Simple Morphology, Complex Models: A Benchmark Study and Error Analysis of POS Tagging for Martinican Creole

Ludovic Mompelat
Department of Modern Languages and Literatures
University of Miami
Miami, FL, USA
lvm861@miami.edu

Abstract

Part-of-speech (POS) tagging is a foundational task in NLP pipelines, but its development for Creole languages remains limited due to sparse annotated data and structural divergence from high-resource languages. This paper presents the first POS tagging benchmarks for Martinican Creole (MC) as well as a linguistically motivated evaluation framework, comparing three fine-tuned transformer-based models (mBERT, XLM-Roberta, and CreoleVal). Rather than focusing solely on aggregate metrics, we perform detailed error analysis, examining model-specific confusion patterns, lexical disambiguation, and out-of-vocabulary behavior. Our results yield F1 scores of 0.92 for mBERT (best on the X tag and connector distinctions), 0.91 for XLM-Roberta (strongest on numeric tags and conjunction structures), and 0.94 for CreoleVal (leading on both functional and content categories and lowest OOV error rate). We propose future directions involving model fusion, targeted and linguistically motivated annotation, and reward-guided Large Language Models data augmentation to improve our current tagger. Our linguistically grounded error analysis for MC exposes key tagging challenges and demonstrates how targeted annotation and ensemble methods can meaningfully boost accuracy in under-resourced settings.

1 Introduction

Despite significant progress in multilingual language modeling (Qin et al., 2024; Huang et al., 2024), natural language processing (NLP) for Creole languages remains underdeveloped. This is largely due to the scarcity of annotated resources and the unique linguistic features of Creoles such as morphosyntactic restructuring and realignment (Mufwene, 2013), and frequent code-switching/code-mixing (Vaillant, 2023) which challenge existing models trained on high-resource languages (Mompelat et al., 2022). This paper introduces the first benchmark part-of-speech (POS) tagging dataset for Martinican Creole (MC) and presents a comparative evaluation of three transformer-based models: XLM-Roberta (Conneau et al., 2019), mBERT (Devlin et al., 2018), and CreoleVal (Lent et al., 2024), the latter being a recent adaptation specifically designed for Creole NLP tasks.

Accurate POS tagging is an important aspect of NLP pipelines, directly affecting the performance of downstream applications such as dependency parsing (Zhou et al., 2020), machine translation (Hlaing et al., 2022), and information extraction (Chiche and Yitagesu, 2022). Yet for languages like MC, both the lack of training data and the linguistic divergence from typologically dominant languages like French present ongoing obstacles. In previous work, Mompelat et al. (2022) demonstrated that cross-lingual transfer from French improved syntactic parsing performance for MC, but it also introduced cascading errors in the POS tagging stage, therefore highlighting the need to treat POS tagging as a distinct problem.

Rather than focusing solely on accuracy metrics, this paper takes a linguistically informed approach to model evaluation. We analyze tagging errors across the three models, with particular attention to phenomena such as lexical and syntactic ambiguity, as well as out-of-vocabulary (OOV) words. By examining classification reports, confusion matrices, support-F1 dynamics, and error patterns linked to specific linguistic features, we reveal model-specific strengths and weaknesses that would be obscured by aggregate scores alone.

This analysis not only benchmarks current POS tagging performance for MC but also informs future work on multi-model strategies, such as weighting predictions across models or implementing multi-task learning. It also guides annotation efforts, helping to determine which linguistic phenomena and patterns most merit attention in resource-constrained settings. Ultimately,

Proceedings of the 2025 CLASP Conference on Language models And RePresentations (LARP), pages 1–10
September 8-9, 2025 ©2025 Association for Computational Linguistics

this study sets a precedent for linguistically motivated model evaluation and resource development for underrepresented languages. As part of a larger project aimed at improving parsing tools for MC and other Creole languages, it illustrates how targeted, linguistically grounded interventions, beginning with POS tagging, can incrementally strengthen performance and serve as a foundation for future applications, including rule-based data augmentation, transfer learning, and hybrid symbolic-neural modeling.

2 Review of the Literature

2.1 Martinican Creole and Its Relation to French

Martinican Creole (MC) is a French-lexified Atlantic Creole spoken predominantly in Martinique. Despite heavy lexical overlap with Modern French, MC exhibits several typological divergences: it is largely isolating, with minimal inflectional morphology for tense, aspect, or agreement (see tense marker *ké* and mood marker *ka* in example (1)), and it uses post-posed determiners and possessives (see definite marker *an* in example (1)).

French and Creole have historically been described to coexist in a diglossic relationship, with French historically associated with institutional and high-prestige functions, and Creole with informal and oral domains. However, this diglossic split is starting to break down and we see both languages sharing space in virtually all functions of society (Prudent, 1981; Bernabé, 1983; Managan, 2016). This constant language contact situation in all functions results in the two languages to often be mixed within the same discourse. This functional break down between MC and French leads to prevalent code-mixing and code-switching, necessitating models that are robust to lexical variation. This close genealogical and historical link justifies cross-lingual transfer approaches, yet the subtle orthographic and phonological differences between French and MC, together with MC's unique syntactic patterns, pose challenges for direct transfer of standard POS tagsets.

