languages.

7 Conclusion

JamPatoisNLI is a natural language inference dataset in an English-based creole, constructed from existing and novel examples of Jamaican Patois. Our experiments show that the language's relatedness to English significantly boosts the effectiveness of cross-lingual transfer, even for the higher order task of natural language inference in both zero-shot and few-shot settings. We hope that the creation of this dataset encourages further research in the field on methods to improve cross-lingual transfer for creole target languages, and the creation of other low-resource language and creole language datasets.

Acknowledgements

We thank Roxanne Dobson, Ghawayne Calvin, Danielle Roberts, Khaesha Brooks, Dominique Lyew and Ana-Katrina Donaldson for volunteering to be dataset annotators. We also thank Prof. Christopher Potts (cgpotts@stanford.edu) for comments on the paper. RA was partially supported by a Siebel Scholarship, a Google Generation Scholarship and the SWE Motorola Solutions Foundation Engineering Scholarship. JH was supported by an NSF Graduate Research Fellowship under grant number DGE-1656518.

8 Limitations

One limitation of our research is related to the fact that Jamaican Patois is a low-resource language. The size of the dataset splits (particularly, the validation and test sets) are much smaller than those of high-resource language datasets.

Further, the differences observed between the AmericasNLI and JamPatoisNLI datasets are not necessarily solely due to differences in language similarity to the source languages: another contributing factor might be differences in difficulty for the two datasets.

References

Tosin Adewumi. 2022. Itakúroso: Exploiting cross-lingual transferability for natural language generation of dialogues in low-resource, African languages. In *3rd Workshop on African Natural Language Processing*.

Divyanshu Aggarwal, V. Gupta, and Anoop Kunchukuttan. 2022. IndicXNLI: Evaluating multilingual inference for Indian languages. *ArXiv*, abs/2204.08776.

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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. *CoRR*, abs/2004.04721.

Peter Bakker and Aymeric Daval-Markussen. 2013. Creole studies in the 21st century: A brief presentation of the special issue on creole languages. *Acta Linguistica Hafniensia*, 45(2):141–150.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326.

Emrah Budur, Riza Özçelik, Tunga Güngör, and Christopher Potts. 2020. Use of machine translation to obtain labeled datasets for resource-constrained languages. *CoRR*, abs/2004.14963.

Ernie Chang, Jesujoba Oluwadara Alabi, David Ifeoluwa Adelani, and Vera Demberg. 2022. Dialogue pidgin text adaptation via contrastive fine-tuning. In *3rd Workshop on African Natural Language Processing*.

Koel Dutta Chowdhury, Mohammed Hasanuzzaman, and Qun Liu. 2018. Multimodal neural machine translation for low-resource language pairs using synthetic data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 33–42.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. *CoRR*, abs/1809.05053.

The Fracas Consortium, Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the framework.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Arjun Das, Debasis Ganguly, and Utpal Garain. 2017. Named entity recognition with word embeddings and Wikipedia categories for a low-resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(3):1–19.

Cecelia Davidson and Richard G Schwartz. 1995. Semantic boundaries in the lexicon: Examples from Jamaican Patois. *Linguistics and Education*, 7(1):47–64.

Harm de Vries, Dzmitry Bahdanau, and Christopher D. Manning. 2020. Towards ecologically valid research on language user interfaces. *CoRR*, abs/2007.14435.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, Gustavo A. Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *CoRR*, abs/2104.08726.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. *CoRR*, abs/1803.02324.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. *arXiv preprint arXiv:2009.09359*.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online. Association for Computational Linguistics.

Isa Inuwa-Dutse. 2021. The first large scale collection of diverse Hausa language datasets. *arXiv preprint arXiv:2102.06991*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomás Mikolov. 2016. Fasttext.zip: Compressing text classification models. *CoRR*, abs/1612.03651.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020. A new dataset for natural language inference from code-mixed conversations. *CoRR*, abs/2004.05051.

Gokul Karthik Kumar, Abhishek Singh Gehlot, Sahal Shaji Mullappilly, and Karthik Nandakumar. 2022. Mucot: Multilingual contrastive training for question-answering in low-resource languages. *arXiv preprint arXiv:2204.05814*.

