

# HT WMT 2026 CreoleMT System Description: Learning Hierarchies for Low-Resource Creole Language Identification

Ananya Ayasi

Hytton Technologies

Stockholm, Sweden

ananya.ayasi@gmail.com

## Abstract

This work is a submission to the Creole MT Creole Language Identification Shared Task conducted as a part of the Eleventh Conference on Machine Translation (WMT '26) colocated with EMNLP 2026, investigating whether hierarchical classification can improve language identification for closely related French-lexifier Creoles. We compare expert-designed linguistic hierarchies with data-driven hierarchies learned from model confusions and representation similarity, together with both hard and soft routing strategies. All models are built on a shared character-level TF-IDF representation with linear classifiers and are evaluated against a strong flat LinearSVC baseline and the multilingual GlotLID system. The best-performing approach is a data-driven combined hierarchy with soft routing, achieving a Macro-F1 of 87.54 while maintaining one of the lowest Macro False Positive Rates (0.00485). Analysis shows that training data availability remains the strongest predictor of language-level performance, with recall significantly correlated with training set size ( $\rho = 0.685$ ,  $p = 0.0139$ ). Manual inspection further reveals that most remaining errors arise from named entities, lexical borrowing, noisy text, and confusions between closely related Creole varieties. Overall, our results demonstrate that learned hierarchical organization provides a practical and effective alternative to expert-designed taxonomies for low-resource language identification while offering improved balance between classification performance and false-positive control.

## 1 Introduction

Language identification (LID) is a fundamental preprocessing step for multilingual natural language processing, enabling applications such as web-scale corpus construction, machine translation, and multilingual information retrieval. While language identification for high-resource languages is largely considered a solved problem, distinguish-

ing closely related low-resource languages and language varieties remains challenging due to lexical overlap, limited annotated data, non-standardized orthographies, and code-switching.

French-lexifier Creoles exemplify these challenges. Many varieties share a substantial proportion of their vocabulary with French and with each other while exhibiting geographically and historically driven linguistic variation. Recent work has introduced Kreyòl-MT, the first cumulative multilingual benchmark for French-lexifier Creole machine translation (Robinson et al., 2024), and highlighted the linguistic relationships between these languages in multilingual translation systems (Ayasi, 2025). However, comparatively little attention has been paid to language identification for this language family.

In this paper, we investigate language identification for the WMT 2026 Creole MT Shared Task (Robinson et al., 2026). Motivated by linguistic groupings proposed in prior work, we compare expert-designed hierarchical classifiers with data-driven hierarchies learned from model confusions and representation similarity. We further investigate hard versus soft hierarchical routing strategies and compare them against strong flat classification and multilingual language identification baselines. Finally, we analyze the remaining error cases to better understand the challenges of language identification for closely related Creole languages.

## 2 Related Work

Recent work has substantially expanded resources for French-lexifier Creoles. Robinson et al. (2024) introduced Kreyòl-MT, the first cumulative multilingual machine translation benchmark for several French-based Creoles, while Ayasi (2025) demonstrated that exploiting linguistic similarities between geographically related Creoles can improve multilingual translation.

Language identification for low-resource languages has increasingly been viewed as a data mining problem rather than purely a multiclass classification task. [Dent et al. \(2025\)](#) proposes web-scale corpus construction pipelines for French-based Creoles, highlighting challenges arising from lexical overlap with French and orthographic variation. Similar challenges have been observed for other low-resource languages, where standard language identification systems often perform poorly on noisy web text ([Caswell et al., 2020](#); [Bapna et al., 2022](#)). Neural language identification systems such as AfroLID ([Adebara et al., 2022](#)), GlotLID ([Kargaran et al., 2023](#)), and multilingual transformer-based approaches ([Sindane and Marivate, 2024](#)) have substantially expanded language coverage, but accurately distinguishing closely related language varieties remains difficult. More broadly, similar-language identification has long been recognized as a challenging problem because related languages frequently exhibit overlapping vocabularies, shared orthographic conventions, and gradual dialect continua ([Zampieri et al., 2020](#)).

