

Gwo-K Kreyol Classifier: Rock Pidgins (RP) Submission to the WMT 2026 CreoleMT Language Identification Shared Task

Emmett Strickland

National Institute for Oriental Languages and Civilizations (Inalco)
Research Group on Texts, Informatics, and Multilingualism (ERTIM)

Paris, France

emmett.strickland@inalco.fr

Abstract

This paper describes ongoing efforts to build an n-gram text classifier for several major French-lexifier creoles. Our model is based on a simple text processing pipeline, limited training data, and a deliberately lightweight architecture. Despite this, it achieves substantially stronger performance than existing baselines on our target languages. These results suggest that long-established machine learning approaches remain effective for new language identification needs.

1 Introduction

Language identification (LID) is a classic text processing task that involves automatically determining the language used in a written string. Compared to machine translation or speech synthesis, categorizing a text into a limited set of candidate languages is relatively straightforward. However, the reliability of LID systems can be critical to the proper execution of these more sophisticated downstream tasks. For example, automatic translation tools fail when the language of a source text is incorrectly identified. Meanwhile, multilingual and code-switching text-to-speech (TTS) systems require modeling how text maps to speech in each language, and correctly applying these models to the relevant segments. In linguistically diverse regions, LID can streamline the task of categorizing and preprocessing multilingual texts on websites and social media. Existing LID tools achieve excellent performance on the 100 or so best-resourced languages, but support is limited for the marginal majority without large text corpora.

Among these languages, creoles are broadly underserved by language classifiers when considered as a category.¹ As comparatively recent products

of the early modern spice and slave trades, creoles lack the long written traditions of their mostly European lexifiers, and have historically been overlooked as nonstandard varieties of *them* rather than as distinct languages in and of themselves. Like many other low-resource languages, they often lack an official administrative status,² while their speakers mainly reside in historically agricultural economies with delayed access to digital resources.

In spite of these challenges, or perhaps because of them, creoles present a genuinely interesting use case for LID tools. Many creoles today coexist with their lexifiers in a complex state of diglossia, with intermediate syntactic constructions sometimes blurring the line between creole and lexifier grammars.³ Formal and informal spelling conventions can also complicate matters. Some creoles, like West African Pidgin English, spell words as they are written in their language of origin. Others have adopted orthographies distinct enough to obscure otherwise straightforward etymologies, as in the English-based Tok Pisin, or ‘Talk Pidgin’. Some have no codified orthography whatsoever, leading to multiple competing standards in practice.⁴

The preference for creoles in informal contexts complicates matters further. While French’s diagraphs and silent letters easily distinguish it from the phonemic spellings of its creole offshoots, this distinctiveness evaporates among the abbreviated stylizations of informal writing. A simple *qu’est-ce que tu fais* ‘what are you doing’ can be rendered as a superficially far more creole-like *kesk tu fai* in

²Haitian Creole and Papua New Guinea’s Tok Pisin being two major exceptions.

³See the concept of creole continuum popularized by Bickerton (1973).

⁴In some cases, competing norms can reflect both linguistic and identitarian concerns. See the tension between a pragmatic English-style orthography and a distinctly ‘nativized’ phonemic system proposed for Nijár, or Nigerian Pidgin (Ofulue, 2017; Ekundayo and Enwerem, 2021).

¹Whether creoles represent a coherent typological class or a loose sociohistoric construct grouping disparate colonial varieties is a matter of debate. See McWhorter (2018) and Mufwene (2015) for two contrasting visions of creoleness.

the very informal contexts where a creole is most likely to be used.

In this paper, we present early steps towards addressing these issues with Gwo-K, a lightweight n-gram classifier for differentiating French-lexifier creoles. This model was produced as part of the 2026 WMT Shared Task on Creole Machine Translation. Indirect comparison with the current state-of-the-art suggests that our model outperforms existing LID tools for creoles, demonstrating the effectiveness of long-established machine learning approaches when applied to task-specific datasets supported by targeted augmentation strategies.

2 Model design

As specified in the Shared Task challenge, our language classifier was designed to categorize texts into twelve distinct language categories, nine of which are French-lexifier creoles spoken in the Caribbean and the Indian Ocean. The remaining categories are reserved for English and French texts, and a catch-all ‘Other’ category for any other language.

