Identifying Rare Languages in Common Crawl Data is a Needles-in-a-Haystack Problem

Rasul Dent[1], **Pedro Ortiz Suarez**[2], **Thibault Clérice**[1], **Benoît Sagot**[1]
[1]Inria, Paris, {firstname.lastname}@inria.fr
[2]Common Crawl Foundation, Paris, pedro@commoncrawl.org

Abstract

Automatic language identification is frequently framed as a multi-class classification problem. However, when creating digital corpora for less commonly written languages, it may be more appropriate to consider it a data mining problem. For these varieties, one knows ahead of time that the vast majority of documents are of little interest. By minimizing resources spent on classifying such documents, we can create corpora covering previously overlooked languages faster than existing pipelines. To demonstrate the effectiveness of the targeted mining perspective, we introduce a new pipeline that can filter a single snapshot in two hours. We also provide web corpora for several French-based Creoles.

1 Introduction

As Natural Language Processing (NLP) technologies gain prominence, so does the demand for corpora. Filtered versions of Common Crawl data, such as OSCAR (Ortiz Suárez et al., 2019; Abadji et al., 2021, 2022), MADLAD-400 (Kudugunta et al., 2024), GlotCC (Kargaran et al., 2024), and Fineweb-2 (Penedo et al., 2025), are one solution to this demand. For English and Mandarin, this approach has yielded terabyte-sized datasets. Yet, even for widely written languages like Romanized Arabic and Hindi, filtering Common Crawl has yielded modest results.

As such, language identification (LID) for the thousands of varieties with little representation in web corpora has become a prominent research agenda (Caswell et al., 2020). For example, Kreutzer et al. (2022) document some types of noise that affect less common languages. Similarly, OpenLID Burchell et al. (2023), GlotLID Kargaran et al. (2023), and Adebara et al. (2022), try to improve the coverage of LID models.

At the same time, Creole languages have garnered considerable attention (Lent et al., 2022). They generally have strong lexical overlap with certain widely-spoken languages, such as English, French, and Portuguese, but differ in morphosyntax. Although often lumped into a broad "low-resource" category, contact varieties bring unique challenges and opportunities for NLP (Bird, 2022).

For French-based Creoles (FCs), a group of 10-20 closely-related languages[1] spoken by approximately 15 million people, NLP in general remains challenging. These languages share many linguistic traits with French and even more with each other, which greatly limits the effectiveness of existing LID solutions. Table 1 shows how FCs draw on French words, but do not inflect verbs or adjectives for person or number, unlike the French *naissent* '(they) are born' or *égaux* 'equal (plural)'.[2] Their written standards also have much lower grapheme-phoneme ratios, which distances them from French but not necessarily from each other.

At the same time, these languages exist in distinct sociolinguistic contexts. At one end of the spectrum lies Haitian Creole, which accounts for roughly 2/3 of FC speakers. It is the national language of Haiti and spoken by sizable diaspora communities in countries like the United States. As such, it has a robust online presence, and is well-represented in LID benchmarks like FLORES-200 (NLLB_Team et al., 2022). At the other end, the Creoles of Louisiana and Trinidad are critically endangered and have limited online presence. In the middle, Lesser Antillean,[3] French Guianese Creoles, Mauritian and Réunionese Creoles are partly institutionalized.

Although models like GlotLID and AfroLID (Adebara et al., 2022), try to take more FCs into

[1]The number depends on how we enumerate dialect chains.

[2]The Haitian *egal ego* shows such forms were often repurposed, but a full discussion of this is beyond our scope.

[3]The FCs of the Lesser Antilles form a dialect chain. Due to the lack of differentiation between the islands in previous work, here we treat this chain as one language.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 1460–1473

Language	Text
Haitian	Tout moun fèt lib, egal ego pou diyite kou wè dwa.
Antillean	Tout moun né lib èk égal an dignité èk dwa.
Mauritian	Tou imin vinn lor later lib ek egal an drwa ek an dignite.
French	Tous les êtres humains naissent libres et égaux en dignité et en droits.
English	All human beings are born free and equal in dignity and rights.

Table 1: The First Sentence of Article 1 of the UDHR shows the lexical similarity of FCs to French and even closer affinity to one another. (https://www.ohchr.org/)

account, even recent filters of Common Crawl snapshots, like GlotCC (Kargaran et al., 2024) identify as few as 49 pages of content in Réunion Creole and 100 in Lesser Antillean Creole, each of which have vibrant online speech communities. While such efforts have improved the situation for some languages, they still generally approach corpora creation within a framework intended for the most common varieties, overlooking text distributions.

More specifically, pipelines like Ungoliant (Abadji et al., 2021, 2022) and GlotCC break every document into sentence-to-paragraph-length segments. They then use multiclass classifiers, and especially the fastText architecture (Joulin et al., 2017) on the resulting segments. Yet, when we target a very small fraction of the Web, most data can be discarded with much less effort.

With this in mind, we reframe (Creole) LID as a Needles-in-a-Haystack problem, and propose discriminative feature mining as a solution. Our main claim is that we can efficiently identify a small French Creole cluster in large webcrawls by using a document-level Bag-of-Types strategy. To demonstrate this, we first introduce our threshold-based filtration system and then benchmark speed, recall, and false positive rate on clean Wikipedia data. Next, we estimate the recall capabilities on noisy web data by applying it to Creole subcorpora from three recent projects. After that, we eliminate 99% or more of distracting documents in a 2.6 billion page Common Crawl snapshot in a few hours on a medium-sized cluster while maintaining competitive recall. Finally, we explore additional passes and the remaining "last kilometer problem" of fine-grained LID. We will release our source code and filtered versions of Fineweb-2 and the results of first pass filtering on the December 2024 Common Crawl snapshot for each target label. Additionally, we will offer second pass corpora for Lesser Antillean and Mauritian Creoles.[4]

[4] https://github.com/DEFI-COLaF/LanguageMining.