### 3 Task

The WMT 2026 Creole Language Identification (LID) shared task requires assigning each input sentence to one of twelve language labels. The label inventory comprises nine French-lexifier Creole languages: Saint Lucian Creole (acf), Seychellois Creole (crs), Guadeloupean Creole (gcf), French Guianese Creole (gcr), Haitian Creole (hat), Louisiana Creole (lou), Martinican Creole (mart), Mauritian Creole (mfe), and Réunion Creole (rcf); the lexifier languages English (eng) and French (fra); and an oth class representing all remaining languages. The task is particularly challenging due to the high lexical overlap between related Creole varieties, similarities with their lexifier languages, and the limited availability of training data for many classes.

Following the shared task, systems are primarily evaluated using macro F1 score and macro False Positive Rate (Macro FPR), which jointly assess classification performance while accounting for class imbalance and the tendency to over-predict particular languages. For completeness, we additionally report overall accuracy, weighted F1 score, and per-language precision, recall, and F1 scores to provide a more detailed analysis of model performance across individual language varieties.

Our objective is to improve language identification performance for closely related Creole languages while maintaining low false positive rates through linguistically and data-driven hierarchical classification strategies.

## 4 Methods

All models are trained and evaluated using a unified experimental pipeline to ensure a fair comparison. Every trainable model uses the same training data, held-out test set, feature representation, preprocessing steps, and evaluation protocol, differing only in the classification strategy.

### 4.1 Unified Data Preparation

Training data is constructed from the official Kreyòl-MT training splits for all target languages. A single held-out evaluation set is similarly constructed from the official test splits. To model open-set rejection, the oth class is created using test examples from non-target languages that are not included in the target language inventory.

To prevent data leakage, duplicate sentences within each language are removed, examples assigned conflicting labels are discarded, and exact train-test overlaps are eliminated using normalized text matching. The resulting held-out evaluation set is frozen and identified using a SHA-256 fingerprint to ensure reproducibility across all experiments.

### 4.2 Character-Level Representation

All trainable models operate on a shared character-level TF-IDF representation. Character n-grams are used instead of word-level features as they are more robust to spelling variation, inconsistent orthography, and limited training data, which are common characteristics of low-resource Creole languages.

A single TF-IDF vocabulary is learned from the complete training set and reused by every model.

### 4.3 Flat Baseline

The flat baseline consists of a multiclass Linear Support Vector Machine (LinearSVC) trained directly on the shared TF-IDF representation. This model serves as the primary non-hierarchical baseline against which all hierarchical methods are compared. To reduce the effect of class imbalance, balanced class weights are used during training.

| Language                     | Flat  | Combined Soft | GlottLID |
|------------------------------|-------|---------------|----------|
| Saint Lucian Creole (acf)    | 93.41 | 93.64         | 73.46    |
| Seychellois Creole (crs)     | 96.64 | 96.85         | 81.58    |
| English (eng)                | 94.44 | 94.45         | 89.05    |
| French (fra)                 | 96.93 | 97.09         | 90.68    |
| Guadeloupean Creole (gcf)    | 81.21 | 80.83         | 45.73    |
| French Guianese Creole (gcr) | 48.98 | 48.98         | 38.24    |
| Haitian Creole (hat)         | 92.90 | 93.41         | 89.80    |
| Louisiana Creole (lou)       | 90.86 | 91.48         | 0.00     |
| Martinican Creole (mart)     | 92.37 | 92.17         | 0.00     |
| Mauritian Creole (mfe)       | 96.52 | 96.57         | 87.96    |
| Other (oth)                  | 96.75 | 96.88         | 86.10    |
| Réunion Creole (rcf)         | 63.41 | 68.09         | 29.03    |

Table 1: Per-language F1 scores for the strongest proposed model, the flat baseline, and GlottLID.

Table 2: Shared TF-IDF representation used by all trainable models.

| Parameter                  | Value        |
|----------------------------|--------------|
| Analyzer                   | Character    |
| N-gram range               | 2–6          |
| Maximum features           | 500,000      |
| Minimum document frequency | 2            |
| TF scaling                 | Sublinear TF |
| Normalization              | L2           |

Table 3: LinearSVC hyperparameters.

| Parameter              | Value     |
|------------------------|-----------|
| Classifier             | LinearSVC |
| Regularization ( $C$ ) | 2.0       |
| Class weights          | Balanced  |
| Maximum iterations     | 10,000    |

#### 4.4 Expert Linguistic Hierarchy

Our first hierarchical model incorporates expert linguistic knowledge based on the historical development and geographical distribution of Frenchlexifier Creoles. Languages are grouped according to established linguistic relationships before performing fine-grained language identification.

