

Automatically Detecting Code-Switching Between Nigerian Pidgin English and English

Godspower Ogbonda

gdg24xtn@bangor.ac.uk

Bangor University

William Teahan

Bangor University

Research Article

Keywords: Code-switching, Nigerian Pidgin English, Language identification, Compression-based models, Natural language processing, Low-resource languages

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9706554/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Automatically Detecting Code-Switching Between Nigerian Pidgin English and English

Godspower Ogbonda^{1*} and William Teahan¹

¹School of Computer Science and Electronic Engineering, Bangor University, Dean Street, Bangor, LL57 1UT, UK .

*Corresponding author(s). E-mail(s): gdg24xtn@bangor.ac.uk;
Contributing authors: w.j.teahan@bangor.ac.uk;

Abstract

This research explores the automatic identification of code-switching between Nigerian Pidgin English (NPE) and Standard English, a frequent linguistic practice in multilingual contexts. The absence of standardisation and the regional differences in NPE contribute to significant under-resourcing in natural language processing (NLP) research. To address this gap, we collected a multi-domain corpus comprising literary texts, forum discussions, and online media content. A manually verified ground truth dataset was subsequently created to support fine-grained evaluation. Multiple modelling paradigms were employed, including traditional machine learning classifiers such as Support Vector Machine, Random Forest, Logistic Regression, and Naïve Bayes, a Bidirectional LSTM (BiLSTM) deep learning model, and a novel application of compression-based models trained using the Prediction by Partial Matching (PPM) algorithm via the Tawa toolkit. The evaluation was performed using a character-level comparison between model-generated annotations and gold standard XML-tagged data, with accuracy, precision, recall, and F1-score used as primary evaluation metrics. The results reveal that compression-based PPM models offer competitive performance compared to supervised approaches, particularly in low-resource settings, and demonstrate significant computational efficiency. These findings highlight the effectiveness of compression-based methods in segmenting mixed-language text without extensive linguistic resources, and the study contributes a reusable annotated corpus to support future NLP research on Nigerian Pidgin English.

Keywords: Code-switching, Nigerian Pidgin English, Language identification, Compression-based models, Natural language processing, Low-resource languages

1 Introduction

Code-switching refers to the practice of alternating between two or more languages within a single conversation and is prevalent in multilingual contexts. It provides insight into cognitive processes, social interactions, and cultural identities (Myers-Scotton, 1993; Poplack, 1980). In Nigeria, where over 500 indigenous languages coexist alongside English, Nigerian Pidgin English (NPE) plays a crucial role as a lingua franca and helps people communicate across diverse ethnic and linguistic communities. (Eberhard et al., 2021). However, the linguistic variability and lack of standardisation inherent to NPE pose significant challenges for natural language processing (NLP) applications, particularly in tasks involving code-switching detection.

Developing effective computational models for code-switching in NPE is complicated by the scarcity of sufficiently annotated corpora that reflect the language’s natural variability and contextual versatility (Lin et al., 2023; Winata et al., 2021). To address these limitations, this research involves creating a comprehensively annotated corpus and employing machine-learning and compression-based models to accurately detect and segment code-switched text.

Our primary objective is to develop resources for NPE, including a comprehensive corpus and a code-switching detection model using the Tawa toolkit (Teahan & Harper, 2003). Specifically, the research evaluates the performance of compression-based models employing the Prediction by Partial Matching (PPM) algorithm (Tarmom et al., 2020; Teahan & Harper, 2003). We hypothesise that these models can achieve performance comparable to supervised approaches in segmenting mixed-language text without relying on extensive linguistic datasets. An experimental evaluation is performed to assess the language identification and segmentation accuracy between NPE and standard English. This evaluation involves multiple modelling paradigms, including traditional machine learning classifiers, compression-based models, and deep learning approaches, which ensures comprehensive analysis across linguistic dimensions in the compiled corpus.

The remainder of this paper is organised as follows: Section 2 provides a literature review on code-switching, covering theoretical foundations, computational methods, and challenges associated with Nigerian Pidgin English (NPE) and other creole languages. Section 3 describes our methodological approach, including corpus compilation, preprocessing steps, feature extraction techniques, and details on model training and evaluation procedures. Section 4 presents and analyses our experimental results; as a result, the reader gains insight into model performance, classification patterns, and linguistic characteristics identified within the corpus. Finally, Section 5 summarises the key findings of this study and proposes directions for future research so that the potential for further advancements in NLP for multilingual and low-resource language contexts becomes evident.

2 Related works

Research on code-switching traditionally centres on theoretical frameworks explaining linguistic, cognitive, and sociocultural foundations. Sociolinguistic studies have extensively explored how speaker identity, discourse functions, and community norms

influence language-switching behaviour (Myers-Scotton, 1993, 2002; Poplack, 1980). Cognitive perspectives highlight models such as the Matrix Language Frame (MLF) model (Myers-Scotton, 2002), Activation Threshold Hypothesis (Green, 1998), and the Inhibitory Control Model (Green & Abutalebi, 2013); these models shed light on the cognitive control mechanisms that enable language selection and switching in bilingual contexts.