2.1 Data selection

As training data, we primarily used the train subset of the Kreyòl-MT dataset, which includes text-translation pairs in numerous creoles and several European lexifiers (Robinson et al., 2024). We extracted all texts in the Creole-English and Creole-French subsets of the corpus. All creole language tags not included in our list of target languages were mapped to oth (other), while examples of English and French were sourced from the translated texts. Any duplicate creole texts were excluded without exception. Texts underwent minimal preprocessing: case distinctions were eliminated, and common punctuation marks were discarded. The same set of treatments were applied to the validation set. While exploring the training data, we noticed that a large portion of Haitian texts were composed of product numbers, HTML tags, and other linguistically uninformative strings. To eliminate such texts, we discarded all entries where common HTML tags were present, or where numerals, punctuation marks, and upper-case letters comprised more than 30% of characters.

The full list of categories and dataset sizes is outlined in Table 1. Note that Guianese Creole (gcf) and Reunion Creole (rcf) were especially underrepresented in the Kreyòl-MT dataset. To

overcome this shortcoming, we augmented our training set with external texts manually gathered online. Reunion Creole was supplemented with 2163 sentences provided by Cellier (1985) and McLellan (2024), while the Guyanese data was reinforced with 1000 sentences from geography-focused Wikipedia entries. The majority of the supplemental texts were far longer and more formal than the brief and syntactically simple examples used in the Kreyòl-MT dataset.

As a self-imposed limitation, we restricted our supplemental data to sources explicitly published under a permissive license. Potentially useful texts published by religious and cultural organizations were excluded because of unclear reuse conditions. However, we did secure access to an additional 2823 examples of prose and poetry by collaborating with the Reunion Creole advocacy organization *Lofis*.⁵ These were included in the augmented training set.

Code	Language	Samples	
		Train	Test
acf	Saint Lucian	3757	222
crs	Seychellois	1963	104
gcf	Guadeloupean	4400	204
gcr	French Guianese	346 (+1k)	78
hat	Haitian	814k	1544
lou	Louisianan	1532	92
mart	Martinican	4346	281
mfe	Mauritian	26.1k	619
rcf	Reunionese	189 (+5k)	34
eng	English	576k	3800
fra	French	319k	1970
oth	Other	38.4k	2592

Table 1: Language categories and dataset sizes. Figures in parentheses represent text samples gathered from external sources.

2.2 Model training and variants

After preprocessing our train and test sets, we trained a classifier based on character-level TF-IDF representations combined with a linear support vector machine (SVM). Our system extracts character sequences of length three to five from each text, allowing it to capture orthographic patterns, grammatical monosyllables, and frequent subword structures unique to each language class. The resulting feature vectors are weighted using

⁵<https://lofislangkreollarenyon.re/kisa-ni-le/>

term frequency-inverse document frequency (TF-IDF) and passed to a LinearSVC classifier with balanced class weighting to mitigate the effects of uneven training data distribution.

For this paper, we evaluated two model variants on our test set. Gwo-K_1 was trained using only the texts available in the Kreyòl-MT corpus, while Gwo-K_2 included the additional data gathered for the data-poor Reunion and Guyanese creoles.

3 Results and analysis

3.1 Global performance

Both Gwo-K_1 and the data-augmented Gwo-K_2 models achieve accuracy around 97%, though this figure is skewed by the overrepresentation of Haitian, English, French, and Other classes. Average and language-specific precision, recall, and F1 scores for both models are presented in Table 2.

Table 2: Classification performance comparison between Gwo-K_1 and Gwo-K_2. Bold values indicate better performance between models; * indicates a notable absolute difference (≥ 0.03).

Class	Gwo-K_1			Gwo-K_2		
	P	R	F1	P	R	F1
acf	0.93	0.91	0.92	0.94	0.91	0.92
crs	0.97	0.96	0.97	0.97	0.94	0.96
gcf	0.80	0.92	0.86	0.83*	0.92	0.87
gcr	0.95	0.47	0.63	0.94	0.58*	0.71*
hat	0.94	0.99	0.97	0.95	0.99	0.97
lou	0.98	0.89	0.93	0.98	0.88	0.93
mart	0.93	0.82	0.87	0.94	0.84	0.89
mfe	0.97	0.98	0.97	0.97	0.98	0.98
rcf	0.92*	0.35	0.51	0.82	0.79*	0.81*
eng	0.97	0.99	0.98	0.97	0.99	0.98
fra	0.99	0.99	0.99	0.99	0.99	0.99
oth	0.99	0.96	0.98	0.99	0.96	0.98
Accuracy	0.9699			0.9720		
Macro Avg.	0.95	0.85	0.88	0.94	0.89*	0.91*
Weighted Avg.	0.97	0.97	0.97	0.97	0.97	0.97