The hierarchy consists of four groups:

Table 4: Expert-designed linguistic hierarchy.

| Group   | Languages               |
|---------|-------------------------|
| Group 1 | acf, gcf, mart          |
| Group 2 | rcf, mfe, crs, lou, gcr |
| Group 3 | hat                     |
| Group 4 | eng, fra, oth           |

A first-stage classifier predicts the language group, after which a specialist classifier trained only on languages within the predicted group performs the final language identification.

Two variants are evaluated:

- **Hard hierarchy:** the predicted group determines which specialist classifier is used.
- **Soft hierarchy:** posterior probabilities from the group classifier are concatenated with the TF-IDF representation before final language prediction, allowing uncertainty in group membership to influence the final decision.

#### 4.5 Data-Driven Hierarchies

Rather than relying on expert knowledge, we also induce hierarchical structures directly from the training data. Three different language similarity measures are investigated.

1. **Confusion hierarchy:** languages are grouped according to their pairwise confusion patterns obtained from out-of-fold predictions of the flat classifier.
2. **Representation hierarchy:** languages are clustered according to cosine distances between their TF-IDF class centroids.
3. **Combined hierarchy:** an equally weighted combination of the normalized confusion and representation distance matrices.

For each distance matrix, agglomerative hierarchical clustering with average linkage is used to partition the languages into four groups. These learned hierarchies are subsequently used to train both hard and soft hierarchical classifiers using the same procedure described for the expert hierarchy.

#### 4.6 GlottLID Baseline

As a multilingual baseline, we evaluate GlottLID (Kargaran et al., 2023), a FastText-based language identification model covering over

1,600 languages. Predictions outside the target label inventory are mapped to the oth class to ensure compatibility with the shared evaluation protocol. Unlike the proposed approaches, GlotLID is used without retraining or adaptation.

## 5 Experimental Setup

### 5.1 Dataset

Experiments are conducted using the Kreyòl-MT dataset (Robinson et al., 2024), the first large-scale benchmark for machine translation across French-lexifier Creole languages. We formulate the language identification task over twelve labels consisting of ten target Creole languages together with English, French, and an additional oth class representing unsupported languages.

Although the shared task does not prescribe a specific evaluation split, we construct both the training and evaluation sets exclusively from the official Kreyòl-MT splits to ensure a controlled comparison across all models. Training examples are drawn from the official training splits, while evaluation uses only the official test splits.

The oth class is constructed from languages present in Kreyòl-MT that are not part of the target label set. To avoid leakage, oth training examples are taken only from the corresponding training splits, whereas evaluation examples are taken only from the official test splits. This provides a realistic closed-set benchmark while ensuring that every model is evaluated on exactly the same held-out data.

It should be noted that the oth category is inherently open-ended. Since Kreyòl-MT also contains additional Creole languages outside the target label inventory, some oth examples may be linguistically similar to the target Creoles and therefore represent particularly challenging negative examples.

### 5.2 Data Preprocessing

Several preprocessing steps are applied before training.

First, duplicate sentences within each language are removed. Sentences appearing under multiple language labels are discarded to eliminate ambiguous supervision. Exact train-test overlap is then removed using normalized text matching to prevent evaluation leakage.

All text is normalized using Unicode NFC normalization and whitespace normalization while preserving the original orthography. No stemming,

| Language     | Train         | Test          |
|--------------|---------------|---------------|
| acf          | 3,733         | 435           |
| crs          | 2,048         | 222           |
| eng          | 3,230         | 3,096         |
| fra          | 3,218         | 2,060         |
| gcf          | 3,272         | 257           |
| gcr          | 53            | 37            |
| hat          | 49,692        | 2,108         |
| lou          | 1,557         | 174           |
| mart         | 4,259         | 514           |
| mfe          | 22,135        | 874           |
| oth          | 5,173         | 4,594         |
| rcf          | 184           | 28            |
| <b>Total</b> | <b>98,554</b> | <b>14,399</b> |

Table 5: Number of training and held-out test examples for each language after preprocessing. The held-out test set is constructed exclusively from official Kreyòl-MT test splits after removing duplicate instances, cross-label conflicts, and exact train-test overlap.

tokenization, or language-specific preprocessing is applied.