Empirical research has illustrated the complexity of detecting and modelling code-switching, particularly in computational linguistics. Early computational models, primarily heuristic or rule-based, struggled with linguistic variability inherent in code-switched contexts, especially in less standardised languages (Poplack & Walker, 2003; Winata et al., 2022). Recent advancements in supervised machine learning and deep learning approaches have substantially enhanced language identification and segmentation accuracy in multilingual texts. For instance, by demonstrating significant improvements using deep neural networks and acoustic-based methods, Adel et al. (Adel et al., 2013) and Rivera (Rivera, 2019) show how contextual features and sequential dependencies are crucial in language-switching detection.

Despite these advances, computational studies that address low-resource languages such as Nigerian Pidgin English (NPE) remain underdeveloped (Hedderich et al., 2021). NPE is characterised by significant regional variability, extensive integration with Standard English, and the absence of standard orthographic conventions, all of which create substantial challenges for traditional supervised NLP approaches based on large annotated datasets (Marchal et al., 2021; Winata et al., 2021). Moreover, computational approaches specifically designed for creole languages and their distinctive linguistic features are scarce. As a consequence, there is an urgent need for alternative methodologies capable of handling sparse linguistic data.

Compression-based language models like Prediction by Partial Matching (PPM) provide promising alternatives for processing under-resourced languages. These models leverage probabilistic and entropy-based techniques to effectively capture language patterns without extensive annotated resources (Tarmom et al., 2020; Teahan & Harper, 2003). Samanta et al. (Samanta et al., 2019) and Piergallini et al. (Piergallini et al., 2016) have further shown how combining advanced computational techniques, such as deep generative models and sequence modelling methods, can significantly enhance the performance of NLP systems in identifying and segmenting multilingual texts.

A crucial barrier limiting NLP progress in languages like NPE is the scarcity of comprehensive annotated corpora capturing authentic code-switching instances (Deuchar, 2020; Sitaram et al., 2019). Addressing this gap involves developing datasets reflective of natural multilingual communication and integrating diverse methodologies for annotation, such as manual, crowd-sourced, and semi-automated processes (Agüero-Torales et al., 2023; Deuchar, 2020; Muhammad et al., 2022). This study aims to address these challenges by compiling a diverse corpus of NPE and employing both machine learning and compression-based models.

3 Methodology

An experimental evaluation methodology is employed to detect language identification and code-switching between Nigerian Pidgin English (NPE) and Standard English. The evaluation of machine learning models, compression-based approaches, and deep learning methods provides a comprehensive analysis of linguistic patterns across a compiled corpus of NPE and English.

3.1 Corpus Compilation and Structure

The corpus used in this study was compiled from a variety of sources to ensure diverse linguistic representations of both formal and informal usage of Nigerian Pidgin English (NPE) and Standard English. It consists of three datasets: a Pidgin corpus, an English corpus, and a Mixed corpus containing both languages.

The Mixed corpus, which serves as the foundation for this study, was curated to capture authentic code-switching between NPE and English. It was compiled using content extracted from three primary sources: 30 literary books, the Nairaland online discussion forum, and articles from BBC Pidgin. Table 1 provides a structured overview of these sources, which lists the genre, number of entries, and purpose of each dataset.

Table 1 Overview of Data Sources Used to Construct the Mixed Corpus

Source	Entries	Domain	Description
Literary Books	30 books	Fiction, Drama	Nigerian and African literary texts featuring conversational and narrative code-switching. Authors include Chinua Achebe, Wole Soyinka, Buchi Emecheta, Ngugi wa Thiong’o, among others. Texts were obtained from publicly accessible archives.
Nairaland Forum	~5,000 comments	Informal Discussion	Forum threads from https://www.nairaland.com/ focused on politics, technology, relationships, and social commentary. Data reflects real-world, user-generated dialogue in English and NPE.
BBC Pidgin News	~300 articles	News and Media	News articles scraped from https://www.bbc.com/pidgin , including sections like Nigeria, Africa, Sport, and Entertainment. Content provides a structured, formal use of NPE.

To enhance transparency and reproducibility, selected literary texts are accessible via open digital repositories such as Project Gutenberg (<https://www.gutenberg.org>), Internet Archive (<https://archive.org>), and Open Library (<https://openlibrary.org>). Titles like *The Palm-Wine Drinkard* (Tutuola), *Efuru* (Nwapa), and *A Man of the People* (Achebe) are among those publicly available through these platforms.

The literary texts were processed using a rule-based sampling method that targeted direct quotations, where code-switching or Pidgin expressions were most likely to appear. This method incorporated Pidgin markers: lexical indicators such as *na*, *dey*, *abi*, *wetin*, *oga*, to flag likely Pidgin content. Upon identifying a relevant phrase, the extraction was extended to a full stop to preserve sentence coherence. Importantly, the aim was not to isolate pure Pidgin expressions but to capture genuine instances of code-switching. The inclusion criteria allowed for any sentence that began with or contained Pidgin, even if it concluded with English.

Web data was collected using custom-built scrapers to extract informal language from Nairaland and BBC Pidgin articles. These online platforms contributed significantly to the real-world, conversational nature of the corpus. The English corpus was derived from English-only segments of this web-scraped material and the literary books. Meanwhile, the Pidgin corpus was constructed by isolating clearly identifiable Pidgin content from the Mixed dataset. Ground truth datasets used for evaluation were also derived from these same sources, but the selected samples were excluded from the Mixed corpus and, therefore, not present in the derived Pidgin or English corpora. This ensured an unbiased evaluation.