2.2 Polyfunctionality in MC versus Homonymy

In MC, many closed-class items exhibit *polyfunctionality*: the capacity to serve multiple grammatical categories without any overt morphological change. We follow the terminology of Wang et al. (2021) whereby words that have more than one part of speech are called *polyfunctional words*, while words with only one part of speech are called *monofunctional words*. Polyfunctionality differs from polysemy, since all senses of a polysemous word may belong to the same POS category.

In our MC corpus, several high-frequency items display PF > 1, driving much of the POS-tagging ambiguity. Below are illustrative examples for the marker *ki* (PF = 3):

(1) KI = PRON, SCONJ, DET

a. Sé pa mwen **ki** pou réparé
COP not me who must fix
lektrisité [...]
electricity
'I'm not the one **who** must fix the electricity [...]' PRON

b. chak moun-an ja ka di **ki**
each person-DET already PROG say that
yo pa dakò.
3PL not agree
'Each of them are already saying **that** they disagree' SCONJ

c. Jik **ki** tan nou ké asepté yo
until which time we FUT accept 3PL
fè nou wont kon sa ?
make 2PL shame like that
'Until when will we let them embarass us like that?' DET

Similarly, *kon* alternates among coordinating conjunction (CCONJ), subordinating conjunction (SCONJ), and adposition (ADP) as shown in example (2).

(2) KON = CCONJ, SCONJ, ADP

a. Rad-maré anni balié lakot Atlantik
tidal-wave only sweep coast atlantic
kon lakot karayib
as_well_as coast caribbean
'The tidal wave only swept the Atlantic Coast **as well as** the Caribbean Coast' CCONJ

b. Kon di Kolo, 'si ou pa ri yo,
like say Kolo, 'if 2SG not laugh 3PL,
yo ké ri'w'.
3PL FUT laugh'2PL
'**Like** Kolo says, "If you don't laugh at them, they'll laugh at you' SCONJ

c. Jik ki tan nou ké asepté yo
until which time we FUT accept 3PL
fè nou wont **kon** sa ?
make 2PL shame like that
'Until when will we let them embarass us like that?' ADP

By contrast, true *homonymy* involves two or more unrelated lexical entries sharing form but with distinct, non-overlapping etymologies and meanings (e.g. MC *sé*$_1$ "it is" from French "c'est" acting as a copula vs. *sé*$_2$ from French "ces" acting as a plural marker).

(3) Kataloy, **sé** réjion ki pi rich adan tout
Catalonia, COP region that most rich in all
sé réjion l'Espay la
PLUR region Spain DET
'Catalonia, **it's** the richest region in all **the** region**s** of Spain' COP/DET

In example (3), we see that the homonym *sé* serves as a determiner and a copular predicate marker.

The frequent polyfunctionality of closed-class words is typical of isolating Creoles and underscores the importance of a UD-compliant annotation that preserves each usage and motivates our linguistically informed error analysis.

2.3 Neural Approaches to POS Tagging and Its Role in NLP

Part-of-speech (POS) tagging is an valubale task in natural language processing, as it provides essential grammatical structure that enables models to make more informed linguistic predictions. For example, Hlaing et al. (2022) showed that POS tags can be leveraged as syntactic signals to improve neural machine translation in low-resource language pairs.

While POS tagging systems for high-resource languages now achieve near-human accuracy, their development and evaluation are relatively straightforward due to the abundance of annotated corpora. In contrast, the task assumes greater importance in low-resource settings, where POS tags may be the only structured representation available. They can serve as scaffolding for downstream tasks such as dependency parsing (Mompelat et al., 2022) or Machine Translation, and as a way to stabilize training in scenarios where full syntactic or semantic annotations are lacking.

However, POS tagging for low-resource languages faces challenges on multiple fronts. First, the scarcity of labeled data makes it difficult for supervised models to learn robust tag distributions or to use unsupervised training methods. Multilingual languages models such as mBERT and XLM-R have become widely adopted as they excel in transferring knowledge from high-resource languages (like English) to low-resource ones, even without parallel data (Pires et al., 2019). However, their performance is uneven across languages, particularly when faced with typological distance, orthographic variability, or underrepresentation in pretraining data. Many low-resource languages, especially Creoles, diverge typologically from the languages on which these multilingual models have been trained as they often exhibit minimal inflection, fluid category boundaries, and weak morphological cues that may be underrepresented in current multilingual models (Hedderich et al., 2020).

These challenges warrant the need for typological and linguistically-aware modeling choices, whether neural, symbolic, or hybrid, using insights from more detailed parsing error analysis.