Gwo-K_1 performs poorly on both French Guianese Creole (gcr) and Reunion Creole (rcf), with recall scores below 0.5 and F1 scores in the low 50s and 60s. This likely reflects the shortage of training data mentioned earlier. Languages with lower amounts of training data are broadly associated with lower F1 scores, though the relationship is not fully predictive. For example, Louisiana Creole (lou) is identified far more effectively than its Guadeloupean (gcf) and Martinican (mart) counterparts despite having less than one third as much training data. This is somewhat unsurprising given the language’s geographic and historic isolation from the Caribbean creoles, and suggests that more

linguistically distinct classes require fewer examples to build effective LID tools.

Gwo-K_2, which was trained on expanded Guianese and Reunion Creole datasets, far outperformed Gwo-K_1 on those languages. These respectively exhibit 11- and 44-point improvements in recall, and 8- and 30-point improvements in F1 measure. The difference in gains is roughly proportionate to the amount of supplemental data included in each language’s training set. Despite marked improvements over Gwo-K_1, performance on these languages lags behind the other classes. It is especially interesting that gains made on Reunion Creole were partially offset by a lower overall precision. We attribute this performance gap to the fact that we chose not to expand the evaluation sets for these languages. It is therefore likely that differences in genre between the training and test sets contributed to higher error rates. This is especially true for Reunion Creole, whose supplemental data contained an orthographic standard not used in the Kreyòl-MT corpus. Still, the high overall improvement does suggest that the additional data is largely generalizable beyond its original sources.

Overall, the augmented model enjoys a 3-point macro average improvement thanks to the superior performance on these two creoles, with significantly higher recall and slightly lower precision.

3.2 Error analysis

To better understand sources of errors, we produced confusion matrices for each model.

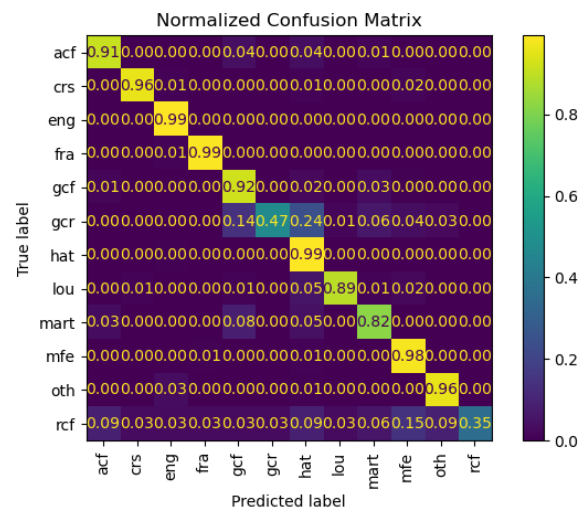


Figure 1: Confusion matrix for the Gwo-K_1 classifier on the test set. Values are normalized by true class.

The matrix produced from Gwo_K-1 in Figure 1

shows that French Guianese Creole (gcr) is almost exclusively confused with Guadeloupean, Martinican, and Haitian creoles, all of which are spoken in the Caribbean. Meanwhile, Reunion Creole is most likely to be confused with Mauritian Creole (mfe), also spoken in the Indian Ocean. To some extent, geographic and linguistic proximity appears to reflect the types of errors this model produces.

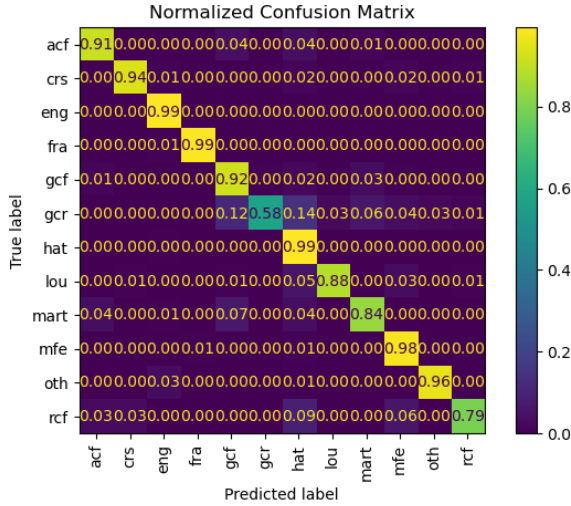


Figure 2: Confusion matrix for the Gwo-K_2 classifier on the test set. Values are normalized by true class.

