Jamaican Patois Constituency Treebank and Parser

Angelic McPherson
Dartmouth College
angelic.mcpherson.25@dartmouth.edu

Abstract

Natural Language Processing (NLP) models have become integral to today's society, however, they disproportionately serve high-resource languages such as English, Mandarin, and Spanish. This imbalance raises concerns about equity and bias regarding low-resource languages that lack the annotated data necessary to train NLP models. The lack of annotated syntactic data presents a challenge for NLP tasks like constituency parsing, which are foundational to many NLP systems. In this paper, we address the problem of building a constituency parser for Jamaican Patois, a low-resource language, using a cross-lingual approach. Here we show that a RoBERTa-based CRF constituency parser, pre-trained on English treebanks, can be effectively adapted to Jamaican Patois using only a small, manually annotated dataset of 135 sentences. Our results suggest that linguistic similarity between Jamaican Patois and English supports successful transfer parsing. This furthers prior research on low-resource language processing and methodologies leveraging related high-resource languages. Overall, this work contributes to the development of inclusive language technologies by offering methods for building syntactic tools for low-resource languages using state-of-the-art models.

1 Introduction

Data is the foundation of NLP models. As these models become increasingly integrated into everyday technology, a critical challenge has emerged: What happens when we lack sufficient data for a specific task? This concern has increased discussion regarding high-resource and low-resource languages. High-resource languages, such as English, Chinese, and Spanish, benefit from large annotated corpora and linguistic tools. Comparatively, low-resource languages lack the necessary datasets to support training and evaluation of NLP models, creating a significant barrier.

This barrier introduces systemic bias in NLP. Models trained primarily on high-resource languages often fail to generalize to low-resource languages. This results in poor performance and complete exclusion of underrepresented communities. For example, indigenous and other underrepresented communities, often report struggling with popular technologies, such as Siri, Alexa, autocorrect, etc. Speakers of low-resource languages are frequently under-served, as their languages are mishandled, misinterpreted, or ignored by existing technologies.

To address this challenge, this paper explores whether leveraging the linguistic relationship between a high-resource and a related low-resource language can help bridge the gap. In this paper, I investigate constituency parsing, a foundational NLP task, by adapting a CRF-based parser pre-trained on English to parse Jamaican Patois. Given that Jamaican Patois draws heavily on English as its lexifier language, this pairing has a theoretical basis for cross-lingual syntactic transfer (Campbell, 2020). This study contributes to NLP by testing whether linguistic proximity to a high-resource language can support and inform a model's understanding of low-resource linguistic properties.

1.1 Related Work

This work draws on research in constituency parsing, low-resource language processing, and creole linguistics to the methodology for Jamaican Patois.

Fast and accurate neural CRF constituency parsing by Zhang et al. proposes a fast and accurate neural CRF-based constituency parser that improves efficiency and performance through a novel batchification and backpropagation strategy, including a two-stage bracketing and labeling technique. Their model achieves parsing speeds exceeding 1,000 sentences per second on English and Chinese treebanks, and maintains or surpasses the accuracy of prior top-performing models (Zhang et al., 2020).

Constituency Parsing by Cross-Lingual Delexicalization by Kaing et al proposes a cross-lingual approach to constituency parsing for low-resource languages by leveraging delexicalized parsing techniques. They explored English, French, Chinese, Khmer, Japanese, Myanmar, and German, paired with their respective treebanks. By training delexicalized parsers on source languages and evaluating on low-resource target languages, they show that selecting an appropriate source language can produce results comparable to state-of-the-art models (Kaing et al., 2021).

Chapter 10, Section 4.1 of Historical Linguistics by Campbell examines the linguistic properties of creoles and pidgins, and evaluates the validity of traits commonly associated with these languages. Campbell presents arguments both supporting and challenging these assumptions, ultimately suggesting that creoles and pidgins may exhibit greater complexity than previously acknowledged by linguistic scholars (Campbell, 2020).

2 Methodology

Dataset

To evaluate constituency parsing on Jamaican Patois, I constructed a dataset consisting of 135 sentences drawn from two major sources: Dancehall and Reggae lyrics, and the Jamaican Bible. Lyrics were retrieved using the Genius API and from Spotify and YouTube playlists. Sentences from the Bible were chosen from randomly selected books in the New Testament. All texts were manually cleaned to remove repeated lines and tokenized. Utilizing two sources allowed for increased linguistic complexity and topic diversity, from the inherent characteristics and differences of the data.

The dataset was then split into train, validation, and test sets with an approximate 70%, 20%, 10% split, shown in Table1.

Dataset	Sentences
Train	94
Validation	15
Test	26

Table 1: Sentence counts for each dataset split.