The resulting training and evaluation sets are frozen for all experiments. Every model, including hierarchical variants and external baselines, is evaluated on exactly the same held-out examples to ensure a fair comparison.

Table 5 summarizes the final dataset after preprocessing. The training set contains 98,554 examples across twelve labels, while the held-out evaluation set contains 14,399 examples. During preprocessing, 740 training instances with conflicting labels were removed. For the evaluation set, 80 duplicate instances, two cross-label conflicts, and 538 exact train-test overlaps were discarded, resulting in a leakage-free held-out benchmark.

**Class imbalance and resampling.** No resampling, oversampling, undersampling, or synthetic data augmentation techniques were applied during training. Apart from constructing comparable English and French negative examples, the original class distributions were preserved. This decision was made intentionally to evaluate language identification under realistic low-resource conditions rather than artificially balancing the dataset. In particular, several target languages (e.g., gcr and rcf) contain orders of magnitude fewer training examples than high-resource varieties such as Haitian Creole (hat) and Mauritian Creole (mfe). Preserving this imbalance enables a more faithful assessment of model robustness across varying resource levels and allows us to investigate the relationship between training data availability and language identification performance.

### 5.3 Baselines

We compare the proposed hierarchical classifiers against two baselines.

**Flat LinearSVC.** A character-level TF-IDF representation followed by a balanced Linear Support Vector Machine serves as the primary task-specific baseline. This architecture has consistently shown strong performance for low-resource language identification while remaining computationally efficient.

**GlottLID.** We additionally evaluate GlottLID (Kargaran et al., 2023), a FastText-based multilingual language identification system covering more than 1,600 languages. GlottLID is used without any task-specific retraining. Predictions outside the target label inventory are mapped to the oth class.

### 5.4 Training Details

All trainable models share the same character-level TF-IDF representation using character n-grams of lengths 2–6. The TF-IDF vocabulary is learned only from the training data and reused across every experiment.

LinearSVC classifiers are trained using balanced class weights with regularization parameter  $C = 2.0$ . Hierarchical models reuse the same feature representation, differing only in their classification strategy. Group classifiers within the soft hierarchy are implemented using logistic regression optimized with stochastic gradient descent, while specialist language classifiers use LinearSVC.

Data-driven hierarchies are induced exclusively from the training data using three-fold out-of-fold predictions, ensuring that no evaluation labels influence hierarchy construction.

### 5.5 Shared-task submissions

For the official WMT 2026 Creole Language Identification shared task, we submitted three systems selected from the data-driven soft-routing variants. Our primary submission used the combined soft hierarchy, in which hierarchical groups are induced from an equally weighted combination of confusion-based and representation-based distances. Contrastive submission 1 used the representation soft hierarchy, where groups are derived from cosine distances between character TF-IDF language centroids. Contrastive submission 2 used the confusion soft hierarchy, where groups are induced from out-of-fold confusion patterns of the flat classifier. All three submissions use the same

character-level TF-IDF representation and soft-routing architecture, differing only in the signal used to induce the language hierarchy. This selection allows the official submissions to compare a combined hierarchy against two complementary sources of structure: distributional similarity in the representation space and empirical classifier confusion patterns.

### 5.6 Evaluation Protocol

Performance is evaluated using Accuracy, Macro-F1, Weighted-F1, and Macro False Positive Rate (Macro-FPR).

Following the WMT 2026 shared task, Macro-F1 and Macro-FPR are treated as the primary evaluation metrics. Macro-F1 measures balanced performance across all languages, while Macro-FPR quantifies the tendency of a model to incorrectly assign languages across the label inventory. Accuracy and Weighted-F1 are additionally reported to provide a more complete picture of overall classification performance.

To ensure reproducibility, all experiments are performed on a single frozen held-out evaluation set, and the corresponding test set fingerprint (SHA-256 hash) is reported.

## 6 Results

Table 6 summarizes the performance of all evaluated systems on the held-out Kreyòl-MT test set. The proposed hierarchical approaches are compared against a flat TF-IDF + LinearSVC classifier and the multilingual GlottLID baseline.

