Caribbean
Educational
Research Journal

Redefining AI for Caribbean Creole: a framework for text and speech recognition

Christon Nunes, Matthew Mohammed, Keston Smith, Akash Pooransingh and Daniel Joseph Ringis
Department of Electrical and Computer Engineering, The University of the West Indies, St Augustine, Trinidad and Tobago

Received 24 March 2025
Revised 18 August 2025
Accepted 17 November 2025

Abstract

Purpose – Large language models are versatile enough to understand almost any high resource language perfectly. However, low resource languages, such as Trinidadian English Creole, are not as easily recognized, either as input or output. This regularly stems from a lack of available labelled data. This study discusses how local data combined with augmented data can improve the recognition and translation of Trinidadian English Creole by AI systems.

Design/methodology/approach – We propose a unique framework in which a small number of localized sample data were used to generate a larger appropriate dataset for training. This aided in the data availability problem often seen with languages such as Trinidadian English Creole. This generated dataset then had to be manually verified or fine-tuned to be used in the retraining of chatbots.

Findings – Using the fine-tuned AI system, the accuracy of translation had jumped up to a BLEU score of 75% from 59% of the evaluation set, with a 1-s response time. This is further verified with subjective testing results. This was using 3000 sentences of training data, and it is expected that with more data this result can be improved even further.

Originality/value – The framework presented is transferable and can be applied to any low resource language. This has great benefit to the wide range of Creole languages in the Caribbean region, thereby underscoring the importance of creating natural language processing tools that serve linguistically minority communities. Future work can expand this approach to include speech-based systems, code-switched data and regional Creole variants, all while continuing to engage native speaker communities in a participatory design process.

Keywords Generative AI, Large language models, Speech recognition, Retraining, Trinidad Creole

Paper type Technical paper

Creole languages are natively spoken by millions and represent complex, fully developed linguistic systems with rule-governed structures (DeGraff, 2003; Bakker, Daval-Markussen, Parkvall, & Plag, 2011). Originating from intense contact among diverse linguistic communities, Atlantic Creoles, such as those found in the Caribbean, often exhibit lexical overlap with their European lexifiers (e.g. English, French, Spanish), but develop distinct phonological, morphosyntactic and pragmatic systems (Kouwenberg & Singler, 2008). Rather than viewing these languages as simplified or "broken" forms of their lexifiers, current linguistic consensus acknowledges their status as independent languages shaped by sociohistorical context.

Creole languages are rich and expressive, yet they are still sidelined in most natural language processing (NLP) systems. A big reason is history: colonial attitudes branded them as "lesser languages," a stigma that has lingered (Alleyne, 1971; DeGraff, 2003). On the technological side, mainstream NLP tends to favour high-resource languages with fixed spellings and large, tidy datasets, which leaves many Creoles, often low-resource and not formally recognized, out of the picture (Joshi, Santy, Budhiraja, Bali, & Choudhury, 2020; Tatariya, Lent, & de Lhoneux, 2023). Add to that the fact that Creoles differ widely by region and lexifier (for example, English-based Caribbean vs. French-based varieties), each with its own morphosyntax, making a one-size-fits-all NLP rare (Joshi *et al.*, 2020).

Caribbean Educational Research Journal
Vol. 10 No. 1, 2026
pp. 99-114

2978-6177
DOI 10.1108/CERJ-03-2025-0003

Using Trinidad and Tobago as an example illustrates this complexity. Centuries of colonization, migration and cultural mixing have produced a layered linguistic landscape where multilingualism and multidialectalism thrive along a Creole continuum (Ferreira, 1997). At its core there are two main varieties: Trinidad and Tobago Standard English (TTSE) and Trinidad English Creole (TEC) (Ferreira, 1997). TTSE dominates schools, media and government, while TEC is what most people use day to day. This is classic diglossia, different varieties for different social settings, and most speakers shift between them with ease based on who they're talking to and why. This is also known as "code switching" (Winford, 1997).

Even so, TEC has long been dismissed as "bad English," a view rooted in colonial ideologies that linked non-European speech with low status or poor education (Alleyne, 1971; DeGraff, 2003). Those attitudes helped keep TEC out of writing, out of formal institutions and, more recently, out of tech (Lent, Bugliarello, de Lhoneux, Qiu, & Søgaard, 2021). As a result, it barely shows up in digital corpora, which makes training NLP systems hard. From an NLP standpoint, the challenges are clear: TEC is mostly oral, it doesn't have a standardized orthography and it often appears in code-mixed forms. To represent it well, projects need native speakers involved in annotation and an understanding that its variability follows real sociolinguistic patterns, not random noise. These principles are key to building AI tools that are both culturally aware and linguistically accurate for Creole languages.