Dataset Example

The example below illustrates the parsing methodology for Jamaican Patois and shows a matching constituency grammar.

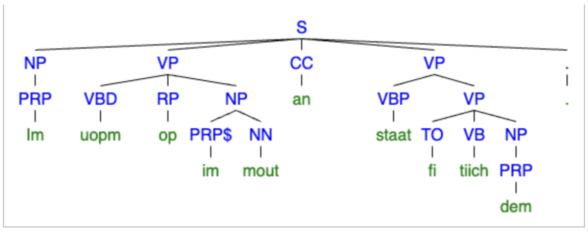

Figure 1: Constituency tree for example sentence.

Example Sentence: Im uopm op im mout an staat fi tiich dem.

Translation: He opened up his mouth and started to teach them.

I chose to adopt Penn Treebank-style constituency parsing and part-of-speech labeling for this project. While this paper presents only a single annotated example, it clearly demonstrates that Jamaican Patois shares similar, if not identical, grammatical structures with English. This is likely due to English serving as its lexifier language, meaning many of its linguistic characteristics are inherited directly from English (Campbell, 2020).

Grammar Rules
$S \rightarrow NP\,VP\,CC\,VP$
$VP \rightarrow VBD\,RP\,NP$
$VP \rightarrow VBP\,VP$
$VP \rightarrow TO\,VB\,NP$
$NP \rightarrow PRP$
$NP \rightarrow PRP\$\,NN$

Table 2: Grammar rules (excluding terminals) used for Jamaican Patois constituency parsing.

Constituency Parser

I utilized the SuPar CRF constituency parser, specifically the `crf-con-roberta-en` model, which is built on a RoBERTa encoder backbone with a Conditional Random Field (CRF) layer to predict syntactic tree structures. This parser was chosen due to its strong performance and its open-source release on GitHub[1] (Zhang et al., 2020). For replication, the seed value for each model is 31.

Evaluation Metrics

To assess the parser's performance, SuPar offers both labeled and unlabeled evaluation metrics. These include Unlabeled Complete Match (UCM) and Labeled Complete Match (LCM), which measure the percentage of predicted trees that match

[1] https://github.com/yzhangcs/parser

gold-standard trees exactly, with and without labels. It also reports Unlabeled and Labeled Precision, Recall, and F1 scores. For evaluation, unlabeled measures were prioritized over labeled to focus on general constituency parsing performance.

3 Results

Metric	JP Model	EN Model
Batch Size	4000	4000
BERT LR	5e-05	5e-05
Dropout	0.4	0.4
Epochs	15	15
Learning Rate	5e-07	5e-07
LR Interval	15	15
Update Steps	1	1
Warmup Steps	100	100
Weight Decay	0.2	0.001
Best Val F1	**0.1220**	**0.5980**

Table 3: Training hyperparameters and best validation F1 scores for JP (Jamaican Patois) and EN (English) models.

The best hyperparameter settings and corresponding validation F1 scores for each model are shown in Table 3. The English Pretrained model achieved a validation F1 of 59.80, while the Jamaican Patois model achieved only 12.20, resulting in a significant gap of 47.6 points. This trend is consistent in the test results. As shown in Table 4, The English Pretrained model achieved a UP score of 56.39, UR of 66.06, and UF of 60.84. Comparatively, the Jamaican Patois model only reached 31.70 UP, 49.64 UR, and 38.69 UF, showing a significant 25-point difference in UP, 17-point difference in UR, and 22-point gap in UF (rounded).

The English Pretrained model performance is below the Baseline English CRF model, whose UF is 94.80, which is expected. The Baseline model is trained on the English Penn Treebank, a high-resource dataset with over 40,000 sentences (Zhang et al., 2020). Therefore, the performance gap between the two is understandable. In fact, this suggests the English Pretrained model's performance on Jamaican Patois is significant due to the discrepancy in amount of training data (Table1).

These differences show that the English Pretrained model successfully leveraged its prior training on English to learn useful patterns in Jamaican Patois. The Jamaican Patois model, however, failed to generalize and had poor performance.

4 Discussion

4.1 Jamaican Patois Model

The Jamaican Patois model, built on the base CRF Constituency RoBERTa infrastructure, underperformed in comparison to other models. Manual inspection of the predicted trees revealed that the model defaulted to labeling nearly all constituent spans as noun phrases (NP), regardless of context. This behavior indicates the model failed to learn and generalize meaningful patterns from the training data.

As shown in Table 4, most metrics are notably low, with UCM, LCM near zero. However, recall values are somewhat higher. This is likely due to the model over-predicting NP constituent spans, leading to some coincidental matching with gold-standard constituent spans. This most likely inflated the recall score, suggesting it is not a reliable measure of genuine syntactic understanding in this case.