2 Related Work

Due to its important role in multilingual NLP, LID has a long history, which Jauhiainen et al. (2024) resume in depth. By the 90s, approaches based on n-gram frequencies, including Cavnar and Trenkle (1994) had achieved 99% accuracy on monolingual documents of sufficient length in several common languages.

However, many subproblems like closely-related varieties, short texts, and code-switching remained open (Da Silva and Lopes, 2006). During the late 2000s and early-mid 2010s, alternative approaches were explored, culminating in the adoption of sentence-level linear classifiers as a *de facto* standard. In this section, we briefly review the usage of linear classifiers for language identification, and then explore how keyword methods provide a useful alternative.

2.1 LID with Linear Classifiers

Since the late 2010s, fastText (Joulin et al., 2017), which relies on Continuous Bag-of-Words vectors, a single hidden layer, and a multiclass output layer, has been the basis of notable web pipelines. Google's CLD3, a popular alternative, utilizes many of the same principles, such as character n-grams, the hashing trick, and a shallow network.[5] Given the similar advantages of fastText and CLD3, these have become backbone LID baselines (Burchell et al., 2023; Kargaran et al., 2023).

Grave et al. (2018) also introduced a corpus creation pipeline based on applying the fastText model to each line in a raw data dump, and then appending each line that passed a certain LID confidence threshold to the relevant corpus. Shortly thereafter, the OSCAR project refined this idea by parallelizing the transfer of data, and moving cleaning to occur before language identification (Ortiz Suárez et al., 2019), and then incorporating metadata and new optimizations (Abadji et al., 2021), ultimately

[5] https://github.com/ropensci/cld3, see Botha et al. (2017).

resulting in the document-level Ungoliant pipeline (Abadji et al., 2022).

2.2 Problems with LID at Scale

Although OSCAR and older initiatives like C4 (Habernal et al., 2016) worked for the most common languages, they overlooked most other languages. Caswell et al. (2020) detail conditions that cause models with high coverage under test conditions to be effectively unusable for many ostensible targets when applied at scale. Beyond obvious issues such as non-Unicode encoding, high-resource/out-of model related languages, and short texts, there are mis-rendered PDFs, scripts mixed for visual effect, text with spaces between every character, and improbable repetitions of n-grams in high resource languages. For these issues, they suggest post-filtering using lists words uncommon in the high-resource cousin, which can be arbitrarily precise. In addition, they also suggest self-supervised Transformer-based models, but these have the disadvantage of much slower runtimes. Kreutzer et al. (2022) expand upon this work by identifying additional sources of error, such as inconsistent and incorrect use of language codes.

As noted previously, Creoles share lexical affinities with more well-resourced languages, which, when combined with orthographic instability, make LID particularly challenging (Lent et al., 2024). For example, Caswell et al. (2020) show that even when ensembling complementary LID strategies, Naija (Nigerian Pidgin), remains one of the hardest languages to detect at scale due to its high overlap with English.

With the exception of Haitian Creole, LID for French-based Creoles has largely been pursued within the framework of models meant to detect at least one hundred languages. MADLAD-400 (Kudugunta et al., 2024), built from multiple snapshots of Common Crawl, included several Creoles such as the French-based St. Lucian, Mauritian, and Seselwa, and the English-based Eastern Maroon and Belizean Creoles among their 419 languages. At the time of writing, GlotLID Kargaran et al. (2023) offers the most extensive coverage across Creole languages, and has recently been used to create two large scale corpora, GlotCC (Kargaran et al., 2024) and FineWeb-2 (Penedo et al., 2025), which will serve as points of reference for our system and are further described in Section 4.2.

2.3 Keyword Search

Keyword methods date back to the early days of LID (Prager, 1999; Jauhiainen et al., 2024). An important example was the Crúbadán Project, which revolved around modeling languages as search queries for a custom web crawler (Scannell, 2007). More specifically, they identified languages like Irish by searching combinations of at least one common, yet distinctive stopword and at least one other word common to the language using the 'AND' and 'OR' operations. For validating the results of the query and outgoing links, the crawler augmented simple character trigram frequencies with basic metadata about the relevant languages, such as languages they are likely to co-appear with.[6]

More recently, Lau et al. (2024) used key-string methods to distinguish written Cantonese, Standard Written Chinese, and intermediate varieties (either mixed or unmarked). They emphasize that focusing on string operations allows for LID decisions to be made up to 4x faster than the commonly used fastText lid.176 model. Similarly, the Molyé project implemented a keyword-based approach to identify historical examples of nonstandard French-related varieties (Dent et al., 2024).

3 Multilevel Feature Mining

The phrase "needle in a haystack" is a longstanding metaphor for search scenarios where one or more target "needle(s)" is/are embedded in a very large number of distracting "hay" records (Crammer and Chechik, 2004). The heart of a Needle(s)-in-a-Haystack problem is to eliminate hay quickly. For language identification, this means ignoring documents that lack (quasi)-unique features of our target language(s). To find the few documents that **do** contain clusters of such features, we create a kind of search engine that uses lexicons as queries and returns ranked document- and line-level corpora. To implement this data triage, we introduce an indexing-scoring system that operates in two phases: first at the document-level, and then at the sentence level. The main components of this method are shown in Figure 1 and formalized as pseudocode in Algorithm 1. The remainder of this section details the logic of each component.

[6]For example, pages with content in Lingala would likely also contain French, and English might skew any language.