English-lexicon Creoles such as TEC exhibit a range of morphosyntactic features that distinguish them from their lexifier languages (Lent, Ogueji, de Lhoneux, Ahia, & Søgaard, 2022). These include preverbal tense, mood and aspect (TMA) markers, serial verb constructions, null copula in equative contexts and the use of invariable pronouns (Kouwenberg & Singler, 2008). For instance, rather than using inflected auxiliary verbs for tense (e.g. "has eaten"), TEC might use completive markers like "done" (e.g. "He done eat"). Word order in interrogative sentences remains declarative, and passive constructions are typically agentless or use lexical strategies rather than auxiliary verbs. Such features are not simplifications but reflect typologically consistent grammatical strategies commonly found in Creole languages (DeGraff, 2003). It uses a fixed sequence of preverbal markers to express TMA, in contrast to MSE's inflectional system. Negative concord with a single negator is common, and distinctions are made between equative, attributive and locative copulas. Syntactically, Creole maintains a consistent word order in statements and questions, lacking subject-auxiliary inversion. Focused elements are often fronted for emphasis, and passive equivalents exclude agentive phrases and auxiliary verbs, while relative clauses may involve pronoun copying (Kouwenberg & Singler, 2008).

Despite recent advances in NLP research, the focus has largely remained on about 20 of the world's 7,000 languages (Magueresse, Carles, & Heetderks, 2020), leaving Creole languages significantly underexplored (Lent *et al.*, 2022). Low-resource languages require a large corpus to train models; thus, they are affected by this limitation, as they lack large-scale corpora for training monolingual models. Developing such corpora requires language experts and is a time-consuming process, creating a significant bottleneck in NLP development for these languages. This gap is particularly evident when considering NLP tools for Creole subsets such as Trinidadian Creole and Jamaican Patois, which exhibit unique structural differences from their base languages. These differences have resulted in underdeveloped and ineffective language models for Creole (Bancu *et al.*, 2024).

Furthermore, although interest in low-resource languages has grown within the NLP community (Ebrahimi *et al.*, 2021; Inuwa-Dutse, 2021; Agić & Vulić, 2019), yet Creole languages still remain critically underrepresented. However, Creole languages especially offer an entry point for cross-lingual transfer learning, a technique in NLP where knowledge from high-resource languages is transferred to related low-resource ones. In the case of TEC, this technique may bridge the resource gap by leveraging English-trained models while accounting for Creole-specific grammar.

Recent research has highlighted challenges in accurately translating and recognizing low-resource languages, including Creoles. Various techniques have been explored to bridge the

performance gap of large language models (LLMs) with Creole languages, such as sentiment analysis, word embedding and dataset partitioning (Mohammed & Prasad, 2023; Le, Moeljadi, Miura, & Ohkuma, 2016; Ghafoor *et al.*, 2021; Basiri & Kabiri, 2018). While previous studies have made significant strides in developing NLP resources and models for specific Creole languages, there remains a gap in creating a unified framework that addresses both text and audio recognition across diverse Caribbean Creole languages.

Caribbean languages, particularly Creoles, are at risk of linguistic erosion due to the dominance of English, French and Spanish in formal education and governance. NLP plays a crucial role in the preservation and revitalization of endangered and low-resource languages by enabling digital representation, documentation and accessibility (Zhang, Benjamin, & Bansal, 2022). The use of NLP techniques such as automatic speech recognition and machine translation (MT) can bridge the digital divide by integrating Creole languages into mainstream technology. In the Caribbean context, structured datasets can improve automated translation, speech-to-text systems and chatbots capable of processing Creole text. Additionally, the integration of NLP into educational tools allows speakers to learn, write and communicate more effectively in their native languages, promoting cultural identity.

The aim is to build on previous insights by presenting a framework for retraining AI tools specifically for Caribbean Creole–based text and audio recognition. By leveraging the linguistic commonalities and differences between Caribbean Creole languages and their base languages, this study employs a hybrid data augmentation technique by combining generative AI with meticulous manual correction. This approach not only expands the dataset efficiently but also ensures high linguistic integrity by preserving the cultural and contextual nuances of TEC. Unlike conventional automatic translation pipelines, this method effectively addresses inaccuracies typical of generative models, thereby enhancing the quality and authenticity of the dataset.