The annotation of the Mixed dataset was carried out through a semi-automated pipeline. First, a language segmentation prompt using OpenAI’s GPT-3.5 Turbo model was applied to automatically insert XML-style tags indicating language boundaries. This automated output was then reviewed by human annotators using a custom-built annotation interface. OpenAI’s GPT-3.5 Turbo was used as an AI-assisted tool in the initial annotation phase only. All outputs were manually reviewed and verified by human annotators, and the authors take full responsibility for the final annotations. Three native speakers of Nigerian Pidgin participated in the annotation process. Two of them had formal backgrounds in natural language processing (NLP) and experience with tools like NLTK. Initial tagging disagreements were resolved through a three-stage adjudication process. First, the two primary annotators independently reviewed the samples. Disagreements were logged, and a third expert annotator, also a native speaker, examined only the conflicting cases. Final annotations were determined by majority agreement or third-party adjudication.

A comprehensive breakdown of the composition of the corpus is presented in Table 2, with a focus on the number of sentences, tokens of words, characters, unique word types, and the average length of sentences.

These statistics highlight the linguistic diversity and size of the corpora used for training and evaluation. The Mixed dataset, being the most extensive, played a central role in model development. However, the benchmark for evaluating classification and code-switching detection performance was the manually annotated ground truth dataset.

Table 2 Corpus Statistics Summary

Corpus	Sentences	Words	Characters	Unique Words	Avg. Word Length
Pidgin	6,157	110,608	590,267	10,668	4.21
English	5,746	97,100	571,746	12,586	4.78
Mixed	13,699	263,869	1,573,812	18,558	4.70

To further illustrate how language boundaries are represented in the dataset, a sample of code-switching annotation is shown below:

```
<pidgin>How mama be today? You no sabi book but you sabi plenty thing wey
pass book,</pidgin> <English>my dear girl, what a Waste of effort.</English>
```

This sample comes from the Mixed dataset, where segments were tagged to indicate a switch from Pidgin to English. These annotated samples were essential in training models to detect and segment multilingual content at a fine-grained level.

3.1.1 Comparison with Related Corpora

To contextualise the scale and scope of the corpus developed in this study, Table 3 presents a comparison with existing corpora that address code-switching or low-resource Nigerian Pidgin English (NPE) in related NLP tasks. The corpora selected for comparison represent the closest relevant prior work cited in this paper, spanning Arabic dialect code-switching, Swahili and English language identification, NPE sentiment analysis, NPE discourse annotation, and NPE adaptive training.

Table 3 Comparison of the Present Corpus with Related Code-Switching and NPE Corpora

Study	Language Pair	Size	Domain
This study	NPE and English	263,869 tokens	Literary, forum, news
(Tarmom et al., 2020)	Arabic and English	218,149 tokens	Social media
(Piergallini et al., 2016)	Swahili and English	N/A	Social media
(Muhammad et al., 2022)	NPE and English	14,000 tweets	Social media
(Marchal et al., 2021)	NPE and English	N/A	Spoken (Treebank)
(Lin et al., 2023)	NPE and English	176,843 sentences	News, misc.

As shown in Table 3, the corpus developed here is distinctive in several respects. Unlike the Arabic dialect corpus of (Tarmom et al., 2020), which focuses on a single social media platform, or the sentiment-labelled Twitter data of (Muhammad et al., 2022), the present corpus spans multiple domains: literary, forum, and news, thereby capturing a broader stylistic and register range of NPE usage. The Swahili and English dataset of (Piergallini et al., 2016) is a methodological parallel in its use of word-level annotation, but no token count is reported and the data is not publicly available. While (Lin et al., 2023) assembled a substantially larger parallel corpus, their focus is machine translation rather than code-switching detection, and annotation is primarily automatic. The discourse-annotated Naija Treebank-derived corpus of (Marchal et al., 2021) addresses a complementary linguistic dimension but

does not target code-switching segmentation directly. In contrast, this study contributes the first multi-domain, manually verified, XML-annotated corpus specifically designed for code-switching detection between NPE and Standard English, with a ground truth dataset withheld for evaluation. The manually annotated ground truth dataset developed in this study is publicly available via Zenodo at <https://doi.org/10.5281/zenodo.20117988>.

3.2 Preprocessing and Feature Extraction

Preprocessing and feature extraction were critical steps in preparing the dataset for classification. The preprocessing pipeline involved a series of operations aimed at cleaning and standardising the raw text data. Initially, Unicode normalisation was performed to correct corrupted or inconsistent characters, ensuring textual integrity across all corpora. This was followed by the removal of extraneous content such as URLs, usernames, and special symbols, allowing the models to focus exclusively on linguistically relevant tokens. The text was then converted to lowercase, and white-space inconsistencies were resolved to create a uniform input structure. Tokenisation was applied at both the word and character levels to accommodate the varying requirements of different models. Feature extraction techniques were tailored to the classification architecture: traditional machine learning models utilised character-based n-grams ranging from 1 to 8, while deep learning models relied on word embeddings to capture semantic relationships. For the compression-based approach, specifically the Prediction by Partial Matching (PPM) model, character-level PPMD models of orders 2 to 8 were applied, and features were derived through sequential character probability distributions, enabling the system to model language patterns based on contextual dependencies.