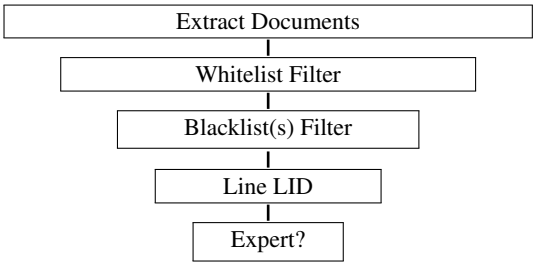

Figure 1: Multilevel Feature Mining Pipeline

3.1 Document-level

Since we want to identify a very small portion of the overall data, we take an approach of staggered filtration. Importantly, filters can have very different runtimes. The task then is to order them according to mean runtime per example rejected, while repeating as little work as possible. For the initial filter, the most important step is to see if the characteristic traits of the rare class appear in the document at all, which can be accomplished by extracting features and comparing them to whitelists of distinctive features. As methods that depend on a few key features are liable to capture noise characterized by the same features, the concomitant use of blacklists helps to reduce false positives.

3.1.1 Whitelists

For languages with distinct orthographies and relatively little bound morphology, as is the case for French-based Creoles, lexicon-based filters are a very fast and (potentially) high-precision way to establish that a document likely contains data in a target language. For the prototype, we focus on whitespace tokens, which align closely with words in the case of FCs (and many, but not all, other languages), because this allows us to use fast built-in string methods. However, whitelist methods do not inherently require tokenization. Through hashing, they can be generalized to work with any substring. For languages that are heavily inflected or written in a script that relies less on whitespace, fixed-length character windows are a fast, generalizable, and readily available alternative (Google Research, 2025). We further note that Caswell et al. (2020) have created lists of 1000 tokens using term-frequency inverse-internet frequency (TF-IIF) analysis for several of our target languages, based using a wide sample of general Internet data for background frequencies. In the interest of open science, we use these lists as our whitelists.

Ranking	Antillean	Haitian	Mauritian
1	épi	pou	zot
2	pou	fè	pou
3	mwen	moun	bann
4	èk	yon	finn
5	zòt	mwen	dimoun

Table 2: Top 5 disjunctive words for 3 TF-IIF lists

To concretely demonstrate the intuitions and limitations of the whitelist approach, we briefly review the top 5 words from the Lesser Antillean (acf), Haitian (ht) and Mauritian (mfe) wordlists. In line with the overview presented in Section 1, these words all have direct French etymologies (e.g. *pou* < Fr. *pour* 'for'), but are easily distinguishable from their etymons when written in the normative orthographies. Despite being on opposite sides of the planet, identical matches like (*pou*) are not rare, words differing only by a diacritic like *zòt/zòt* '2PL' or incorporated article (*(di)moun* 'person/people') are very common. These considerations mean that distinction is possible, but nonstandard and especially etymologizing spelling can degrade performance. This issue is especially acute for English-lexifier languages, a point we explore further in Section 6.3. In Section 4.3.1, we briefly consider the impact of punctualization as well by adding spaces before commas and ending punctuation.

3.1.2 Blacklists

At the line level, it is well-established that one or two collisions of rare n-grams can be enough to lead a classifier to treat noise as the target language (Caswell et al., 2020; Kreutzer et al., 2022). However, this problem is much easier to manage with document-level blacklists, because we directly characterize and filter specific kinds of noise using the same extracted features. For whitespace tokenization, this method, is effective for eliminating various kinds of spam, especially adult websites. After initial testing of the whitelist approach yielded substantial amounts of content from machine-translated pornographic content, we introduced a simple 5 word blacklist that effectively eliminated this king of noise.[7]

In some scenarios, distractor languages can also be filtered with blacklists (Ljubešić et al., 2007). However, we do not use this approach for the

[7]The five words are: "porn", "porno", "porna", "sex", and "xxx". Any document with two or more of these words is discarded.

document-level first pass because it would harshly penalize the multilingual content that makes up a substantial part of the overall language precess (e.g. bilingual French and Creole books).

3.1.3 Scoring, Ranking and Indexing

To identify candidate documents, we first calculate both a whitelist score (wsc). This score can be computed in two ways: using a boolean match per type, or using type or token frequencies. Additionally, one could follow Scannell (2007) and make certain types mandatory. For the initial implementation, we use simple the sum of boolean matches, as this does not require calibrating relative frequency information. For efficiency, we store each list as a set and calculate this score as the intersection of the list and the document's token-types.

Having calculated whitelist scores, we exclude the documents whose scores are below a certain value (threshold). For the documents that pass this threshold, we then calculate a blacklist score (bsc) using the same scoring mechanism, and eliminate scores above a second value called the tolerance. Next, we rank the remaining documents by whitelist score. This allows us to prioritize the highest scoring ones, which are likely either in the target language or special edge cases (see Algorithm 1).

If we have sufficient space and are interested in eventually exploring other languages, we can greatly reduce the runtime of future searches by saving the vocabularies of all (or most) documents as indices. When space is limited and/or we are sure that we are only interested in specific languages, we only need to save the highest scoring documents.

Additionally, after the first pass, we can inspect the ranked data to identify specific sources of reliable or confusing data. For closely-related languages, this is an easy to way capture human intuitions that may be difficult to model succinctly.

3.2 Line-level

At the line-level, we have a wide range of possible filters. The simplest of all is a length check, which can help remove common boilerplate (Kohlschütter et al., 2010). However, this may exclude list-based content like dictionaries. Script checks and line-level keyword filters are slightly more expensive, but still lightweight options. Beyond this, we can still classify sentences using more intensive models, such as fastText or even Transformers. The main difference between the staggered filtration approach and the more established pipelines is that we aim to only use the intensive models when needed.

```
input : W whitelist, B blacklist(s), D documents
for i ← 0 to len(D) do
    First get token-types
    tokens ← tokenize(D);
    types ← set(tokens);
    Then score
    wsc ← score(types, W);
    Optionally cache all vocabularies
    Then filter
    if wsc[i] ≥ threshold then
        bsc ← score(types, B);
        if bsc < tolerance then
            save((D[i], wsc[i]))
        end
    end
end
Sort saved documents by score (descending)
```

Algorithm 1: Double List Filtering Algorithm

In this work, we explore the extent to which line-level filtering can also be performed with wordlist-based approaches. For ranking lines, we start with a type-based score, like at the document-level. However, observing that naive line-tokenization on less-structured data can yield extremely long lines, we normalize the score by dividing by the length of the string. This allows short but meaningful strings with types to appear at the top of the pile while pushing long strings that contain a few target types by chance to the bottom. Duplicates also cluster together and can be removed if desired.