Overall, the proposed hierarchical models consistently outperform GlottLID across all evaluation metrics. Among the proposed methods, soft hierarchical routing consistently achieves higher Macro-F1 and lower Macro-FPR than the corresponding hard hierarchical variants. The best Macro-F1 is obtained by the data-driven combined soft hierarchy (87.54%), while the expert soft hierarchy achieves the lowest Macro-FPR (0.004846). The flat LinearSVC baseline remains highly competitive, obtaining a Macro-F1 of 87.04%, demonstrating the strength of character-level TF-IDF representations for Creole language identification.

Table 1 reports the per-language F1 scores. Performance remains high for the majority of languages, particularly English, French, Haitian Creole, Mauritian Creole, and the oth category. Lower-resource languages such as French Guianese Creole

| Model                             | Accuracy     | Macro-F1     | Weighted-F1  | Macro-FPR      |
|-----------------------------------|--------------|--------------|--------------|----------------|
| Flat TF-IDF + LinearSVC           | 94.93        | 87.04        | 94.91        | 0.00501        |
| Expert hierarchy (hard)           | 94.83        | 81.43        | 94.71        | 0.00510        |
| Expert hierarchy (soft)           | <b>95.13</b> | 86.14        | <b>95.08</b> | <b>0.00485</b> |
| Data-driven confusion (hard)      | 94.06        | 84.53        | 94.04        | 0.00584        |
| Data-driven confusion (soft)      | 94.92        | 86.79        | 94.89        | 0.00502        |
| Data-driven representation (hard) | 94.54        | 83.09        | 94.48        | 0.00536        |
| Data-driven representation (soft) | 95.06        | 86.79        | 95.03        | 0.00487        |
| Data-driven combined (hard)       | 94.14        | 84.34        | 94.12        | 0.00577        |
| Data-driven combined (soft)       | 95.08        | <b>87.54</b> | 95.06        | 0.00485        |
| GlottLID                          | 83.96        | 59.30        | 82.53        | 0.01743        |

Table 6: Overall performance on the held-out Kreyòl-MT test set. The official shared-task metrics are Macro-F1 and Macro-FPR.

(gcr) and Réunion Creole (rcf) remain substantially more challenging across all systems.

## 7 Analysis

The proposed data-driven combined hierarchy with soft routing achieved the best overall performance among all evaluated systems, obtaining a held-out accuracy of **95.08%** (708 errors out of 14,399 examples) and the highest Macro-F1 (**87.54**). Although the expert hierarchy achieved the lowest Macro-FPR, the absolute difference between the two systems was only  $5 \times 10^{-6}$ , while the combined hierarchy improved Macro-F1 by approximately 1.4 points. This demonstrates that the learned hierarchy provides a better overall trade-off between balanced classification performance and false-positive control. All remaining analyses therefore focus on the best-performing data-driven combined soft hierarchy.

### 7.1 Training data remains the dominant performance bottleneck

Figure 1 shows the relationship between retained training examples and language-level performance. Performance is strongly associated with data availability. The lowest-resource languages, Guianese Creole (GCR; 53 training examples) and Réunion Creole (RCF; 184 training examples), exhibit substantially higher error rates than higher-resource languages.

Across languages, retained training size is significantly correlated with recall (Spearman  $\rho = 0.685$ ,  $p = 0.0139$ ). Likewise, error rate decreases significantly as training data increases (Spearman  $\rho = -0.685$ ,  $p = 0.0139$ ). In contrast, the correlation with Macro-F1 is weaker and not statistically significant ( $\rho = 0.427$ ,  $p = 0.167$ ), indicating that factors beyond dataset size—such as lexical overlap and linguistic similarity—also contribute

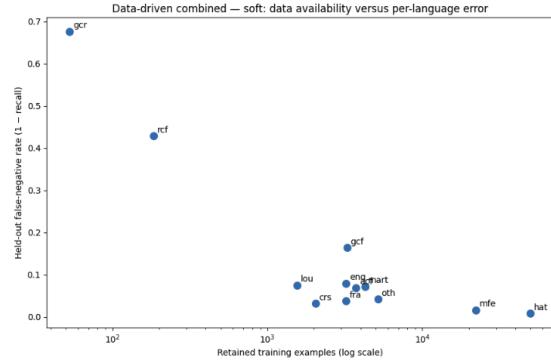


Figure 1: Relationship between retained training examples and language-level performance. Larger datasets generally yield lower error rates and higher recall. Spearman correlation confirms a significant association between training data size and both recall ( $\rho = 0.685$ ,  $p = 0.0139$ ) and error rate ( $\rho = -0.685$ ,  $p = 0.0139$ ), indicating that data availability remains the primary determinant of classification performance.