3.3 Model Training and Implementation

This study employed a multi-paradigm approach to model development by incorporating traditional machine learning, deep learning, and compression-based techniques. For the traditional models, four classifiers: Support Vector Machine (SVM), Random Forest (RF), Multinomial Naive Bayes (NB), and Logistic Regression (LR) were each trained independently using character-level n-gram features ranging from 1 to 8. To enable comparative analysis across different linguistic granularities, an additional set of models was trained using word-level token representations. All models were developed using a unified Python script that managed tokenisation, feature extraction, and training procedures. Input data from the Mixed dataset was vectorised using scikit-learn's `CountVectorizer`, with n-gram ranges specified per configuration, from 1 to 8 for character-based models and unigrams for word-based models. Each model type and feature representation was treated as a distinct training run, with all models trained from scratch on the full dataset. Instead of extensive hyperparameter tuning, widely adopted default settings were used uniformly across classifiers to maintain consistency and comparability. Each trained configuration produced a standalone markup output for evaluation.

The compression-based models were developed using the Tawa toolkit, which implements Prediction by Partial Matching (PPM), specially using the PPMD algorithm (Tarmom et al., 2020). Separate models for English and Pidgin were trained across PPMD orders 2 to 8 using the `train` application from the toolkit. Each trained model pair (one for Pidgin and one for English) was used collectively in evaluation via the toolkit’s `markup` function; the detailed evaluation procedure is explicitly described in Section 3.4. The use of character-level compression probability allowed these models to detect and classify language segments based on contextual statistical patterns. This approach provides a novel and data-efficient alternative to traditional supervised machine learning methods.

Additionally, a Bidirectional LSTM (BiLSTM) model was implemented using TensorFlow and Keras for sequence level classification. Input data consisted of tokenised and padded word sequences derived from the Mixed dataset, using a custom tokeniser with a vocabulary size capped at 10,000. The network architecture comprised two stacked BiLSTM layers, each followed by dropout regularisation to reduce overfitting, and a final dense layer activated by a sigmoid function to perform binary classification. Word embeddings were learned during training rather than being preloaded. The model was compiled using the Adam optimiser and trained with a class-weighted binary cross-entropy loss to address label imbalance. Training was conducted over multiple epochs with performance monitored on a validation split to ensure generalisation. This configuration enabled the model to effectively capture contextual dependencies between tokens and adapt to the intricacies of language switching across sequences.

3.4 Evaluation Procedure

All evaluation processes in this study were conducted using a manually annotated ground truth dataset, which was extracted from the broader corpus during data preparation and withheld from training. This ground truth resource consisted of sentences tagged with XML markers to indicate regions of code-switching between Nigerian Pidgin English and Standard English. Although the annotations were applied at the sentence level, each model’s output was converted into character-level labels **P** for Pidgin and **E** for English to enable fine-grained comparison. These reference labels served as the benchmark for assessing language segmentation accuracy across all models. The compression-based approach utilised the Prediction by Partial Matching (PPM) model, applying character-level PPMD models of orders 2 to 8. Features were derived through sequential character probability distributions, allowing the system to capture language patterns based on contextual dependencies. A dedicated processing script controlled this markup process, supporting various configurations including classifier type, token representation (word or character), and n-gram size. The second stage involved a custom evaluation script that aligned the model-generated tags with the gold-standard annotations. This alignment operated at the character level and computed metrics such as accuracy, precision, recall, and F1-score, along with visual outputs like confusion matrices and performance bar charts to aid in model comparison. Compression-based model annotation of the untagged ground truth was performed using the Tawa toolkit’s built-in `markup` function, which simultaneously

applied Pidgin and English PPM models based on statistical likelihoods. A configuration file defined the model-to-language mappings, which enabled automatic selection during the tagging process. The evaluation of these outputs, however, was performed using a separate custom script tailored for the compression-based setup. This script converted tagged segments into character-level labels, conducted alignment with the gold-standard tags, and computed a similar suite of evaluation metrics. Sentence structure from the original text remained intact across pipelines, enabling detailed character-level analysis of segmentation accuracy and classification performance. This consistent evaluation approach allowed for meaningful comparison across model types and representation strategies.

3.5 Challenges and Limitations

Several methodological and linguistic challenges arose during the project, shaping both the model development process and the interpretation of results. A primary constraint was the scarcity of large-scale, high-quality annotated datasets for Nigerian Pidgin English (NPE). Although a mixed dataset was constructed through careful extraction from literary texts, web forums, and news articles, the annotation process, particularly for identifying code-switching boundaries, was highly labour-intensive. While GPT-assisted tagging accelerated the initial labelling phase, considerable manual verification was required to correct errors and inconsistencies. The use of a three-annotator adjudication framework improved annotation reliability but also highlighted the inherent subjectivity involved in labelling mixed-language data.

Dialectal variation posed an additional challenge. NPE exhibits considerable regional and stylistic diversity, leading to inconsistencies in lexical usage, sentence structure, and orthographic conventions. This variation made it difficult for models to generalise effectively, especially when exposed to previously unseen linguistic patterns. Furthermore, code-switching itself remains a complex phenomenon to model, particularly within informal discourse where language boundaries are fluid and often context-dependent. Despite these limitations, the methodology developed here, which combines traditional supervised learning, compression-based models, and deep learning within a unified evaluation pipeline, provides a good reference for future work in low-resource multilingual classification and segmentation.