4 Experiments and Results

Comparing our method with other systems is difficult for two reasons. Firstly, building LID test sets for long-tail languages is complicated in general, since the data available are often overly clean, skewed to a few reliable sources, and/or themselves collected with LID. Secondly, because our primary focus is document-level filtration, sentence-level datasets like FLORES-200 are likely to severely underestimate our performance. These problems are further aggravated when seeking to compare our method to some other list-based filtering approaches mentioned in Sections 2 and 3, as Scannell (2007) and Dent et al. (2024) further rely on the indexing of external search engines, while Caswell et al. (2020) and Ljubešić et al. (2007) use wordlists as secondary filtering to support multinomial classifiers, rather than first-pass models.

With these concerns in mind, we adopt a four-pronged approach. First, we first benchmark the

speed, recall, and false positive rate of Feature Mining and GlotLID. To get an idea of how much content we should expect to find in a snapshot, we then apply our document-level filtration system to subcorpora from three recent, large-scale filtered corpora that feature our target languages, namely MADLAD-400, GlotCC, and Fineweb-2. For all three corpora, we report our recall on the 'clean' portions of these corpora, as well as our ability to find usable data in the discard portion of the largest, Fineweb-2. After that, we test document-level filtering on a full Common Crawl snapshot with different wordlists at two thresholds. Lastly, we briefly explore the potential of type-based filtering for the second pass, and qualitatively estimate corpora quality.

For now, we focus on improving LID for Lesser Antillean Creole(s) *acf* and *gcf*,[8] French Guianese *gcr*, and Mauritian *mfe*, Seychellois *crs*, and Réunionese *rcf*. We do not include endangered varieties because the reference corpora do not address them. As Haitian Creole is already covered by many systems, we include it mainly as a point of reference.

4.1 Benchmarking Speed, Recall, and FPR

4.1.1 Benchmark Creation

For benchmarking the speed, recall, and false positive rate (FPR) of our method on clean data we create a dataset of 10,000 entries from two different languages: French and French Guianese Creole. By selecting data from Wikipedia, we are able to obtain pseudo-gold language labels without requiring an extensive annotation compaign. To reflect the typical class imbalance between target languages, 9800 articles in French and 200 in French Guianese Creole, taken from their respective Hugging Face Wikipedia datasets.[9] To reduce the chance of Creole data appearing in the French portion, which could potentially lead to true positives being mistaken for false positives, we skip articles that contain the word "créole".

4.1.2 Settings Compared

We compare our document-level Feature Mining approach to the full GlotLID model (Kargaran et al., 2023), which covers over 1600 linguistic varieties, as well as a version of GlotLID with the output space reduced to 3 languages (French, French Guianese Creole, English). We compare these two GlotLID models' speed and performance to those of document-level Feature Mining when used to mine data for a single language ('Min-1' setting) as well as for 3 languages (French Guianese, Lesser Antillean, Mauritian) simultaneously ('Min-3' setting), based on our wordlists for the respective languages. To convert line-level fastText predictions into document-level labels, we implement the Ungoliant document-scoring procedure (Abadji et al., 2022) adopted by Kargaran et al. (2024). We also explore both the impact of regex-based punctuation normalization, as mentioned in Section 3.1.1, and the impact of the value of the score threshold defined in Section 3.1.3.

4.1.3 Benchmark Results

We first compare the speed of GlotLID and document-level Feature Mining.[10] Document-level Feature Mining takes 0.46s for a single target language (Min-1) and achieves 79.0% recall with a base threshold of 5 disjunctive types per document (Table 3). Adding two languages (Min-3) only adds 0.05s on average compared to Min-1, and has no effect on recall or false positive rate (FPR) because the scores for each language are independent. Punctuation-aware tokenization (Min-1-P) nearly doubles the runtime for a single language and increases recall by 0.5% (a single document) but doubles the FPR. Using the full GlotLID model (Glot-Full) takes over 200 times longer than Min-1, but achieves a high 91% recall. Since GlotLID was trained in part on Wikipedia, this high performance is to be expected. After restricting the output space of GlotLID to 3 languages (Glot-3), speed improves, but remains 45 times slower. Recall rises to 100 % and FPR falls to 0.

Model	Glot-Full	Glot-3	Min-1	Min-3	Min-1-P
Mean Speed (s)	114.27	21.45	0.46	0.51	0.83
gcr Recall %	91.00	100.00	79.00	79.00	79.50
gcr FPR %	0.0	0.0	0.04	0.04	0.09

Table 3: Benchmark on the mixed Wikipedia corpus.

When we test Min-1 at different score thresholds (Table 4), we see a gradual decrease in recall as the required number of disjunctive types approaches 10. With a conservative threshold of 5 types, we are able to eliminate almost all of the French documents, while keeping nearly 80% of the Creole

[8]The ISO codes *acf* and *gcf* are listed as Saint Lucian and Guadeloupean. Other islands, particularly Martinique and Dominica, are inconsistently split between these labels.

[9]https://huggingface.co/datasets/wikimedia/wikipedia.

[10]We run these experiments in serial with an Intel i7-1370P.

articles. Beyond a threshold of 10, performance continues to decrease, but without additional benefits on the two-language dataset. Overall, we validate the wordlist approach of Caswell et al. (2020) for quickly eliminating high-resource cousins using choosable thresholds.