2.4 Approaching POS Tagging for Creole Languages

Despite advances in low-resource NLP, Creole languages remain severely underrepresented in the development of computational resources and tools. Only a handful of projects have produced annotated corpora for Creoles, and among these, part-of-speech (POS) tagging has received limited focused attention. The recent CreoleVal benchmarks (Lent et al., 2024) introduced a multilingual POS tagging dataset for Haitian, Mauritian, and 26 other Creoles, alongside a transformer-based model trained on these data. We note that MC is not included in the dataset from the CreoleVal project. Therefore, although this represents a significant step forward for Creole languages, the datasets remain modest in size, often domain-restricted, unevenly distributed across tasks, and the model necessitates fine-tuning for MC.

Prior work by Mompelat et al. (2022) proposed a dependency parser for MC that leveraged French as a support language via cross-lingual transfer. Although this approach improved parsing performance, it also revealed notable shortcomings in the POS tagging layer. The overall accuracy scores for dependency parsing relying on TAG embeddings showed evidence that differences in morphosyntactic structure between French and MC may be the source of the tagging inconsistencies. These results underscore the need to treat POS tagging in Creole languages as a task in its own right, rather than a secondary artifact of parsing models trained on other languages.

However, relying solely on coarse metrics such as overall accuracy or macro-F1 scores, especially in structurally complex or data-scarce settings has its limitations. In this paper, we propose a more comprehensive evaluation framework aimed at providing a deeper understanding of model behavior, uncovering patterns of misclassification that would otherwise be obscured by aggregate performance metrics. As Schöffel et al. (2025) show in their study of Old Occitan, low-frequency categories such as interjections, proper nouns, or borrowed terms are especially prone to misclassification by neural taggers, even when macro metrics remain high. In doing so, it also helps identify gaps in the annotated data and motivates more targeted annotation strategies.

2.5 Error Analysis in POS Tagging

Error analysis is a critical tool for understanding the systematic failures of NLP models, especially in contexts where structural ambiguity and low-resource constraints compound the difficulty of robust language modeling. Studies such as Garcia and Gamallo (2010) have demonstrated that error-driven rule-based correction can significantly improve tagging accuracy, particularly when errors are concentrated in predictable linguistic contexts such as confusion between adjectives and nouns, or misinterpretation of closed-class items like determiners, adpositions, and subordinating conjunctions. This suggests that error analysis not only diagnoses model weaknesses but can actively guide remediation strategies through symbolic or hybrid interventions.

In our study, we build on these precedents by introducing a comprehensive statistical error analysis of POS tagging in MC. This includes confusion matrix interpretation, per-tag F1 tracking, identification of homonymous/polysemous tokens, Out-Of-Vocabulary (OOV) error reports, and a support-F1 LOESS analysis to capture the interaction between tag frequency and performance. Our proposed metrics are motivated by MC's high levels of lexical ambiguity/polyfunctionality, frequent borrowing, and minimal morphological marking. To date, no comprehensive error analysis has been conducted for POS tagging in MC or other French-lexified Creoles.

Importantly, our goal is not only to better understand model behavior, but to use these findings to inform the design of targeted data augmentation strategies via large language models. In contrast to Schöffel et al. (2025) who use LLMs directly for tagging evaluation, we intend to leverage them as generative and augmentation tools, guided by linguistic insights extracted through error analysis.

3 Methodology

Our approach is organized into two main phases: first, the creation of a POS tagging benchmark for Martinican Creole (MC) using transformer-based models, and second, the creation of an evaluation framework for fine-grained error analysis to inform linguistically and symbolically future annotation efforts, automatic data augmentation methods and new, hybrid strategies for the development of NLP tools for MC and other low resource languages.

3.1 Dataset and Annotation Process

The dataset used in this study combines the depedency parsing corpus introduced in Mompelat et al. (2022), containing 236 manually dependency-annotated sentences in MC and from which we only extracted the POS annotations, with 298 additional sentences annotated for POS for this project. Although ideally multiple native-speaker linguists would adjudicate, all 298 new sentences were annotated by the author, a heritage speaker of MC with formal training in Universal Dependencies schemas, due to a severe lack of Martinican UD experts, typical for underrepresented languages. To ensure quality, we conducted two full consistency passes over the data and spot-checked ambiguous tokens against a small panel of native speakers. The full dataset thus comprises 534 sentences, 9470 tokens, and 1780 types, making it the most extensive POS-annotated corpus for MC. The data are drawn from online news sources, blogs, and social media, reflecting contemporary usage and the frequent presence of code-mixed French elements. These mixed tokens are essential for capturing the diglossic and bilingual nature of Martinican linguistic practice.

Annotation followed the Universal Dependencies (UD) guidelines for POS tagging (De Marneffe et al., 2021). A key annotation tag concerns the treatment of foreign words, particularly French lexical items. When a French-origin word is syntactically integrated into the MC sentence, functioning as a noun, verb, or modifier, it was tagged according to its grammatical role using standard UPOS categories. However, when the foreign word appeared as a translation equivalent or gloss, not syntactically integrated into the clause structure, it was assigned the X tag, in line with UD conventions for unclassifiable or extragrammatical tokens. Although the "X" tag is often excluded or minimized in evaluation tasks, we chose to preserve it as a focus in our error analysis since this decision enables us to track annotation ambiguity and model behavior around edge cases, rather than suppressing them from the training and evaluation process.