Thus, the following research questions and objectives guide the framework of this study:

(1) How can existing pretrained language models be adapted to effectively translate TEC into modern standard English?

(2) What data collection and augmentation strategies best capture the linguistic and cultural nuances of TEC for training translation models?

(3) How does the performance of a retrained model compare to existing models in terms of accuracy and cultural relevance when translating from TEC to TTSE?

Research objectives:

(1) To develop a structured and culturally accurate dataset representing TEC aligned with TTSE.

(2) To fine-tune a pretrained language model, specifically BART, to translate TEC to modern standard English or TTSE.

(3) To evaluate the retrained model through both quantitative metrics and qualitative assessments by native speakers.

In addition to engaging with native speakers to gather authentic linguistic data, it is essential to ensure that the developed model remains culturally and contextually relevant. This framework has the potential to be generalized to other Caribbean Creoles, leveraging their lexical and syntactic similarities. Given the shared linguistic features among Caribbean Creoles, this approach could pave the way for developing NLP technologies across the region, thus addressing the broader challenge of low-resource language processing.

1. Limitations of existing commercial systems

While this study focuses specifically on TEC, we include comparative examples from Jamaican and Vincentian Creoles to illustrate the broader typological landscape of Caribbean

English-lexicon Creoles. These comparisons are not intended to generalize translation strategies across Creoles but rather to contextualize TEC's grammatical uniqueness within a family of related languages. Tables 1–3 highlight areas where these English-lexicon Creoles diverge from Modern Standard English and from one another in meaningful, linguistically structured ways.

The linguistic variations observed in Tables 1–3 showed significant structural and grammatical differences between Standard English and between the respective Creole languages. These differences were seen through various linguistic mechanisms, including the removal or substitution of prepositions, modifications to pronoun structures, variations in sentence construction, addition of plural markers, alterations in the placement of tense indicators and simplifications in verb conjugation. While these Creoles share a number of structural similarities in their divergence from Standard English, they also exhibited unique characteristics that distinguish them from one another, further demonstrating the complexity of Creole linguistic evolution (Mühleisen, n.d.; Farquharson, n.d.; Prescod, n.d.).

One of the more notable distinctions between MSE and the English-lexicon Creole is the presence of invariable pronouns and reduced use of contractions (Kouwenberg & Singler, 2008). In MSE, contractions such as "He's" and possessive adjectives like "his" were commonly used, as seen in "He's going to his parents." For instance in TEC, "He going by he parents" functions grammatically with "he" serving both subject and possessive roles, a pronominal feature of Creole languages attested by Mühleisen (n.d.). Jamaican Creole followed a similar pattern but further modified pronouns by incorporating "-dem" for pluralization, as seen in "Di pikni-dem a ramp ina di house" instead of "The children are playing in the house" (Farquharson, n.d.). Vincentian Creole generally maintained a more linear syntactic structure while still omitting auxiliary verbs, as illustrated in the example "Mi ha a Inglish tung in mi mout" ("I speak English"), where a possessive construction substitutes for the English auxiliary *have* (Prescod, n.d.). Although all three Creoles exhibit pronoun simplification, Jamaican Creole displays more overt morphological marking of plurality (e.g. dem as a plural marker), whereas Trinidadian and Vincentian Creole tend to rely more heavily on pragmatic context and discourse cues to convey plurality.

Another major difference was their question formation and interrogative structures. MSE typically relied on auxiliary verbs and specific word order, as seen in "Where do you live?" In contrast, TEC replaced "where" with "which part" and removed the auxiliary verb, producing "Which part yuh livin'?" (Mühleisen, n.d.). Many Atlantic Creoles form "wh-" questions using compound analytic structures such as "what person" or "what place." This is reflected in TEC expressions like "Which part yuh livin" ("Where do you live?"), aligning it with the broader Creole English typology. Jamaican Creole adopted a similar approach but included additional question particles, as seen in "Wich-paat yu waak gu dong de?" for "Where do you walk to get down there?" (Farquharson, n.d.).Vincentian Creole, however, restructured the question by adding "Hu badi" literally translating to "Whose body" in "Hu badi du dat?" for "Who did that?" (Prescod, n.d.). This suggested that while all three Creoles favour directness over the more complex structure of MSE, Vincentian Creole exhibited a more complex

Table 1. Lexical and grammatical constructions in Trinidad English Creole (Mühleisen, n.d.)