Threshold	1	3	5	10	15
True Positives	197.0	176.0	158.0	118.0	44.0
False Positives	1068.0	38.0	4.0	0.0	0.0
Recall %	98.5	88.0	79.00	59.0	22.0
FPR %	11.6	0.5	0.1	0.0	0.0

Table 4: The effect of threshold on recall and false positive rate (FPR) for French Guianese Creole

4.2 Relative Recall

As mentioned in the introduction to this section, we wanted to compare our with recent SOTA pipelines on real, noisy data. From this point on, we do not measure FPR because we do not have gold labels for the original raw web crawl data and creating such labels would be prohibitively costly.

4.2.1 Reference Corpora

MADLAD-400 (Kudugunta et al., 2024) is the oldest of the three corpora, with a cutoff date of August 1st, 2022. Built from all Common Crawl snapshots up to that point, it has sizeable subcorpora for five of our target Creoles: Lesser Antillean, Haitian, Mauritian, Réunionese, and Seychellois. Since the Transformer-based LID system used to create MADLAD-400 is not open-source, we are unable to directly compare its speed, recall or false positive rate to those of our model.

GlotCC (Kargaran et al., 2024) combines the Ungoliant pipeline with GlotLID, and covers the February/March 2024 snapshot and portions of the September/October and November/December 2023 ones. In addition to the five target languages covered by MADLAD-400, GlotLID and GlotCC also cover Guadeloupean and French Guianese Creoles. Although GlotCC is smaller than the other two, it also comes with details about speed.

Fineweb-2 (Penedo et al., 2025) uses GlotLID to cover our targets. Like MADLAD-400, it is built from all Common Crawl snapshots released prior to its cutoff date of April 2024 (96 in total). By using a high coverage model on several petabytes of raw data, Fineweb-2 has created what are, to our knowledge, the largest *open* web-scraped corpora for several FCs, and thus represents the state-of-the-art at the time of writing.

4.2.2 Relative Recall Results

In Table 5, we show that, with a threshold value of 5, recall is well over 95% for most of the 'clean' datasets, which shows that our method quickly finds documents that would score well on slower models. For GlotLID-acf, the corpus size was very small (6). We verified that the last document was noise, and thus our algorithm correctly excluded it. This issue was most pronounced with the FineWeb-2-mfe clean subcorpus, where most of the raw data appears to be noise of the "repetitive ngram" type, especially from commercial websites, and thus the should have been excluded the first time. For examples, see Appendix 6.3.

Corpus	acf	gcf	gcr	hat	crs	mfe	rcf
GlC	83.3	100.0	100.0	100.0	100.0	95.5	95.9
MAD	100.0	NA	NA	99.9	99.5	98.1	99.0
FW2	77.9	99.8	98.5	99.3	89.1	38.0	93.7

Table 5: Recall percentages on comparable corpora

We also applied our type-score filtering to the 'removed' subcorpora of Fineweb-2, which consist of data that received a target language label, but was of questionable quality. As shown in Table 6, the 'removed' piles range from roughly the same size as the 'clean' piles in the case for Réunionese Creole (rcf) to 66x bigger for Seychellois Creole (crs). When we filter the 'removed' documents, we find that even though passing documents are but a small percentage of the 'removed' data in several cases, they nearly double the Mauritian (mfe) and Haitian (hat) corpora, and are over 9 times more numerous than the filtered 'clean' documents for Lesser Antillean Creole (acf). Thus, beyond post-filtering, we can pre-filter noise too.

4.3 Filtering Common Crawl

We use the December 2024 (CC-MAIN-2024-51) Common Crawl snapshot, which contains 21 TB

Language	acf	gcf	gcr	hat	crs	mfe	rcf
Clean-raw	1.1	2.8	0.9	224.4	3.5	20.4	7.9
Rem-raw	109.0	10.9	5.6	4466.7	233.9	807.0	7.8
Clean-filt	0.9	2.8	0.9	222.8	3.1	7.8	7.4
Rem-filt	7.4	3.0	1.8	351.0	2.5	6.9	3.5
Total-filt	8.3	5.8	2.7	573.8	5.6	14.6	10.9

Table 6: Documents (thousands) in Fineweb-2 Corpora clean and noisy subcorpora before and after filtering

of raw data, to test Multilevel Feature Mining at scale. Early on, we found that parallel document-level sorting in Rust was nearly twice as fast as Python. For an efficiently parallelized first pass on the full 21 TB, we used the Rust implementation and configured the program to be run by a SLURM job scheduling system in order to achieve further parallelism by working on different divisions of raw data at the same time. In 4.3.1, we report the runtime on two kinds of node, Broadwell Xeon e5-2650 v4 processors and Broadwell Xeon e5-2695 v3/v4. Like GlotCC, we do not count the time required to transfer and decompress the data towards our runtime, as these are beyond the scope of the LID algorithm. However, we do include the time needed to parse WET files, which are aggregates of raw data from many web pages, into individual documents. This inflates the LID runtime somewhat, but our method is nevertheless many times faster than GlotLID, for which Kargaran et al. (2024) estimated that treating a similar snapshot would require 340 hours.

In addition to document-level filtering, we also rank each line as described in Section 3.2. This isolates quotes embedded in other languages and penalizes long strings of noise. To explore subsequent passes on smaller data, we returned to Python as absolute speed was less crucial after filtering out most of the initial data.

4.3.1 First Pass

Using the configuration described in Section 4.3, we ran the pipeline several times to measure the impact of different wordlists and thresholds (5 and 10). Overall, we were able to index 21 TB of raw data using 9 parallel jobs with 32 CPUs in 2 to 4.5 hours of wall time, with the exact runtime depending on the cluster. More specifically, on nodes with Broadwell Xeon e5-2650 v4 processors consistently finished their respective jobs in two hours, which translates to an average speed of roughly 1258 pages per CPU per second.[11] On nodes using other processors, including Broadwell Xeon e5-2695 v3/v4, were closer to the 4-hour mark, but it is also possible that external factors within the cluster affected the exact runtime. However, even when jobs were not simultaneous due to cluster conditions, the pre-filtering stage ran at least 50-100x faster than GlotCC's reported wall time of 340 hours (Kargaran et al., 2024). This is congruent with our results in Section 4.1 especially if we note that a considerable percentage of our actual runtime is merely reading the input data.