Rashi Kumar, Piyush Jha, and Vineet Sahula. 2019. An augmented translation technique for low resource language pair: Sanskrit to Hindi translation. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 377–383.

Elmurod Kuriyozov, Sanatbek Matlatipov, Miguel A Alonso, and Carlos Gómez-Rodríguez. 2022. Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek. In *Language and Technology Conference*, pages 232–243. Springer.

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

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022b. What a creole wants, what a creole needs. ArXiv preprint arXiv:2206.00437.

Heather C. Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. 2021. On language models for creoles. *CoRR*, abs/2109.06074.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264.

Christian Mair. 2003. Language, code, and symbol: The changing roles of Jamaican Creole in diaspora communities. *AAA: Arbeiten aus Anglistik und Amerikanistik*, 28(2):231–248.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Abdullahi Salahudeen, Aremu Anuoluwapo, Alípio Jeorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. *CoRR*, abs/2201.08277.

Pieter Muysken, Norval Smith, et al. 1995. The study of pidgin and creole languages. *Pidgins and creoles: An introduction*, pages 3–14.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis M. Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. *CoRR*, abs/2004.10643.

Peter Patrick. 2019. Jamaican Creole. In *The Mouton World Atlas of Variation in English*, pages 126–136. De Gruyter.

Edoardo Maria Ponti, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. *CoRR*, abs/2005.00333.

Suzanne Romaine. 2017. *Pidgin and creole languages*. Routledge.

William Soto. 2020. Language Identification of Guadeloupean Creole. In *2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)*, pages 54–59, Montrouge (virtuel), France. CNRS.

Shane Storks, Qiaozi Gao, and Joyce Yue Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *ArXiv*, abs/1904.01172.

Hongmin Wang, Jie Yang, and Yue Zhang. 2019. From genesis to creole language: Transfer learning for Singlish Universal Dependencies parsing and POS tagging. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(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.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2020. ANLIzing the adversarial natural language inference dataset. *CoRR*, abs/2010.12729.

Hitomi Yanaka and Koji Mineshima. 2021. Assessing the generalization capacity of pre-trained language models through Japanese adversarial natural language inference. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 337–349, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *CoRR*, abs/2106.10199.

A Appendix

A.1 Finetuning with BitFit

BitFit is a sparse parameter efficient finetuning method introduced for use with small-to-medium sized training datasets which involves finetuning only the bias terms of a pretrained language model (Zaken et al., 2021). As an initial approach for few-shot finetuning, we experimented with using BitFit using the same hyperparameters described in our prior experiments (in Table 7) for the best JamPatoisNLI model (English BERT uncased unfrozen), but increasing the learning rate by one order of magnitude as the authors do in the paper to 5e-04.

In Table 12, we present the results for few-shot finetuning using the BitFit method (Zaken et al., 2021) in comparison with the vanilla finetuning method (in which all model parameters are left unfrozen). In the zero-shot setting and in the cases where there are a small number of few-shot examples, the two techniques perform similarly, but BitFit begins to underperform relative to the vanilla method with more few-shot examples.

Num Examples	Jam	Jam-BitFit
0	56.00	56.00
1	54.50	55.83
2	56.83	55.67
4	51.00	55.83
8	52.17	55.83
16	56.17	55.83
32	61.17	54.67
64	64.50	58.00
83	66.17	58.67

Table 12: Comparison for zero-shot and few-shot finetuning using BitFit and the vanilla finetuning technique. Experiments are averaged over three seeds, and are reported on the test dataset.

Source	Examples
Twitter	634
Anthology: Shelley Sykes-Coley	6
Poetry: Rt. Hon. Dr. Louise Bennett-Coverley	4
Online blog	6

Table 13: Sources for premises in the dataset.

Language	ISO	Family	Dev	Test
Aymara	aym	Aymaran	743	750
Asháninka	cni	Arawak	658	750
Bribri	bzd	Chibchan	743	750
Quechua	quy	Quechuan	743	750
Rarámuri	tar	Uto-Aztecan	743	750

Table 14: Languages used from the AmericasNLI dataset and the sizes of the original splits.