| Metric     | Spearman $\rho$ | $p$ -value |
|------------|-----------------|------------|
| Recall     | 0.685           | 0.0139     |
| Macro-F1   | 0.427           | 0.1667     |
| Error rate | -0.685          | 0.0139     |

Table 7: Correlation between retained training examples and language-level performance.

to classification difficulty.

### 7.2 Errors are highly concentrated

Rather than being uniformly distributed across language pairs, the model’s mistakes are concentrated within a small number of recurrent confusions.

The four largest confusion pairs account for more than half of all held-out errors (52.0%), while only eleven pairs explain over 75% of all errors. The largest confusion is English→Haitian (122 errors), followed by English→Other (85), Other→Haitian (83), and Other→English (78). Together these reflect ambiguity between English, the heterogeneous Other category, and Haitian Creole.

| Confusion pair | Errors | Share (%) |
|----------------|--------|-----------|
| eng → hat      | 122    | 17.2      |
| eng → oth      | 85     | 12.0      |
| oth → hat      | 83     | 11.7      |
| oth → eng      | 78     | 11.0      |
| fra → hat      | 48     | 6.8       |
| eng → fra      | 28     | 4.0       |
| mart → gcf     | 25     | 3.5       |
| gcf → mart     | 24     | 3.4       |

Table 8: Largest held-out confusion pairs.

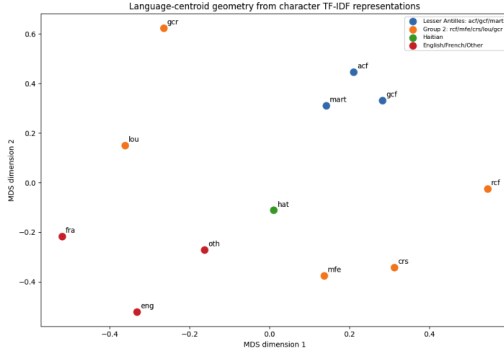


Figure 2: Two-dimensional projection of language centroids derived from character TF-IDF representations using UMAP. Languages belonging to the same expert linguistic family occupy nearby regions of the representation space despite the hierarchy being learned without linguistic supervision. This supports the conclusion that the learned hierarchy captures meaningful linguistic structure.

Within the Lesser Antillean varieties, the strongest bidirectional confusion occurs between Martinican and Guadeloupean Creole (25 and 24 errors respectively), consistent with their close linguistic relationship.

### 7.3 The learned hierarchy reflects linguistic structure

Although the hierarchy is learned entirely from data, it captures meaningful linguistic organization. Figure 2 visualizes language centroids obtained from character TF-IDF representations. Representation-based clustering exhibits the strongest agreement with the manually designed expert hierarchy (ARI=0.224, NMI=0.580), outperforming confusion-based and combined clustering. Furthermore, languages belonging to the same expert family occupy significantly more compact regions of representation space than languages from different families. A permutation test confirms that this compactness is unlikely to arise by chance ( $p = 0.022$ ).

These findings suggest that hierarchical struc-

| Hierarchy      | ARI          | NMI          | Pairwise agreement |
|----------------|--------------|--------------|--------------------|
| Representation | <b>0.224</b> | <b>0.580</b> | 0.667              |
| Confusion      | 0.164        | 0.489        | 0.667              |
| Combined       | 0.071        | 0.418        | 0.652              |

Table 9: Agreement between learned hierarchies and the manually designed expert taxonomy.