After initial submission of this paper, we discovered annotation inconsistencies in the MC POS corpus. We have corrected these and fixed a random seed (seed=42) for our train/dev/test splits. All results below reflect this finalized dataset.

3.2 Model Selection and Experimental Design

To evaluate POS tagging performance on MC, we selected three transformer-based models: 1) XLM-Roberta (xlm-roberta-base), a multilingual transformer pretrained on CommonCrawl data containing 100 languages, 2) mBERT (bert-base-multilingual-cased), a widely used multilingual model pretrained on the BooksCorpus and Wikipedia, and 3) CreoleVal, a domain-adapted XLM-R transformer model fine-tuned on the CreoleVal benchmark (Haitian, Mauritian, and 26 other Creoles; Lent et al. 2024). Its subword vocabulary and pre-training did *not* include any MC data, making it directly comparable to XLM-R and mBERT for MC POS tagging. Although CreoleVal and XLM-R share the same tokenizer and subword vocabulary, their parameter distributions diverge during CreoleVal's in-domain fine-tuning. By adapting XLM-R weights on a multilingual Creole benchmark (Haitian, Mauritian, etc.), CreoleVal becomes highly specialized on core Creole patterns, boosting closed-class and function-word tagging, but can "forget" some of XLM-R's broader multilingual robustness. In practice, this leads XLM-R to outperform CreoleVal on proper nouns and rare connectors: its untouched pre-training retains more general representations for named entities and low-frequency items, whereas CreoleVal's weights have shifted toward the distributions encountered in its fine-tuning data.

Each model was fine-tuned on our annotated dataset using an 80/10/10 split for training (427 sentences), development (53 sentences), and testing (54 sentences). The splits also broadly preserve the overall distribution of common POS tags.

3.3 Evaluation and Error Analysis

To assess the behavior and weaknesses of our POS tagging models in a linguistically informed manner, we adopt a suite of diagnostic metrics tailored to the structural characteristics of MC.

We begin with the standard classification report shown in Table 1, which provides precision, recall, and F1 scores per tag. We further examine the relationship between data availability and tagging performance through a support-F1 analysis shown in Figure 1. By plotting F1 scores against tag frequency using LOESS smoothing, we can estimate the number of examples required for each tag to achieve reliable performance/information balance that directly informs future annotation priorities. Then, to uncover more nuanced patterns of misclassification, we analyze confusion matrices, shown in Figure 2, that reveal frequent tag-level confusions such as NOUN versus PROPN or ADP versus SCONJ. These confusion patterns are particularly relevant in MC, where the absence of morphological cues often makes syntactic functions harder to disambiguate and where homonymy/polysemy is frequent. To evaluate the model's ability to resolve morphosyntactic ambiguity, we conduct a homonymy/polysemy error analysis by tracking tokens that occur with multiple POS tags in the corpus. Finally, we perform an OOV error analysis, examining how models handle test-set tokens not seen during training. This includes measuring the overall OOV error rate and identifying common misclassification patterns, such as overpredicting NOUN or confusing named entities and borrowed forms.

4 Results & Discussion

4.1 Overall Model Comparison: Tagging Accuracy by Category

We begin our evaluation with overall accuracy: fine-tuned mBERT achieves 92% , XLM-Roberta 91%, and CreoleVal tops at 94%. Table 1 presents per-tag F1 scores on the Martinican Creole (MC) test set.

All three models achieve perfect or near-perfect scores on the most frequent, low-ambiguity classes: PUNCT, PART, and PRON. They also perform strongly on PROPN. By contrast, rare tags such as INTJ (support=2) remain challenging.

Mid– and low–frequency tags reveal clear differentiation. CreoleVal leads on most functional and content categories, while mBERT retains the edge on coordinators and interjections. In sum, CreoleVal delivers the highest overall F1 by excelling on both function- and content-POS tags; mBERT and XLM-Roberta bring complementary strengths on rare or subtle categories, motivating the model-fusion strategies described in Section 4.6.

Table 1: Per-tag F1 scores across models. Highest values per row are in **bold**.