Hyperparameter	Values
Batch size	8, 16
Learning rate	1e-05, 5e-05
Number of iterations	100, 200

Table 15: Values used for few-shot hyperparameter sweep. Experiments are averaged over three seeds.

Hyperparameter	Value
Batch size	8
Learning rate	1e-05
Number of iterations	100
Dropout	0.25

Table 16: Hyperparameters used for model type selection. Experiments are averaged over three seeds.

Entailment.
(a) Given the premise, a reasonable reader would conclude that the hypothesis must also be true.
(b) The hypothesis is necessarily consistent with the premise.
(c) If a speaker holds the sentiment or opinion expressed in premise, then a reasonable reader would conclude that they also hold the sentiment or opinion expressed in hypothesis.

Contradiction.
(a) Given the premise, a reasonable reader would conclude that the hypothesis must be false.
(b) The hypothesis is necessarily inconsistent with the premise.
(c) If a speaker holds the sentiment or opinion expressed in premise, then a reasonable reader would conclude that they do not hold the sentiment or opinion expressed in hypothesis.

Neutral
(a) Given the premise, a reasonable reader would conclude that the hypothesis could be either true or false.
(b) The hypothesis is neither necessarily inconsistent nor necessarily consistent with the premise.
(c) If a speaker holds the sentiment or opinion expressed in premise, then a reasonable reader would conclude that it may or may not be true that they hold the sentiment or opinion expressed in hypothesis.

Figure 4: Labelling criteria used to generate each hypothesis based on the premise, and given as labelling guidelines to dataset validators.

Premise	**Hypothesis**	**Label**
Jason mi deh cook and me nah mek u mek di likkle bickle bun up!	Jason neva eat cook food from da restaurant deh inna im life	neutral
And if dem tek everything and all mi have a my breathe , mi happy same way	Nuh matta weh dem waa tek from mi glad as long as mi have life	entailment
Mi nuh bada waa get married... ever	Mi cyaa wait fi get married	contradiction

Table 17: Examples of negation markers in examples from each of the three classes in the dataset.

Num Examples	**aym**	**bzd**	**cni**	**quy**	**tar**	**jam**
0	42.00	44.50	43.00	40.50	40.00	56.00
1	42.33	46.17	40.33	41.50	38.83	54.50
2	42.33	46.83	43.00	41.00	40.17	56.83
4	44.33	47.17	42.17	41.00	38.67	51.00
8	46.17	45.83	41.67	41.17	38.67	52.17
16	47.83	46.83	39.50	42.83	41.17	56.17
32	51.67	47.67	46.50	43.67	42.50	61.17
64	53.67	48.33	49.50	49.17	43.67	64.50
83	53.17	49.50	49.17	50.67	43.67	66.17

Table 18: Zero-shot and few-shot plot for the best BERT-based AmericasNLI model (`mbert-cased-unfrozen`) accuracies for each language in the dataset and the best BERT-based JamPatoisNLI model (`bert-uncased-unfrozen`). Experiments are averaged over three seeds and the best models were chosen based on results for the validation set.

Num Examples	**aym**	**bzd**	**cni**	**quy**	**tar**	**jam**
0	42.50	38.50	42.50	37.00	37.50	67.50
1	42.00	41.00	40.17	37.00	35.67	68.17
2	41.00	41.67	42.50	38.50	33.83	69.17
4	42.33	41.33	42.00	39.00	35.50	66.83
8	43.33	41.67	41.67	38.00	35.00	68.83
16	48.17	41.33	44.83	43.83	35.67	70.17
32	52.33	49.67	45.67	45.83	36.83	73.00
64	52.67	49.67	47.33	47.00	40.33	76.33
83	50.17	54.00	51.33	46.83	41.83	76.50

Table 19: Zero-shot and few-shot plot for the best RoBERTa-based AmericasNLI model (`xlm-unfrozen`) accuracies for each language in the dataset and the best RoBERTa-based JamPatoisNLI model (`roberta-unfrozen`). Experiments are averaged over three seeds and the best models were chosen based on results for the validation set.