The choice of wordlist and threshold did not seem to impact runtime. However, there were appreciable differences in the sizes of the indexed corpora. As seen in Table 7, the Lesser Antillean (acf, gcf) and French Guianese (gcr) lists yield corpora 2-3x smaller than the Haitian and Indian Ocean lists, which is likely because the former use more diacritics.

Language	acf	gcf	gcr	hat	crs	mfe	rcf
5	158	93	116	265	212	260	233
10	13	12	15	34	31	39	46

Table 7: Indexed size (GB) by language and threshold

4.3.2 Second Pass

Once the documents and lines are ranked, the highest-scoring content (for document, from the mid teens upward to hundreds) is reliably in either the target language or a closely-related sister language. Medium scores are sometimes true positives, but a substantial amount of content is in a handful superficially similar distractor languages, such as Catalan, Roman Hindi, and English with erratic spacing. At the document level, low scores are often indicative of noise, but sometimes come from short quotes, either as standalone documents, or embedded in a longer work in another language. At the line level, however, low normalized scores are particularly indicative of noise.

Due to the sort operation, the second pass is run as a single job, which makes identifying a (higher and language specific) data loading threshold important for speed. Focusing on one language from the Americas (Lesser Antillean) and one language from the Indian Ocean (Mauritian), the second pass on the former takes less than an hour with a loading threshold of 10, while the latter needs a threshold of 14 for similar speed (compare with Table 7). We easily remove less-closely related languages at the document-level by using the WARC-identified language [12]. Languages removed this way include Swedish, Romanian, and Turkish. To remove data from closely-related FCs, namely French Guianese,

[11] 2.6×10^9 webpages / (9 jobs $\times$ 32 cpu/job $\times$ 2 hr $\times$ 3600 s/hr).

[12] Common Crawl WARC files are the base archival records. They include preliminary language identification using CLD2. The smaller, text-focused WET files contain these results (see https://commoncrawl.org/blog/august-2018-crawl-archive-now-available.)

Haitian, and Réunionese, we eliminate documents where they score higher than the target, and also filter by source. For example, the first pass for both corpora includes pages from French Guianese Wikipedia and a specific Réunionese newspaper, which can be removed based on the URL.

4.4 Qualitative Evaluation

Due to the complications mentioned in the beginning of this section, we focus on qualitative estimations of corpora quality and diversity. After the first pass, one of the authors familiar with the relevant languages manually examined the resulting document and line level corpora, along similar lines to the audits of Kreutzer et al. (2022). This provided insights which were taken into account for the second pass on two target languages: Lesser Antillean and Mauritian. These two then received a second round of qualitative evaluation.

A small portion of the removed-filtered data from Fineweb-2 consists of false positives in languages such as Hilgaynon and phonemically-spelled French. However, the majority are in the correct language or at the very least, the correct subgroup (Americas or Indian Ocean). In broad terms, the content found within each language corresponded with what Robinson et al. (2024) found through semi-manual collection. Bible translations were very prominent in the Lesser Antillean corpora, but many song lyrics where also found. French Guianese data was predictably mainly from Wikipedia. Large amounts of the Mauritian corpus came from a single prolific language activist, but religious texts, news, and music were also detectable. Réunionese data was similarly dominated by one local newspaper, but web forums and cultural content had a sizable presence. Seychellois data was particularly diverse; parliamentary reports were predictably well represented, but we also found full-length linguistic studies and the other genres already mentioned. Curiously, we found aligned Bible translations for unexpected languages, including Amharic, Arabic, Toba Batak, Biak, and Ghomálá'[13] in the Indian Ocean datasets.

For our Common Crawl filter, the range of content is similar to what is observed in Fineweb-2. Low scoring data is numerically dominant after the first pass, but simple visual inspection suffices to identify a cutoff point beyond which above which the signal-to-noise ratio increases dramatically. Ultimately, the exact subcorpora sizes depend on our threshold, but for both languages, we find several hundred reliable documents even at high (>30) thresholds. In the current experiments, the line-level, length-normalized cutoff is around 0.005. Appendix B demonstrates the Lesser Antillean gradient in more detail.

We will release our filtered versions of FineWeb-2, which significantly increase the amount of readily-usable web data for several languages, as well as the intermediate outputs of the first pass for all varieties and the aforementioned second pass datasets for further LID research.

5 Conclusion

We have introduced Multilevel Feature Mining as an efficient approach to web corpora creation for less commonly written languages. Taking advantage of the large class imbalance among varieties, we are able to eliminate the overwhelming majority of documents by counting how many distinctive types appear in the document. For French-based Creoles, such types can be identified quickly using whitespace tokenization. Rapid LID facilitates new kinds of exploration of crawled data, including but not limited to training language models.

6 Limitations

6.1 Harmful Content

As with any attempt to filter Common Crawl snapshots or similarly massive repositories of raw web data, we run the risk of encountering text which contains explicit and/or derogatory remarks regarding various personal attributes. In this work, we use simple blacklists to filter out multilingual machine-translated pornographic websites, as this requires essentially zero extra resources, but eliminates one of the main sources of noise affecting our target languages. We recognize that for the release of "off-the-shelf" corpora for training models, it would be highly beneficial to develop language-specific tools for identifying more subtle instances of harmful/unwanted content in our target languages.

For example, the Mauritian newspaper comment in Appendix 6.3 comes from a Mauritian complaining about local cleanliness, and contains a negative generalization about Mauritians. Similarly, the French social media comment comes from an online dispute and contains several insults. While there are reasons to exclude such content

[13] Amharic, Toba Batak and Biak, and Ghomálá' are spoken in Ethiopia, Indonesia and Cameroon, respectively.

when training certain kinds of models, they are clearly in a completely different category from mass-translated pornography, and nevertheless remain of interest for studies such as ours.