4 Results and Discussion

This section presents the experimental findings of our classification models applied to Nigerian Pidgin and English text. The results are evaluated across multiple perspectives, such as model performance, classification trends, and linguistic insights derived from corpus analysis. The discussion is structured to ensure logical progression, and each subsection expands on key aspects of the study.

4.1 Experimental Setup and Evaluation Procedure

The classification models were trained and evaluated using three datasets: a mixed dataset containing both Nigerian Pidgin and English text, and two separate datasets

for Pidgin and English. These datasets were compiled from multiple sources, with the mixed dataset originally tagged using XML labels to differentiate between the languages. This tagged dataset was then used to derive the separate monolingual datasets. For this task, five machine learning models: Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), and a BiLSTM model were trained from scratch, alongside multiple order-based PPM models implemented using the Tawa toolkit. A Ground truth dataset was used to evaluate classification performance across both the machine learning models and the PPM models. Each model was tasked with marking up an untagged version of the ground truth, with PPM models utilising the Tawa toolkit’s markup function, while the other models used a custom markup script.

The marked-up outputs were compared against the original ground truth data using a confusion matrix and a character-by-character evaluation approach. Accuracy, precision, recall, and F1-score were used as performance metrics, with the confusion matrix providing insights into classification trends and errors. Each N-gram model (1–8) and each PPM model order was trained separately, producing independent marked-up outputs for this task.

4.2 Model Performance and Classification Trends

The classification models were evaluated based on accuracy, F1-score, precision, and recall, with a particular focus on how different models handled English-Pidgin classification. The impact of N-gram size, classifier type, and corpus characteristics was also assessed. Table 4 summarises the overall classification performance across different models and feature representations.

The N-gram Random Forest (RF) model (3-gram and 5-gram) achieved the highest accuracy (0.89), significantly outperforming other models. Logistic Regression models showed a steady improvement as the N-gram size increased, peaking at 6-grams (0.87 accuracy). However, SVM models failed to scale effectively, stagnating at 0.43 accuracy across all N-gram sizes. PPM-based models, especially Order 2 and Order 5, demonstrated excellent performance, which resulted in strong recall and balanced accuracy and highlights their effectiveness in Pidgin classification.

Table 4 Comprehensive Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
RF (3-gram)	0.89	0.88	0.90	0.89
RF (5-gram)	0.89	0.88	0.90	0.89
LR (6-gram)	0.87	0.86	0.88	0.87
NB (6-gram)	0.85	0.84	0.86	0.85
BiLSTM	0.78	0.75	0.80	0.77
SVM (any N)	0.43	0.35	0.29	0.30
PPM (Order 2)	0.89	0.89	0.89	0.89
PPM (Order 5)	0.88	0.88	0.88	0.88

4.3 The Effect of N-gram Size on Model Performance

The role of N-gram size in classification was further explored by analysing model performance across 1-gram to 8-gram settings. Results in Table 5 illustrate a clear pattern: while larger N-grams (3–6) improve model effectiveness, there is a point of diminishing returns beyond 6-grams, particularly in Logistic Regression and Naïve Bayes.

Table 5 Effect of N-gram Size and Classifier on Performance (Accuracy)

Classifier	1g	2g	3g	4g	5g	6g	7g	8g
Random Forest	0.78	0.89	0.89	0.89	0.89	0.89	0.43	0.61
Logistic Reg.	0.61	0.69	0.76	0.80	0.84	0.87	0.43	0.43
Naïve Bayes	0.60	0.68	0.74	0.78	0.82	0.85	0.43	0.43
SVM	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43

The word-level models, particularly Random Forest at the word level (Word RF), exhibited different trends. While they performed competently in Pidgin classification (0.62 Pidgin F1-score), they struggled significantly in English classification, showing lower precision (0.34 English F1-score). This suggests that Pidgin’s unique lexical markers aid classification, while English requires deeper context for accurate detection.

4.4 The Role of Deep Learning and Compression-Based Models

4.4.1 BiLSTM Model Performance

The BiLSTM model, despite its ability to capture sequential dependencies, performed below expectations in this classification task. Unlike the Random Forest and PPM models, which leveraged explicit N-gram features and probabilistic patterns, the BiLSTM model relied solely on learned representations. The lower recall in Pidgin classification (0.70) and English (0.81) suggests that the model struggled with limited contextual data, requiring more training data or additional fine-tuning. Additionally, since Pidgin often follows a syntactically flexible structure, BiLSTM may not have effectively captured word dependencies compared to structured rule-based classifiers like PPMD Order 5.

4.4.2 PPMD Model Strengths and Efficiency

The PPM Order 2 and Order 5 models demonstrated exceptional recall performance, particularly for Pidgin text. These models rely on compression-based prediction, making them highly sensitive to the structure of repeated patterns. Unlike deep learning models that require extensive labelled data, PPM models excel in low-data environments, thereby enhancing their applicability for low-resource language classification tasks.

Beyond their strong classification performance, the PPM models exhibit remarkable computational efficiency. The training times are notably short, and the duration

required for the markup process, here used as an indicator of testing or inference performance, is minimal. In our evaluations, we refer to this phase as the *Markup (Testing) Time*, which reflects the speed at which the model can annotate unseen data.