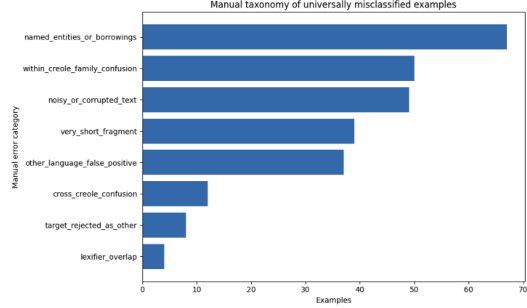


Figure 3: Manual categorization of the 266 evaluation examples misclassified by every evaluated model. Named entities and lexical borrowings constitute the largest category, followed by within-family Creole confusions, noisy or corrupted text, and very short utterances.

ture emerges naturally from distributional similarity without requiring explicit linguistic supervision.

### 7.4 Remaining failures are systematic

Manual inspection of the 266 examples misclassified by every evaluated model reveals that the remaining failures exhibit recurring patterns rather than being uniformly distributed. Figure 3 summarizes the resulting error taxonomy.

The analysis was performed manually by the author rather than through an automated classification procedure. Each universally misclassified example was inspected together with its gold and predicted language labels and assigned to one of eight error categories based on observable properties of the input and the relationship between the gold and predicted labels. Categories included named entities or lexical borrowings, within-family Creole confusion, noisy or corrupted text, very short fragments, other-language false positives, cross-Creole confusion, target languages rejected as other, and lexifier overlap. The categories were defined during an initial inspection of the errors and subsequently applied to the complete set of 266 examples. Each example was assigned a single category according to the most salient apparent source of ambiguity.

The author is not a native speaker or linguistic expert in the Creole varieties considered in this task. Consequently, the analysis does not attempt

| Category                      | Count | Share (%) |
|-------------------------------|-------|-----------|
| Named entities / borrowings   | 67    | 25.2      |
| Within-family confusion       | 50    | 18.8      |
| Noisy or corrupted text       | 49    | 18.4      |
| Very short fragment           | 39    | 14.7      |
| Other-language false positive | 37    | 13.9      |

Table 10: Manual taxonomy of examples misclassified by every evaluated model.

to make fine-grained linguistic judgments about whether individual lexical or grammatical forms uniquely identify a variety. Instead, the taxonomy is intended as a descriptive analysis of recurring properties among errors. Judgments involving language relationships were restricted to the language-family and lexifier information used in the experimental design, while categories such as short fragments, named entities, and corrupted text were assigned from directly observable characteristics of the examples.

Named entities and lexical borrowings constitute the largest category (25.2%), followed by within-family Creole confusions (18.8%) and noisy or corrupted text (18.4%). Very short fragments account for a further 14.7%, indicating that a substantial fraction of universal errors occur in inputs containing limited language-identifying evidence.

Overall, the analysis suggests that errors shared across all evaluated systems are concentrated in identifiable input types, particularly lexical overlap, closely related Creole varieties, noisy text, and context-poor utterances. Because the annotation was performed by a single non-expert annotator, however, the taxonomy should be interpreted as exploratory rather than as a definitive linguistic characterization of the errors.

### 7.5 Prediction confidence is well calibrated

Prediction margin provides an effective estimate of classification uncertainty. Figure 4 shows the coverage–risk trade-off obtained by selectively retaining only the highest-confidence predictions.

Examples in the lowest-confidence quintile exhibit an error rate of 22.3%, whereas the highest-confidence quintile contains no errors. Retaining only the highest-confidence 80% of predictions increases accuracy from 95.1% to approximately 99.4%. This demonstrates that confidence scores provide an effective mechanism for uncertainty estimation and selective prediction, making the classifier suitable for practical deployment scenarios in which uncertain examples can be referred for

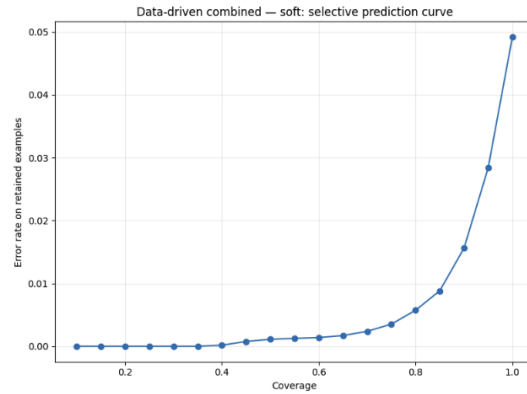


Figure 4: Coverage–risk trade-off for the best-performing data-driven combined hierarchy using soft routing. Predictions are ranked by confidence margin, and progressively lower-confidence examples are discarded. Retaining only the highest-confidence predictions substantially reduces error, reaching nearly perfect accuracy while maintaining high coverage. This demonstrates that prediction confidence provides an effective uncertainty estimate suitable for selective prediction.

manual review.