We hold that, although we are not yet in a position to implement fine-grained content filtering, the intermediate corpora created thus far using our mining approach will ultimately facilitate the necessary research to create such tools in the future.

6.2 Scaling Up

As mentioned in Section 4.3, it is common to report only the time spent on LID, ignoring the transfer and decompression of data. Thus far, we have take a similar approach. However, since our targeted feature mining approach runs several in such little time, one might wonder why we did not process multiple snapshots. Effectively, improving the efficiency of LID causes the entire pipeline for rarer varieties to no longer be bound by CPU-intensive inferences, but rather internet bandwidth.

More specifically, it took 19.25 hours to copy one snapshot (7.37TB compressed) from AWS storage (USA) to our compute cluster (Europe) using the Common Crawl Downloader with the suggested 10 parallel threads.[14] Although we were able to distribute the decompression across several nodes, it took an additional 8 hours of wall time to decompress the dataset. Yet, as we have shown in 4.3.1, we ideally only look at most of the data once (to throw it away), which can be accomplished in a few hours using resources of a similar scale. While copying the raw data is still useful for development purposes, such as debugging very rare conditions, it is likely necessary to process the data *in situ* to truly take advantage of the speed of the method.

6.3 Beyond French-based Creoles

Although the naive Bag-of-Types approach is very effective for varieties that use distinct orthographies, as is this case with the more standardized of the French-based Creoles, there are typologically similar languages for which a whitespace-type based approach is less effective, and there are two obvious reasons for this.

Because the wordlists provided by Caswell et al. (2020) were designed to be used as second-pass filters, some words are only or two characters long. In these cases, a wordlist-based method is particularly vulnerable to 'A N T S P E A K', where texts are broken up by excessive whitespaces. In this work, we mitigated this drawback by introducing a minimum character length when constructing our wordsets.

However, certain languages, notably including Nigerian Pidgin/Naijá, appear to have few distinct, yet common types. In the provided 'pcm' wordlist, many of the ostensibly disjunctive words are standard English words that deal with topics like crime and pop culture. While an obvious first step is to simply remove such words, Naijá and closely related varieties like Jamaican Patois show that we cannot complete dispense with syntax. With this in mind, we also an exploratory regex-based experiment to complement the main work. The preliminary analysis suggests that taking a small number of multiword expressions into account, such as *no dey*, can improve coverage for these languages. Thus, a generalizable way to cross whitespace boundaries without wasting time on superfluous n-grams would be a welcome improvement.

[14] https://github.com/commoncrawl/cc-downloader

Acknowledgements

This work was primarily funded by the Inria "Défi"-type project COLaF. This work was also partly funded by the last author's chair in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

We thank the anonymous reviewers as well as Laurie Burchell for providing feedback that has helped to refine this work.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An Optimized Pipeline for the Generation of a Very Large-Scale Multilingual Web Corpus. In *CMLC 2021 - 9th Workshop on Challenges in the Management of Large Corpora*.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Jan A. Botha, Emily Pitler, Ji Ma, Anton Bakalov, Alex Salcianu, David Weiss, Ryan McDonald, and Slav Petrov. 2017. Natural Language Processing with Small Feed-Forward Networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2879–2885, Copenhagen, Denmark. Association for Computational Linguistics.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An Open Dataset and Model for Language Identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, volume 161175, page 14. Ann Arbor, Michigan.

Koby Crammer and Gal Chechik. 2004. A needle in a haystack: Local one-class optimization. In *Twenty-First International Conference on Machine Learning - ICML '04*, page 26, Banff, Alberta, Canada. ACM Press.

Joaquim Ferreira Da Silva and Gabriel Pereira Lopes. 2006. Identification of document language is not yet a completely solved problem. In *2006 International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06)*, pages 212–212. IEEE.

Rasul Dent, Juliette Janes, Thibault Clerice, Pedro Ortiz Suarez, and Benoît Sagot. 2024. Molyé: A Corpus-based Approach to Language Contact in Colonial France. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 189–199, Miami, USA. Association for Computational Linguistics.

Google Research. 2025. Url-nlp: Fun-langid. https://github.com/google-research/url-nlp/tree/main/fun_langid.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ivan Habernal, Omnia Zayed, and Iryna Gurevych. 2016. C4Corpus: Multilingual Web-size corpus with free license. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 914–922.

Tommi Jauhiainen, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2024. *Automatic Language Identification in Texts*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas 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, Valencia, Spain. Association for Computational Linguistics.

Amir Hossein 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*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. GlotCC: An open broad-coverage commoncrawl corpus and pipeline for minority languages. *Advances in Neural Information Processing Systems*, 37:16983–17005.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 441–450, New York New York USA. ACM.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta

Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

Chaak-ming Lau, Mingfei Lau, and Ann Wai Huen To. 2024. The Extraction and Fine-grained Classification of Written Cantonese Materials through Linguistic Feature Detection. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 24–29, Torino, Italia. ELRA and ICCL.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, Anders Søgaard, and Johannes Bjerva. 2024. CreoleVal: Multilingual Multitask Benchmarks for Creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

Nikola Ljubešić, Nives Mikelić, and Damir Boras. 2007. Language Indentification: How to Distinguish Similar Languages? In *2007 29th International Conference on Information Technology Interfaces*, pages 541–546.

NLLB_Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *Preprint*, arXiv:2207.04672.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language. *Preprint*, arXiv:2506.20920.

John M. Prager. 1999. Linguini: Language Identification for Multilingual Documents. *Journal of Management Information Systems*, 16(3):71–101.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, and Naome Etori. 2024. Kreyòl-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.