Table 6 summarises the average training and markup times in different PPM orders. For instance, the Order 2 models require approximately 1.46 seconds for training and 3.55 seconds for markup, while the Order 5 models take around 1.41 seconds for training and 4.00 seconds for testing. These figures not only attest to the models’ efficiency but also illustrate their suitability for applications where rapid deployment and low computational overhead are critical.

Table 6 Timing Comparison for PPM Models

PPM Order	Training Time (s)	Testing Time (s)
2	1.46	3.55
3	1.14	3.40
4	1.27	3.80
5	1.41	4.00
6	1.39	4.10
7	1.34	4.20
8	1.51	4.30

These timing results demonstrate that PPM models not only achieve competitive classification but also provide significant computational efficiency. Their rapid training and testing capabilities enable quick iteration and deployment, a critical advantage for real-time language identification in resource-constrained settings. Overall, the performance and efficiency of PPM-based approaches make them a solid and pragmatic solution for low-resource multilingual NLP tasks.

4.5 Error Analysis and Confusion Matrix Insights

To further assess model misclassification patterns, we examine the confusion matrices of the top-performing models. Tables 7, 8, and 9 present detailed overviews of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for English and Pidgin across the best classifiers.

Table 7 Confusion Matrix for RF (3-gram)

	Predicted English	Predicted Pidgin
Actual English	TP = 27,257	FN = 2,714
Actual Pidgin	FP = 2,986	TN = 19,732

The 3-gram Random Forest model (Table 7) correctly classified 27,257 English instances, misclassifying only 2,986 cases as Pidgin. It accurately identified 19,732 instances of Pidgin, which shows its strong ability to differentiate between the two languages.

Table 8 Confusion Matrix for PPM Order 2

	Predicted English	Predicted Pidgin
Actual English	TP = 17,246	FN = 2,807
Actual Pidgin	FP = 1,019	TN = 13,429

The PPM Order 2 model (Table 8) exhibited better recall in Pidgin classification, correctly classifying approximately 93% of Pidgin instances, although at the cost of a slightly higher false positive rate (1,019 cases misclassified as English).

Table 9 Confusion Matrix for PPM Order 5

	Predicted English	Predicted Pidgin
Actual English	TP = 16,716	FN = 3,337
Actual Pidgin	FP = 663	TN = 13,785

The PPM Order 5 model (Table 9) showed a similar pattern, accurately identifying 13,785 Pidgin cases, with even fewer false positives (663 cases), but with a slightly higher rate of English misclassification (3,337 cases incorrectly labelled as Pidgin).

4.6 Linguistic Characteristics and Corpus Analysis

To provide further insight into language-specific classification, we analysed the dataset’s linguistic properties using Zipf’s Law (Moreno-Sánchez et al., 2016), Heaps’ Law (Lü et al., 2013), and word length distributions.

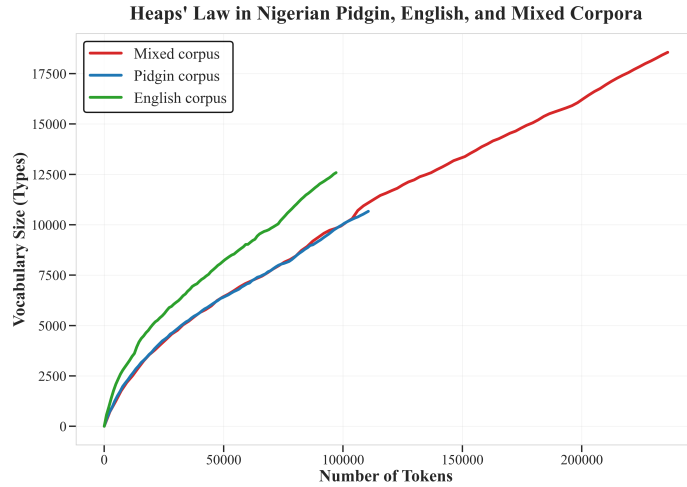


Fig. 1 Zipf’s Law comparison between English, Pidgin, and the Mixed Corpus, illustrating word frequency distributions and lexical compression tendencies.

The distribution in Figure 1 confirms that Pidgin exhibits a more compressed vocabulary compared to English, where a few high-frequency words dominate its lexicon. This aligns with linguistic expectations, as Pidgin relies on a constrained core vocabulary, whereas English possesses a broader range of lexical diversity.

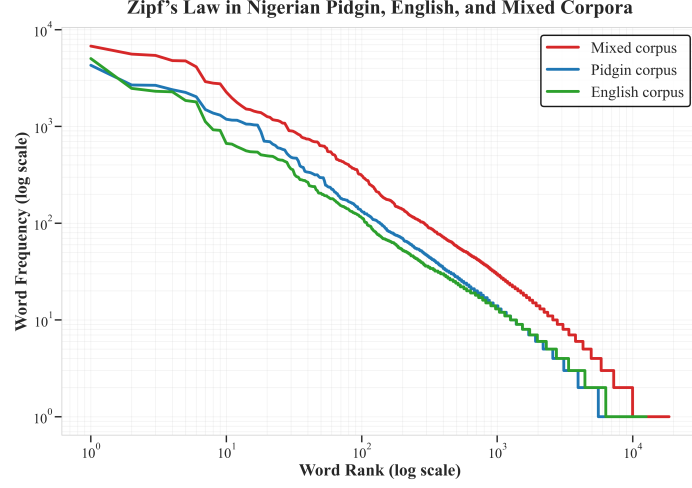


Fig. 2 Heaps' Law representation of vocabulary growth in English, Pidgin, and Mixed corpora, demonstrating lexical diversity trends.