POS Tag	mBERT	XLM-R	CreoleVal
ADJ	0.77	0.77	**0.82**
ADP	0.81	0.74	**0.84**
ADV	0.81	0.82	**0.86**
AUX	0.95	0.95	**1.00**
CCONJ	**0.90**	0.83	0.86
DET	0.90	0.91	**0.95**
INTJ	**0.67**	0.50	0.50
NOUN	0.88	0.89	**0.91**
NUM	0.86	**0.92**	0.89
PART	0.98	0.99	0.99
PRON	0.99	0.97	**1.00**
PROPN	0.93	0.92	0.93
PUNCT	1.00	1.00	1.00
SCONJ	0.74	0.67	**0.77**
VERB	0.94	0.95	**0.96**
X	0.94	0.94	0.94
TOTAL	0.92	0.91	**0.94**

4.2 LOESS F1 vs. Support Analysis

The analysis presented here utilizes Locally Estimated Scatterplot Smoothing (LOESS), a non-parametric regression method, to explore how the number of annotated examples per POS tag (support) relates to model performance (F1-score). LOESS smoothing is particularly beneficial when dealing with limited, unevenly distributed data points.

To derive a practical annotation threshold for each model, we first compute a smoothed F1-vs-support curve via LOESS. Rather than simply picking the support level at which the smoothed curve first crosses our target F1 of 0.90 (which can be susceptible to isolated "spikes" or noise), we identify all contiguous runs of support values where the LOESS-smoothed F1 remains at or above 0.90. We then select the longest such run, namely the largest consecutive region of stable, high performance, then take its minimum support value as our recommended threshold. This ensures the threshold reflects a region where performance truly "stabilizes" above the target, not just a brief local bump.

Although this method only provides an approximate guide (it does not guarantee per-tag minimums in every context), it offers a more robust estimate of how many annotated tokens are needed, on average, before a model can be expected to perform reliably above an F1 of 0.90. These thresholds can then inform future corpus development and annotation planning.

From the LOESS smoothed curves, shown in Figure 1, we observe that models differ significantly in their data requirements for achieving a high-performance threshold. Specifically, the analysis reveals that the CreoleVal model reaches this threshold with fewer annotated instances (support threshold = 54) compared to XLM-R (support threshold = 121) and mBERT (support threshold = 78). Overall, this analysis clarifies the data-driven relationship between annotation volume and model performance, directly informing practical decisions on corpus annotation strategies and model deployment.

4.3 Error Profile per Model : Tag Confusions

In this section, we analyze the confusion matrices and polyfunctionality-driven errors to understand where models tend to confuse tags, and which types of tokens are consistently hard to disambiguate.

The confusion matrices shown in Figure 2 highlight two interrelated sources of error that stem from the structural characteristics of MC (and many other Creoles). First, content words in MC are highly polyfunctional: A single word can freely serve as a NOUN, ADJ or VERB, which leads our models to routinely confuse our ADJ/VERB/NOUN labels. Second, the lack of overt morphological marking on functional tokens makes it difficult to distinguish conjunctions (SCONJ/CCONJ) from prepositional markers (ADP), resulting in systematic SCONJ ↔ ADP ↔ CCONJ errors.

Figure 2 shows that when we aggregate off-diagonal confusions among ADJ, NOUN, and VERB, CreoleVal commits only 12 such errors, compared to 16 by mBERT and 16 by XLM-Roberta. The fact that CreoleVal produces the fewest misclassifications here highlights its superior handling of MC's polyfunctional content items as a Creole-specific language model. Additionally, for the triad ADP–CCONJ–SCONJ, mBERT makes the fewest confusions (11 total), with CreoleVal next (12) and XLM-Roberta last (16). This suggests that mBERT's broader multilingual pre-training better preserves subtle distinctions among functional markers, whereas CreoleVal's in-domain adaptation trades away a small amount of this fine-grained discrimination.

Together, these patterns reveal complemen-

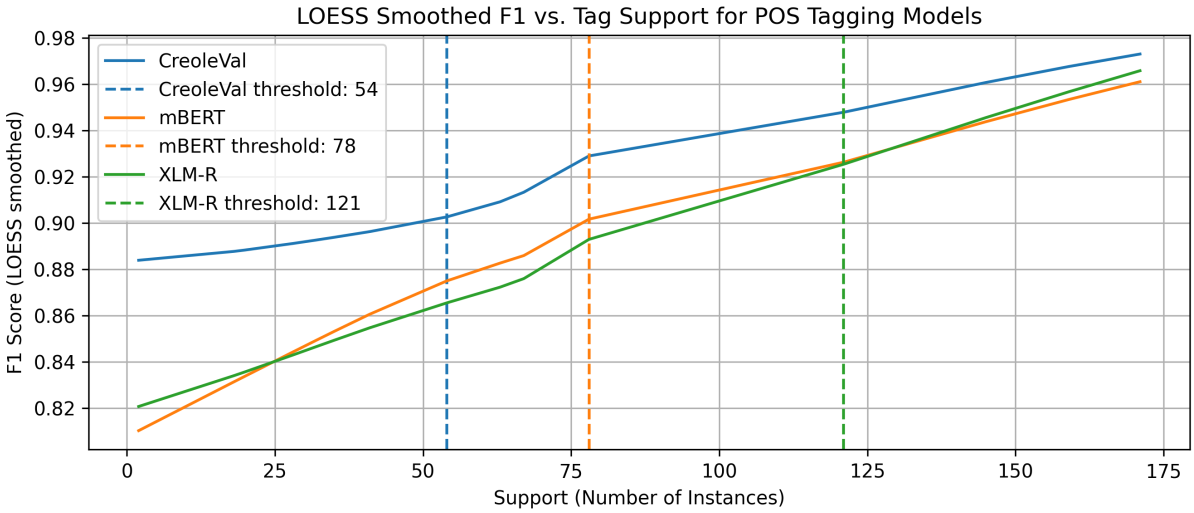

Figure 1: F1-score/support LOESS-smoothed curves for CreoleVal, mBERT, and XLM-Roberta models.