Standard English structure	Trinidadian Creole structure	Literal Trinidadian Creole translation
He's going to his parents [4]	He going by he parents	He going by he parents
Where do you live? [19]	Which part yu livin?	Which part you living?
The cups [24]	Di cups an dem	The cups and them
He has already eaten [43]	He done eat already	He done eat already
She ate the food [51]	She eat the food	She eat the food

Table 2. Lexical and grammatical constructions in Jamaican Creole (Farquharson, n.d.)

Standard English structure	Jamaican Creole structure	Literal Jamaican Creole translation
The children are playing in the house [4]	Di pikni-dem a ramp ina di hous	The child and them are romping in the house
Where do you walk to get down there? [19]	Wich-paat yu waak gu dong de?	Which part you walk go down there?
Rose told her that it was Claris who broke the pots [24]	Ruoz tel im se a Klaris mash di pat-dem	Rose tell them say that Claris smash the pot
This child was always ugly [43]	Dis-ya pikni wehn aalwiez ogli	This here child that was always ugly
Mary cooked the food [51]	Mieri kuk di fuud	Mary cook the food

Table 3. Lexical and grammatical constructions in Vincentian Creole (Prescod, n.d.)

Standard English structure	Vincentian Creole structure	Literal Vincentian Creole translation
I speak English [4]	Mi ha a inglish tuhng ina fomi mout	I have an English tongue in my mouth
Who did that? [19]	Hu badi du dat?	Who body do that?
Rhoda and her friends [24]	Rhoda dem	Rhoda and them
They are so desperate and dishonest that they will do anything to have me locked up [43]	De so despareit an lai dat de go du eniting fo lak mi uhp	They are so desperate and lie that they go do anything to me locked up
She ate the food [51]	She eat the food	She eat the food

restructuring in question formation, whereas Jamaican Creole incorporates additional modifying particles. Pluralization also diverges significantly between Standard English and the Creoles. In Standard English, plurality is typically marked morphologically with an -s suffix, as in "the cups." Trinidadian and Jamaican Creole, however, often use the associative plural marker dem, as in "di cups an dem" (Mühleisen, n.d.) and "di pat-dem" (Farquharson, n.d.), respectively. In both Creoles, they and them are frequently used interchangeably as third-person plural subjects, reflecting reduced morphological differentiation. As seen, Trinidadian Creole tends to rely more on lexical and syntactic strategies than suffixation for marking plural forms (Richards, 1970). Vincentian Creole, by contrast, often omits explicit plural markers and instead depends on contextual interpretation, as in "Rhoda dem" for "Rhoda and her friends" (Prescod, n.d.). Among the three, Jamaican Creole shows the most consistent use of "dem" to pluralize nouns, while Trinidadian Creole applies it selectively and Vincentian Creole prefers pragmatic inference over overt marking.

Further, morphosyntactic features such as the TMA system represent a significant grammatical difference between Creole English and MSE. Tense marking and its placement also distinguished the specified Creoles from MSE and from each other. In MSE, adverbial tense markers such as "already" normally would precede the verb, as in "He has already eaten." In TEC, the completive aspect is expressed analytically using invariant markers like done, as seen in "He done eat already," corresponding to the TTSE equivalent (Mühleisen, n.d.). Jamaican Creole also not only changed the tense placement but also modified the sentence structure further, as seen in "Dis-ya pikni wehn aaliwez ogli" for "This child was always ugly" (Farquharson, n.d.). Vincentian Creole took a more distinct approach, frequently using repeated verbs for emphasis, as seen in "De go go eniting fo lak mi uhp" for "They are so desperate and dishonest that they will do anything to have me locked up" (Prescod, n.d.).

This suggested that while all three Creoles reorder tense markers in contrast to MSE, Vincentian Creole also incorporated verb doubling for emphasis.