Heaps' Law analysis (Figure 2) further supports these observations. The slower vocabulary growth in Pidgin suggests that its lexical diversity is more constrained than that of English. This characteristic likely influenced model performance, particularly in distinguishing language-specific word distributions.

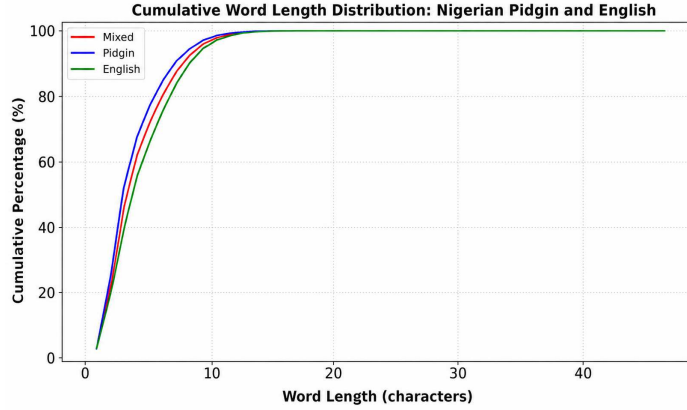


Fig. 3 Cumulative word length distribution across Nigerian Pidgin, English, and Mixed Corpora, highlighting linguistic complexity and aiding error analysis.

Figure 3 presents the cumulative word length distribution across the three corpora. The results indicate that Pidgin words tend to be shorter than their English counterparts, contributing to the misclassification trends observed in some models, particularly those relying on character-based representations.

5 Conclusion and Future Work

This study presents a novel approach to language identification and code-switching detection between Nigerian Pidgin English (NPE) and Standard English by integrating corpus construction, multi-paradigm model development, and character-level evaluation within a unified framework. Through the creation of a custom-annotated mixed-language dataset sourced from literary texts, forums, and news content, alongside the training of traditional, deep learning, and compression-based models, this research addresses the core challenges of processing under-resourced and linguistically fluid languages. The inclusion of a separate manually verified ground truth dataset further enabled robust benchmarking and fine-grained model comparison. Results indicate that compression-based models, particularly those built using the Prediction by Partial Matching (PPM) algorithm, offer competitive alternatives to conventional classifiers, especially in scenarios with limited annotated data. Similarly, the deployment of a BiLSTM architecture and multi-scale n-gram-based classifiers allowed for a broad exploration of lexical and sublexical features. Despite limitations posed by dialectal variation, annotation complexity, and limited language resources, the methodology developed here provides practical insights into low-resource multilingual processing and segmentation.

Looking ahead, future work will benefit from expanding the corpus to include more dialectal and regional varieties of NPE. Scalable annotation techniques, such as semi-automated labelling and community-driven crowdsourcing, can enhance dataset diversity and reduce manual overhead. In addition, exploring advanced neural architectures including transformer-based models, and developing hybrid techniques that combine the statistical robustness of compression methods with the contextual depth of deep learning, may lead to further improvements in classification accuracy. Lastly, applying these models to downstream tasks like speech recognition, conversational AI, and educational tools could support real-world applications while contributing to the broader goals of inclusive and equitable language technology development.

Declarations

Funding

No funds, grants, or other support was received for conducting this study.

Competing Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Author Contributions

G. Ogbonda was responsible for corpus compilation, annotation pipeline development, model implementation, experimental evaluation, and drafting of the manuscript. W. Teahan contributed to the conceptualisation of the research framework, provided supervision, and critically revised the manuscript. All authors read and approved the final version.

Data Availability

The manually annotated ground truth dataset developed in this study is publicly available via Zenodo at <https://doi.org/10.5281/zenodo.20117988> (Ogbonda & Teahan, 2026). Full distribution of the Mixed corpus is not possible due to copyright restrictions on the literary source texts used during corpus construction. Data collection and annotation scripts are available at <https://github.com/Don-Gp/nlp-pidgin-code-switching-full> to support reproducibility of the web-scraped corpus components.

References

- Adel, H., Vu, N.T., Kraus, F., Schlippe, T., Li, H., Schultz, T. (2013). Recurrent Neural Network Language Modeling for Code-Switching Conversational Speech. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8411–8415). Retrieved from <https://doi.org/10.1109/ICASSP.2013.6639306>
- Agüero-Torales, M.M., López-Herrera, A.G., Vilares, D. (2023). Multidimensional Affective Analysis for Low-Resource Languages: A Use Case with Guarani-Spanish Code-Switching Language. *Cognitive Computation*, 15, 1391–1406, <https://doi.org/10.1007/s12559-023-10165-0>
- Deuchar, M. (2020). Code-Switching in Linguistics: A Position Paper. *Languages*, 5(2), 22, <https://doi.org/10.3390/languages5020022>
- Eberhard, D.M., Simons, G.F., Fennig, C.D. (2021). *Ethnologue: Languages of the World* (24th ed.). Dallas, Texas: SIL International. Retrieved from <https://www.sil.org/resources/archives/89573>
- Green, D.W. (1998). Mental Control of the Bilingual Lexico-Semantic System. *Bilingualism: Language and Cognition*, 1(2), 67–81, <https://doi.org/https://doi.org/10.1017/S1366728998000133>
- Green, D.W., & Abutalebi, J. (2013). Language Control in Bilinguals: The Adaptive Control Hypothesis. *Journal of Cognitive Psychology*, 25(5), 515–530, <https://doi.org/https://doi.org/10.1080/20445911.2013.796377>