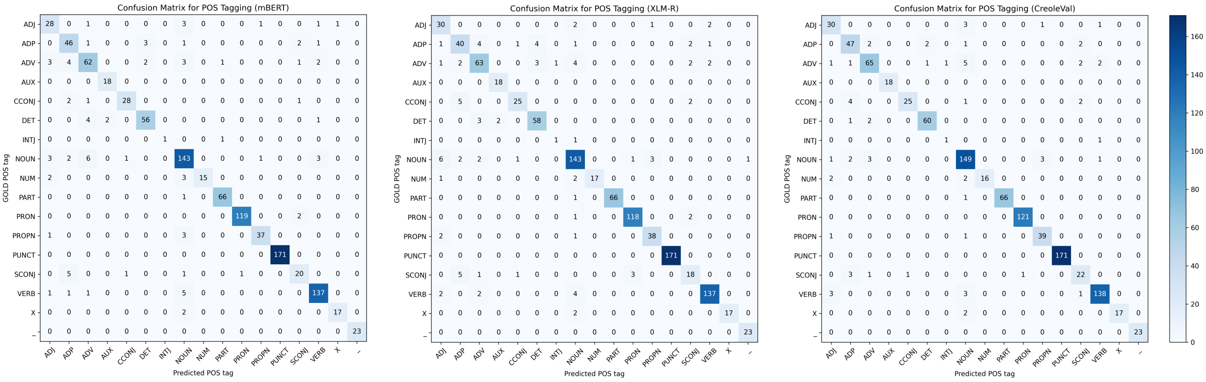

Figure 2: Confusion Matrices for mBERT, XLM-Roberta, and CreoleVal on the Martinican Creole test set.

tary strengths: CreoleVal excels at polyfunctional content-word tagging, mBERT at conjunction/preposition boundaries, and XLM-Roberta lags on both fronts, motivating our fusion strategies to combine their niche advantages in Section 4.6.

4.4 Error Profile per Model : Lexical Ambiguity

Several polyfunctional tokens stand out as persistent error sources, illustrating how MC's category-shifting forms challenge neural taggers. In particular:

- **ki** (30 occurrences): error rates are 3.3% for CreoleVal (1/30), 16.7% for XLM-R (5/30), and 10% for mBERT (3/30). Confusions swap primarily between PRON and SCONJ, with occasional DET misassignments.

- **kon** (6 occurrences): error rates climb to 50% for CreoleVal (3/6), 33.3% for mBERT (2/6), and 66.7% for XLM-R (4/6), reflecting its trifunctional usage as CCONJ, SCONJ, and ADP.

- **pou** (21 occurrences): errors occur at 19% for both CreoleVal and XLM-R (4/21), and 23.8% for mBERT (5/21), with frequent flips between ADP and SCONJ.

Comparing these three confirms a strong and expected correlation between frequency and error rate : the more examples a form has, the more reliably it is tagged. Overall, CreoleVal makes the fewest errors on polyfunctional items, except for *kon*, where its mid-rank performance suggests that exceptionally low support still hampers even the Creole-adapted model.

The homophonous word **sé** ("c'est" vs. "ces", 25 occurrences) exemplifies true homophony rather than just polyfunctionality. Here CreoleVal fully resolves the distinction, while mBERT and XLM-R each mislabels the word on 8% of cases (2/25) as

AUX instead of DET. This difference suggests that in-domain Creole fine-tuning helps models learn language-specific lexical disambiguations that multilingual pretraining alone may miss.

We see that it is these high-frequency polyfunctional and homophonous tokens (and not rare, monofunctional forms) that drive most systematic tagging failures. Remedying them will require richer syntactic context (e.g. dependency relations) or targeted annotation of ambiguous constructions to guide models toward the correct POS distinctions.

4.5 Generalization Limits: OOV Word Behavior

Generalization to unseen vocabulary remains a critical challenge in MC tagging. In our low-resource setting, OOV tokens account for 144 tokens (1.5% of the test set), spanning proper names, phonological variants, idiomatic compounds, and loanwords. CreoleVal mislabels 19 / 144 OOV tokens (13.2%), compared to 23 / 144 (16.0%) for XLM-Roberta and 29 / 144 (20.1%) for mBERT.