These results illustrate the challenges of applying NLP models like RoBERTa to low-resource languages. Although RoBERTa achieves strong results on high-resource benchmarks such as the Penn Treebank, it requires a substantial amount of labeled training data to perform well. Jamaican Patois and other low-resource languages lack this. Overall, this experiment reaffirms that, despite advancements, current NLP models still struggle under low-resourced contexts.

4.2 English Pretrained Model

To explore whether pretraining on a related language would help, I evaluated SuPar's `crf-con-roberta-en` model, pretrained on English treebanks. As English is the lexical base language of Jamaican Patois, this model provides a meaningful comparison. Baseline performance for this model on the Penn Treebank was included for reference.

The English pretrained model showed improved performance on Jamaican Patois data, with precision, recall, and F1 scores exceeding those of the Jamaican Patois model by on average approximately 20 percentage points. When comparing gold trees to predicted trees, we can clearly see the accuracy of the model for certain constituency spans, however Complete Match scores remained low.

These findings suggest that low-resource languages may benefit from models pretrained on their

Model	UCM	LCM	UP	UR	UF	LP	LR	LF
Baseline English CRF Model	50.08	47.56	94.89	94.71	94.80	94.16	93.98	94.07
English Pretrained Model (EN)	11.54	3.85	56.39	66.06	60.84	45.79	53.65	49.41
Jamaican Patois Model (JP)	3.85	3.85	31.70	49.64	38.69	20.28	31.75	24.75

Table 4: Comparison of constituency parsing performance on the test set across models using Unlabeled and Labeled Complete Match (UCM/LCM), Precision (UP/LP), Recall (UR/LR), and F1 (UF/LF). Baseline metrics were obtained using the SuPar CRF parser on the English Penn Treebank.

lexical base language. Despite grammatical differences between the two languages, the English pretrained model performed well and was able to gain linguistic insights from prior training (Campbell, 2020). While these models may not capture all grammatical nuances, they can still provide a strong starting point for syntactic parsing.

Ethical Considerations

This project has important considerations regarding dataset bias. The data is drawn primarily from two culturally specific domains; biblical texts and dancehall / reggae lyrics. These topics do not reflect the full range of everyday language use among Jamaican speakers, as these genres are heavily influenced by religious and sexual themes. This introduces a content bias that may affect the generalizability and neutrality of any NLP models trained on this data.

Therefore, this dataset bias can skew the behavior of NLP models, especially in tasks involving language generation, named entity recognition (NER), or sentiment analysis. Future users of this dataset should be aware of these limitations and avoid assuming full representativeness.

Limitations

A key limitation is the low-resource nature of Jamaican Patois, meaning limited annotated corpora. All constituent trees were manually annotated, a time-consuming process hindered by the lack of annotators. This resulted in a small dataset that may not encapsulate Jamaican Patois's fullness and diversity.

As mentioned previously, model performance was likely constrained by the limited training data. Future work should increase annotation (crowd-sourcing or semi-automated tools) to expand the dataset and support better model training.

Conclusion

This project evaluated constituency parsing performance for Jamaican Patois, a low-resource language, using a base CRF RoBERTa model and an English-pretrained version. The base model trained solely on Jamaican Patois data failed to generalize, defaulting to labeling most constituent spans as noun phrases and achieving very poor metrics. These results highlight the limitations of current NLP methods when applied to low-resource settings without sufficient annotated data.

In contrast, the English-pretrained model, while still underperforming compared to its English baseline, demonstrated significantly higher precision, recall, and F1 scores. This suggests that pretraining on a lexically similar high-resource language can offer meaningful benefits for syntactic tasks in low-resource languages.

These findings reinforce the importance of transfer learning and the potential for leveraging lexical relationships between languages.

Future Work

Future work should not only consider expanding the dataset, but also diversifying it to better capture the fullness of Jamaican Patois. This would solve issues of poor model performance, generalization, and bias in these projects.

Future work could also explore different creole languages with different lexical bases. For example, Haitian Creole and French could yield interesting results for comparison and further study.

References

Lyle Campbell. 2020. *Historical Linguistics: An Introduction*, 4th edition. Edinburgh University Press, Edinburgh, UK.

Hour Kaing, Chenchen Ding, Masao Utiyama, Eiichiro Sumita, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Constituency parsing by cross-lingual delexicalization. *IEEE Access*, 9:137926–137938.

Yu Zhang, Houquan Zhou, and Zhenghua Li. 2020. Fast and accurate neural CRF constituency parsing. In *Proceedings of IJCAI*, pages 4046–4053.