- Hedderich, M.A., Lange, L., Adel, H., Strötgen, J., Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2545–2568). Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.naacl-main.201/>
- Lin, P.-J., Saeed, M., Chang, E., Scholman, M. (2023). Low-Resource Cross-Lingual Adaptive Training for Nigerian Pidgin. *Interspeech 2023* (pp. 3954–3958). Retrieved from <https://arxiv.org/abs/2307.00382>
- Lü, L., Zhang, Z.-K., Zhou, T. (2013). Deviation of Zipf’s and Heaps’ Laws in Human Languages with Limited Dictionary Sizes. *Scientific Reports*, 3, 1082, <https://doi.org/10.1038/srep01082>
- Marchal, M., Scholman, M., Demberg, V. (2021). Semi-automatic Discourse Annotation in a Low-Resource Language: Developing a Connective Lexicon for Nigerian Pidgin. *Proceedings of the 2nd Workshop on Computational Approaches to Discourse* (pp. 84–94). Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.codi-main.8/>
- Moreno-Sánchez, I., Font-Clos, F., Corral, Á. (2016). Large-Scale Analysis of Zipf’s Law in English Texts. *PLOS ONE*, 11(1), e0147073, <https://doi.org/10.1371/journal.pone.0147073> Retrieved from <https://doi.org/10.1371/journal.pone.0147073>
- Muhammad, S.H., Adelani, D.I., Ruder, S., Ahmad, I.S., Abdulmumin, I., Bello, B.S., ... Brazdil, P. (2022). NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis. *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 590–602). European Language Resources Association. Retrieved from <https://aclanthology.org/2022.lrec-1.63/>
- Myers-Scotton, C. (1993). *Social Motivations for Code-Switching: Evidence from Africa*. Oxford, UK: Oxford University Press. Retrieved from <https://global.oup.com/academic/product/social-motivations-for-codeswitching-9780198239239>
- Myers-Scotton, C. (2002). *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford, UK: Oxford University Press. Retrieved from <https://global.oup.com/academic/product/contact-linguistics-9780198299530>
- Ogbona, G., & Teahan, W. (2026). *Annotated Ground Truth Dataset for Code-Switching Detection Between Nigerian Pidgin English and Standard English*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.20117988>

- Piergallini, M., Shirvani, R., Gautam, G.S., Chouikha, M. (2016). Word-Level Language Identification and Predicting Code-Switching Points in Swahili-English Language Data. *Proceedings of the Second Workshop on Computational Approaches to Code Switching* (pp. 21–29). Austin, Texas: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/W16-5803/>
- Poplack, S. (1980). Sometimes I'll Start a Sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a Typology of Code-Switching. *Linguistics*, 18(7/8), 581–618, <https://doi.org/10.1515/ling.1980.18.7-8.581>
- Poplack, S., & Walker, J.A. (2003). Pieter Muysken, Bilingual Speech: A Typology of Code-Mixing. *Journal of Linguistics*, 39(3), 678–683, <https://doi.org/10.1017/S0022226703272297>
- Rivera, G.C. (2019). *Automatic Detection of Code-Switching in Arabic Dialects* (Doctoral dissertation, Massachusetts Institute of Technology). Retrieved from <https://dspace.mit.edu/handle/1721.1/123151>
- Samanta, B., Reddy, S., Jagirdar, H., Ganguly, N., Chakrabarti, S. (2019). A Deep Generative Model for Code-Switched Text. *arXiv preprint arXiv:1906.08972*, , Retrieved from <https://arxiv.org/abs/1906.08972>
- Sitaram, S., Chandu, K.R., Rallabandi, S.K., Black, A.W. (2019). A Survey of Code-switched Speech and Language Processing. *arXiv preprint arXiv:1904.00784*, , <https://doi.org/10.48550/arXiv.1904.00784> Retrieved from <https://arxiv.org/abs/1904.00784>
- Tarmom, T., Teahan, W., Atwell, E., Alsalka, M.A. (2020). Compression versus Traditional Machine Learning Classifiers to Detect Code-Switching in Varieties and Dialects: Arabic as a Case Study. *Natural Language Engineering*, 26(6), 663–676, <https://doi.org/10.1017/S135132492000011X>
- Teahan, W.J., & Harper, D.J. (2003). Using Compression-Based Language Models for Text Categorization. W.B. Croft & J. Lafferty (Eds.), *Language Modeling for Information Retrieval* (Vol. 13, pp. 141–165). Dordrecht: Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-94-017-0171-6_7
- Winata, G.I., Aji, A.F., Yong, Z.-X., Solorio, T. (2022). The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. *arXiv e-prints*, , Retrieved from <https://arxiv.org/abs/2212.09660>

Winata, G.I., Cahyawijaya, S., Lin, Z., Liu, Z., Fung, P. (2021). Are Multilingual Models Effective in Code-Switching? *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Retrieved from <https://aclanthology.org/2021.calcs-1.20/>