A closer look at the misclassified OOV items reveals three dominant error patterns. First, verbs that are morphologically or phonologically similar to the French variants, like **lavé** (*laver*), **payé** (*payer*), **prononsé** (*prononcer*) are persistently mis-tagged as adjectives or nouns. Second, adjectives like **diféran** (*different*) or adverbs like **asé** (*enough*) were frequently flipped between ADJ, ADV, or NOUN. Third, proper-name errors are rare: mBERT alone errs on *Martinique* (PROPN→NOUN), and XLM-R on *guardia* (NOUN→PROPN), underscoring CreoleVal's in-domain fine-tuning advantage for named entities. Finally, foreign words (*galaxie*, *gravitation*) were mis-tagged as NOUN instead of X(2 errors each in CreoleVal and XLM-R, 2 in mBERT), indicating that fully integrated loanwords might confuse the "unclassifiable" category "X" as described in the UD guidelines. Therefore, none of the taggers seem to easily discriminate between fully-integrated borrowed words and foreign insertions. The ability to perform such discrimination is of great importance when dealing with context of enhanced code-switching and code-mixing.

While additional pre-training data can help reduce low-coverage gaps, such large-scale pretraining is often impossible for truly low-resource communities. Our focus here is on fine-tuning existing multilingual models, which offers a more accessible path to improved accuracy. Effective strategies may include explicit linguistically-motivated lexicon augmentation for proper names, enhanced subword tokenization for morphologically complex variants, and hybrid approaches that combine neural tagging with lookup tables for fixed expressions. Only by addressing these structural gaps can we push beyond the generalization limits of our current fine-tuned systems.

4.6 Towards Model Fusion: Leveraging Complementary Strengths

Our analyses confirm that each model brings distinct advantages: CreoleVal delivers the strongest performance on many functional and closed-class tags, leading on DET (0.95 vs. 0.90 vs. 0.91), PART (0.99 vs. 0.98 vs. 0.99), AUX (1.00 vs. 0.95 vs. 0.95), and ADV (0.86 vs. 0.81 vs. 0.82). It also yields the lowest OOV error rate (13.2% vs. 16.0% vs. 20.1%).

XLM-Roberta excels on numeric tags (NUM = 0.92 vs. 0.89 vs. 0.86), reflecting its potential for robust subword representations for rare morphosyntactic constructions. CreoleVal and mBERT both top the named-entity tag (PROPN = 0.93 vs. 0.93 vs. 0.92), while CreoleVal and XLM-R share the highest X performance (0.94 vs. 0.94 vs. 0.92), demonstrating superior handling of fully foreign insertions. mBERT remains the most balanced generalist, with particularly strong scores on CCONJ (0.90), and PRON (0.99).

This clear complementarity suggests several fusion strategies. First, per-token weighted voting, where each model's tag-specific validation F1 determines its vote weight, could improve accuracy on challenging categories like SCONJ (F1 0.67–0.77). Second, per-tag delegation, assigning each POS to its top specialist, would directly leverage CreoleVal's mastery of function words, XLM-Roberta's numeric proficiency, and mBERT's connector expertise. Third, a multi-task architecture combining all three contextual embeddings into a unified classifier may learn to trust each representation dynamically.

Beyond accuracy gains, this fusion approach also informs annotation priorities: SCONJ tags and highly polyfunctional items (e.g. *ki*, *kon*) still incur error rates up to 66.7% under XLM-R. Targeted annotation or data augmentation for these high-ambiguity forms, and enriching their syntactic contexts with dependency relations, may amplify the benefits of any ensemble, ensuring future MC taggers combine the best of each model's strengths.

4.7 Linguistic Insights for Targeted Annotation and Data Augmentation

This study offers key linguistic takeaways that can directly inform future annotation priorities and data augmentation strategies for MC and other contact-influenced low-resource languages. The errors made by even the strongest taggers are not arbitrary; they reveal systematic patterns shaped by the typological characteristics of Creole morphosyntax.

First, the fluidity of Creole categories where the same form can function as verb, noun, adjective, or connective without overt inflection emerges as the root cause of many systematic confusions. Rather than dispersing annotation effort evenly, we should concentrate on sentences that illustrate this fluidity. In practice, this means mining the corpus (or synthetic data) for contexts where high-ambiguity items like *ki*, *kon*, and *pou* appear in each of their roles, then creating compact annotation batches that cover all readings of a single token. By focusing scarce human effort on these multifunctional "edge cases" we ensure the model sees the precise contextual cues needed to resolve category overlap, rather than redundantly tagging unambiguous examples.

Second, our LOESS-informed support thresholds identify which POS tags remain under-supported even in the best model. In total, 10 of the 16 UPOS categories fall below the 54-instance mark required for stable F1 under CreoleVal: specifically ADJ (35), AUX (18), CCONJ (32), INTJ (2), NUM (20), PROPN (41), SCONJ (28), X (19), alongside marginal cases of ADP (54) and ADV (78). A targeted annotation drive that brings each of these tags up to at least 54 examples would allow significant gains rather than spreading effort at random across already well-learned categories.