The confusion matrix produced from the Gwo-K_2 model in Figure 2 reveals similar trends. Even with the marked improvements made from our data augmentation strategy, errors continue to cluster around the same groups of languages. Interestingly, the French-based creoles are almost never confused with their French lexifier, suggesting that the creoles pattern much closer to each other than they do to French. Nor are creoles often confused with ‘Other’ languages. Despite occasionally high error rates *between* creoles, both systems are exceptionally good at differentiating French-based creoles from other categories of language. This could make our system especially useful in contexts where the geographic origins of texts are already known. In such cases, creole classification tasks could be simplified into a matter of differentiating between a generic ‘French creole’ label and a limited set of other likely candidates.

3.3 Comparison with existing baselines

While our models yield excellent results for most creoles and promising results for the rest, it is important to compare our scores to existing baselines. However, this is difficult when it comes to creole

languages. FastText, a popular language classifier, only offers coverage to Haitian Creole (Joulin et al., 2017).

However, GlotLID offers coverage of all languages discussed in this paper, with the exception of Martinican Creole which it does not distinguish from Guadeloupean (Kargaran et al., 2023). This model was tested by Shared Task organizers on a separate evaluation set similar in content and structure to our own. This model yielded a macro F1 score of 0.68, a weighted F1 score of 0.74, and an overall accuracy of 74.87%. While our scores far outperform those of GlotLID, we must also acknowledge that their results are artificially deflated by the lack of a dedicated *mart* label.

4 Perspectives and conclusion

The model presented in this paper is a work in progress, and is likely to undergo substantial revisions in the coming months. The primary contributions of this work are to highlight the usefulness of basic n-gram language classification strategies in data-rich scenarios, and simple data-augmentation approaches for more marginal languages.

Our most significant gains were made from including additional data provided by a language advocacy organization. While exploring other potential sources of data, we identified many texts which we believe could further improve the performance of our system, but these were excluded due to ethical and legal concerns. Going forward, we believe that the most substantial improvements will be made not from architectural modifications or preprocessing strategies, but from our ability to forge similar partnerships with organizations and rights holders willing to contribute their data to open-source language tools.

This and future iterations of our tool will be available on a public GitHub repository⁶ where external partners will be recognized for their data contributions. In the longer-term, we would also like to test our tool against major large language models, both with one-shot prompting and supervised finetuning using our dataset. Regardless, we believe that the simple approach provided in this paper will be more than sufficient for most practical applications. It will carry the additional benefit of low training time, fast inference, and minimal package dependencies and hardware requirements.

⁶https://github.com/emmettstr/gwo-k_kreyol_classifier

Acknowledgments

Thank you to Axel Gauvin and the Reunion Creole advocacy organization *Lofis* for providing the supplemental data used in this project.

References

- Derek Bickerton. 1973. The Nature of a Creole Continuum. *Language*, 49(3):640–669.
- Pierre Cellier. 1985. *Description syntaxique du créole réunionnais: essai de standardisation*. Ph.D. thesis, Aix-Marseille 1.
- Omowumi Olabode Steven Ekundayo and Reliance Chekwubechukwu Enwerem. 2021. Writing Naija Right: A critical comparative review of Naija orthographies. *Journal of the Linguistic Association of Nigeria*, 24(2):82–98.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers*, pages 427–431.
- Amir Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. [GlotLID: Language Identification for Low-Resource Languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 6155–6218. Association for Computational Linguistics.
- Alina McLellan. 2024. [Presentational nana constructions in Reunion Creole: From bi-clausal cleft to monoclausal construction](#). *Isogloss. Open Journal of Romance Linguistics*, 10(1):1–27.
- John H McWhorter. 2018. *The Creole Debate*. Cambridge University Press.
- Salikoko S Mufwene. 2015. Pidgin and Creole Languages. *International encyclopedia of the social & behavioral sciences*, 18:133–145.
- Christine Iyetunde Ofulue. 2017. Sociocultural Perspectives on Orthographic Practices: Implications for Standardisation in Naijá. *Journal of The Linguistic Association of Nigeria*, 20(2).
- Nathaniel R Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Ones, Claire Bizon Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome A Etori, et al. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and Colonial African Creole Languages. *arXiv preprint arXiv:2405.05376*.