Kevin P. Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora (WAC3-2007): Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval*, volume 4, page 5. Presses univ. de Louvain.

A True and False Positives from the "clean" Mauritian Fineweb-2.

Description of Source	Relevant Text Sample	Label	Reason
Mauritian Catholic Diocese	Depi 1995 Komite Diosezin Premie Fevrie organiz la mes. Par sa manier la ek boukou bann lot travay, legliz katolik pe akonpagnn bann desandan esklav dan simin ki zot pe fer dan repiblik Moris avec tou bann lezot organisasion kreol ki sakenn dan so manier pe aport so kontribision dan mem travay.	TP	Monolingual text with standardized spelling.
Mauritian Newspaper Comment	[*Mo* Problem] *Zot* jette tou *zot* saleté derrière *ene* bus-stop dans Vacoas *Lot* fois *la mo p* prend bus dans Vacoas. *Mo ti lor ene* bus-stop près dans centre. L'heure *mo* guette par derrière bus-stop *la*, *mo* gagne vomi. Et pourtant *ti ena* poubelle *lor* bus-stop la. Mais tou dimounes préfère jette *zot* saleté par derrière bus-stop la. *Kifer* Mauriciens malang *koumsa* ?	TP	Monolingual text where most content words use French or English spellings, but the syntax and many *function words* are still distinctly Mauritian.
Mauritian Newspaper Article	[FR] Depuis que l'affaire Kistnen a éclaté, "l'opposition utilise une stratégie pour semer la psychose à travers le pays", est d'avis le ministre de l'Energie, Joe Lesjongard. Ce dernier s'exprimait lors d'une conférence de presse axée sur le bilan 2020 du gouvernement. [MFE] *« Bann avoka sipoze ena bann rezerv alor ki ena enn lanket »*, [FR] a estimé le représentant du gouvernement. Et d'ajouter que [MFE] *« zot fer bann deklarasion ki degrad nou bann institision. Sa fer ditor nou pei »*.	TP	Code-switching between French in the main report and quotes in Mauritian Creole.
French Social Media	CA VOUS DERANGE CES PAS MOI SUR LES TOF OK MARRE DE CES CON KI SAVE KE CE MOKER ... BANDE CONARD ARETTER DE DEMANDER KAN CES KE JE SUCE BANDE DE NAZE DIT LEUX ENFACE	FP	French with nonstandard spelling, especially use of <k> for <qu> and <er> for <ez>; <bande> ressembles Mauritian plural marker
French Insurance Company	Assurance sante etudiant Villeneuve-les-Avignon.Comparatif mutuelle sante etudiant Villeneuve-les-Avignon.	FP	"Santé" [health] is spelled like Mauritian "sante" [health OR to sing]
French Bicycle Vendor	wheels mfg patte de derailleur 92 orbea patte de derailleur usinee cnc br 30 plus resistante qu une patte traditionnelle br br br br strong details strong br ul li vis fournis li ul br br strong pour cadre strong br br	FP	Noisy list of words with an n-gram collision:

B Illustrative examples from the line-level Lesser Antillean Creole (LAC) Corpus.

Norm.	Raw	Text	Description of Source	Remark
0.300	6	Sé nou ki ka pwan fè	Lyrics Site	LAC
0.227	5	An ba latè pa ni plézi	Martinican; Cultural	LAC
0.189	7	Pa janmen fè wè zétwal an ba kout san	Pan-FC; Cultural	LAC
0.153	4	nou - Dèyè bwa ki tini bwa	Political; Blog	LAC
0.115	3	Mo té linmé dé bèl moushwa	Lyrics; Personal	Louisiana
0.114	3	Si nous té pren tan pou nou té palé	Lyrics	LAC
0.103	4	Kiyé tanbouyè, pou woulé tan-la ba mwen	Dominica; Lyrics	LAC
0.094	3	Nou kontan zò vin asi sit-lasa !	Bible	Guianese
0.088	8	SA PA DLO POU MOUYÉ MWEN ! MAN KEY RIVÉ BATJÉ ANLÈ BATO-A POU ALÉ-VIRÉ ADAN TOUT KARAYIB-LA	Pan-FC; Cultural	LAC
0.081	11	Pou yon moun enpòté yon médikaman ki pa apwouvé oswa pa disponib ki nésésè pou sové lavi pou itilizasyon pèsonèl	Global; Medical; Saint-Lucia	LAC
0.071	3	Labitid lapa vantar Ki la di aou yinm pa !	Lyrics; Personal	Réunionese
0.027	7	[FR] Je suis persuadée que parmi les « carnavaliers » y'en a qui vraiment ne peuvent pas rester tranquilles et qui profitent de tout ça pour défiler en robe pété parce que [LAC] si yo pa fè sa yo ka senti yo kay mô! Genre vous pouvez pas rester chez vous cette année quoi!	Caribbean; News; Martinique	FR/LAC
0.0174	3	Dèpeu qu'ill a in ptcho d'quouai, ill s'achte dè souillers. Ill in a ben d'trop, mais y fè ren. Y 'avo ben trop fait malice qu'la pieuge li fasse mèttre sè souillers peus !	Blog; (Mainland) French Regional Culture	Franco-Provençal
0.0167	3	En 1794 bann révolutionnaire la aboli l'esclavage, l'esclavage té interdit par la loi : «Toute de moune lé libe». La envoye deux bougue té i appelle Baco ec Burnel, sanm 2.200 soldats pou allé abolir l'esclavage l'île Maurice ec la Réunion.	History; Indian Ocean	Réunionese
0.005	3	May tenant ako na nag rent since Jan 15, 2018 nag ka problema na kami sa knya noon pa kasi ng volunteer sya na sya na lang mag rerepair ng bahay kasi kailangan na daw [...]	Commercial; Phillipines	Filipino

Table 8: "Raw" indicates the word-type score."Norm." is normalized by the length of the string in characters.