## 8 Conclusion

We presented a systematic comparison of flat, expert-designed hierarchical, and data-driven hierarchical approaches for language identification in the WMT 2026 Creole Language Identification Shared Task. Across all evaluated systems, soft hierarchical routing consistently outperformed hard routing, indicating that propagating uncertainty between hierarchy levels is more effective than committing to a single routing decision.

Among the proposed methods, the data-driven combined hierarchy achieved the highest Macro-F1 while maintaining a Macro False Positive Rate comparable to the best expert-designed hierarchy. Although the improvements over the flat baseline were modest, the learned hierarchy required no linguistic expertise and generalized well across multiple low-resource Creole varieties. Representation analysis further showed that the induced hierarchy captures meaningful linguistic structure despite being learned solely from data.

Our analysis also highlights that data scarcity remains the principal limitation for low-resource language identification. Languages with the smallest training sets exhibited substantially lower recall, and language-level performance was significantly associated with training data availability. The remaining errors were concentrated in a small num-

ber of systematic categories, particularly named entities, lexical borrowing, noisy text, and confusions between closely related Creole varieties.

Overall, our results suggest that lightweight character-level models remain highly competitive for low-resource language identification and that automatically learned hierarchical structure offers a practical alternative to manually designed linguistic taxonomies. We hope that the accompanying analyses provide useful insights for future work on Creole language identification and multilingual modeling for closely related low-resource languages.

## Limitations

Our study has several limitations. First, the proposed models rely exclusively on character-level TF-IDF representations and linear classifiers. While this allows efficient training and strong performance in low-resource settings, it does not exploit contextual information available in modern multilingual encoder models. Future work could investigate whether hierarchical classification provides similar benefits when combined with contextual representations.

Second, the learned hierarchies are induced from a single dataset and language inventory. Their generalizability to other Creole families, additional dialects, or broader multilingual language identification benchmarks remains to be established.

Third, the evaluation reflects the class distribution of the Kreyòl-MT corpus, which is highly imbalanced. Although this provides a realistic low-resource setting, the extremely limited training data available for languages such as French Guianese Creole and Réunion Creole inevitably constrains achievable performance. Additional annotated data would likely provide larger gains than architectural modifications alone.

Finally, our manual error analysis focused only on examples misclassified by every evaluated model. While this identifies systematic failure modes shared across classifiers, it does not capture model-specific error patterns that could further inform future improvements.

## References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. Afrolid: A neural language identification tool for african languages. In *Proceedings of the 2022 Conference on*

*Empirical Methods in Natural Language Processing*, pages 1958–1981.

Ananya Ayasi. 2025. [Krey-all WMT 2025 CreoleMT system description: Language agnostic strategies for low-resource translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1158–1165, Suzhou, China. Association for Computational Linguistics.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan Van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Isaac Caswell, Theresa Breiner, Daan Van Esch, and Ankur Bapna. 2020. Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608.

Rasul Dent, Pedro Ortiz Suarez, Thibault Cl rice, and Beno t Sagot. 2025. Identifying rare languages in common crawl data is a needles-in-a-haystack problem. In *EMNLP (Findings)*, pages 1460–1473.

Amir Hossein Kargaran, Ayyoob Imani, Fran ois Yvon, and Hinrich Schuetze. 2023. Glotlid: Language identification for low-resource languages. *arXiv preprint arXiv:2310.16248*.

Nathaniel Robinson, Ruth-Ann Armstrong, Claire Bizon Monroc, Rasul Dent, Pranav Gupta, Raj Dabre, Andre Coy, and Kenton Murray. 2026. Findings of the second shared task for creole language machine translation, at WMT 2026. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Lo c Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, et al. 2024. Krey l-mt: Building mt for latin american, caribbean and colonial african creole languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110.

Thapelo Andrew Sindane and Vukosi Marivate. 2024. From n-grams to pre-trained multilingual models for language identification. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 229–239.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.