Third, the remaining OOV errors on phonological variants, loanwords, and code-switched items spotlight the need for lexicon-aware augmentation. Instead of relying on larger pre-training corpora, which may be infeasible for MC, we can inject synthetic examples of rare compounds (e.g. *alé-vini*), phonetic spellings (*lwen*, *vré*), and integrated borrowings (*vulgaire*) into the training data. Using few-shot LLM prompting, guided by our linguistic error profile, we can generate minimal pairs that contrast these forms in their correct contexts, helping the tagger anchor them to the right POS class. This is to be tested in a near-future experiment.

Finally, these insights advocate for an active-learning loop in which model disagreement and low-confidence predictions drive both annotation selection and augmentation design. By letting the taggers themselves flag the most contentious tokens, we turn our error analyses into a continuous feedback mechanism. Over successive cycles, this linguistically informed, resource-efficient strategy may allow us to deliver MC taggers that not only achieve higher accuracy but also demonstrate a deeper understanding of the language's typological complexity.

5 Conclusion

In this work we have presented the first fine-tuned transformer models for part-of-speech tagging of Martinican Creole (MC) along with a linguistically grounded evaluation framework. Building on a UD-style POS corpus of 534 sentences, we compared mBERT, XLM-Roberta, and CreoleVal. Our results yield accuracy scores of 0.92 for mBERT, 0.91 for XLM-Roberta, and 0.94 for CreoleVal. Beyond scores, our analyses uncover the linguistic dimensions of tagging difficulty in MC: polyfunctional word classes, blurred syntactic boundaries, and code-switching all contribute to high tag ambiguity.

The complementary model strengths we observe suggest lightweight ensemble strategies: for example, per-tag delegation can assign each POS to its specialist model; weighted voting can boost performance on the most ambiguous categories; and a multi-task fusion architecture can learn to trust each model's representations dynamically.

Looking ahead, we advocate a targeted annotation and augmentation pipeline that focuses scarce human effort where it matters most. By combining linguistically informed strategies with model fusion strategies, we anticipate substantial gains in POS-tagging robustness for MC, and we hope this work serves as a blueprint for other under-resourced, contact-influenced languages.

Limitations

This study is constrained by the relatively small size of the annotated MC corpus, which limits model generalization and makes evaluation sensitive to lexical overlap. Additionally, while we provide detailed post hoc analysis, our tagging architectures remain model-centric.

Future work will address these gaps by 1) Developing a reward-modeling framework to guide annotation and tagging across ambiguous categories; 2) Exploring ensemble and voting-based approaches informed by per-tag performance; 3) Designing linguistically controlled data augmentation pipelines, including LLM-generated MC sentences; and 4) integrating POS tagging with downstream tasks such as dependency parsing and translation, forming part of a unified Creole NLP pipeline.

Together, these directions aim to move Creole NLP beyond passive transfer and toward linguistically-aware, low-resource-first modeling strategies.

References

Jean Bernabé. 1983. *Fondal-natal: grammaire basilectale approchée des créoles guadeloupéen et martiniquais: approche sociolittéraire, sociolinguistique et syntaxique*. L'Harmattan.

Alebachew Chiche and Betselot Yitagesu. 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1):10.

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

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

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.

Marcos Garcia and Pablo Gamallo. 2010. Using morphosyntactic post-processing to improve pos-tagging accuracy. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR'2010). Porto Alegre, RS*.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

Zar Zar Hlaing, Ye Kyaw Thu, Thepchai Supnithi, and Ponrudee Netisopakul. 2022. Improving neural machine translation with pos-tag features for low-resource language pairs. *Heliyon*, 8(8).

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, and 1 others. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, and 1 others. 2024. Creoleval: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

Kathe Managan. 2016. The sociolinguistic situation in guadeloupe: Diglossia reconsidered. *Journal of Pidgin and Creole Languages*, 31(2):253–287.

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

Salikoko S Mufwene. 2013. Simplicity and complexity in creoles and pidgins: What's the metric? *Journal of Language Contact*, 6(1):161–179.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Lambert-Félix Prudent. 1981. Diglossie et interlecte. *Langages*, (61):13–38.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Matthias Schöffel, Marinus Wiedner, Esteban Garces Arias, Paula Ruppert, Christian Heumann, and Matthias Aßenmacher. 2025. Modern models, medieval texts: A pos tagging study of old occitan. *arXiv preprint arXiv:2503.07827*.

Pascal Vaillant. 2023. Noun phrases in mixed martinican creole and french: Evidence for an underspecified language model. *Journal of Pidgin and Creole Languages*, 38(2):207–262.

Lu Wang, Yahui Guo, and Chengcheng Ren. 2021. A quantitative study on english polyfunctional words. *Glottometrics*, 50.

Houquan Zhou, Yu Zhang, Zhenghua Li, and Min Zhang. 2020. Is pos tagging necessary or even helpful for neural dependency parsing? In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 179–191. Springer.