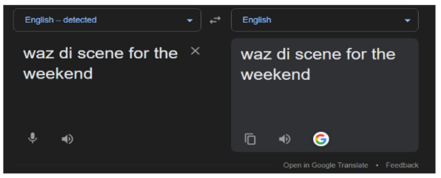
Verb tense simplification was another key difference between Standard English and the specified Creoles. In MSE, verbs are inflected to indicate past tense, as in "She ate the food." In contrast, all three Creoles maintain the same verb form across tenses, producing "She eat the food" (Mühleisen, n.d.; Farquharson, n.d.; Prescod, n.d.). Jamaican Creole, however, shows additional phonological differences in verb realization, as in "Mieri kuk di fuud" for "Mary cooked the food." Meanwhile, Trinidadian and Vincentian Creole tend to retain a more uniform surface form, without significant phonological alteration of the verb. Trinidadian and Vincentian Creole retain a more uniform structure (Farquharson, n.d.). This indicates that while all three Creoles used tense simplification, Jamaican Creole exhibited more phonetic variation in verb forms.

In many Atlantic Creoles like TEC, focus is expressed through fronting, moving the element you want to emphasize to the beginning of the sentence. This is often referred to as clefting when a copula is involved, especially for contrastive focus. The function of contrastive focus is generally to identify some participant, entity, place or time within a sentence (Kouwenberg & Singler, 2008). In many English lexicon Creoles, this is achieved using a focus marker such as "a," but in TEC, the equivalent copula is typically "is." The position of this marker at the front of the sentence allows for multiple grammatical structures that express the same core meaning, while foregrounding different pieces of information.

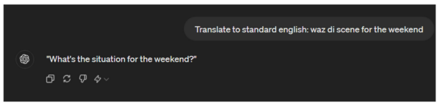
One key syntactic feature in TEC is its strategy for expressing focus through cleft constructions and fronting, often realized with the copula *is* (e.g. Is he cook di food). These constructions serve to mark contrastive or presentational focus and are typologically distinct from equivalent English structures (Kouwenberg & Singler, 2008). For MT systems, misinterpreting this syntax can obscure the speaker's pragmatic intent. Accurate modelling of these constructions thus requires attention not only to lexical equivalence but also to discourse-level grammar. The fronting of focused constituents in Trinidadian Creole introduces a significant challenge for MT. Misinterpreting such fronting can lead to a misrepresentation of the intended pragmatic meaning leading to significant loss in translation.

Although Trinidadian Creole, Jamaican Creole and Vincentian Creole shared multiple linguistic features, such as the removal of auxiliary verbs, the simplification of the verb tense, the restructuring of interrogative sentences and the use of unique pluralization markers, each Creole language exhibited unique characteristics that distinguished them from each other. Jamaican Creole was the most structurally explicit in marking plurality and modifying interrogative forms through additional particles. TEC primarily relies on invariant word order and consistent placement of preverbal tense markers, often leaving plurality to be understood from context rather than overtly marked. Vincentian Creole, however, was the most context-driven, removing auxiliary verbs and tense markers even more explicitly while incorporating repeated verbs for emphasis. These variations highlight the dynamic nature of Creole languages and their evolution as independent linguistic systems, further enforcing the notion that a generalized Creole translator would result in inconsistencies between the different regions and their Creoles.

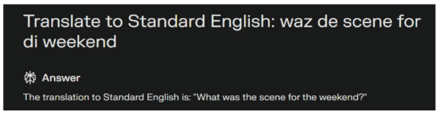
Several TEC sentences passed into GPT-3.5 (Singh *et al.*, 2025) displayed this behaviour and in some instances translated on a word to word basis rather than per phrase. Gemini 1.5 (formerly BARD), an LLM created and used by Google, was slightly more accurate than GPT-3.5, TEC prompts were left unidentified and in some cases misunderstood, some responses given from Gemini even instructed to restate the sentence in plain English. Perplexity AI, another chatbot that uses their own LLM, when tested on TEC prompts, was able to comprehend that the language was Creole but did not recognize that it was specifically TEC and was therefore unable to recognize any of the phrases similar to GPT-3.5. In contrast, similar prompts were also passed into Google Translate and it was unable to recognize or successfully translate any of the TEC prompts. Figure 1 shows the results of these LLMs for a common TEC phrase, "Waz de scene for di weekend."



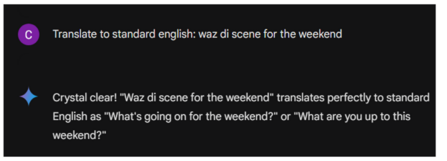
Case 1: Google Translate Trinidad Creole input result



Case 2: GPT-3.5 Trinidad Creole input result



Case 3: Perplexity AI Trinidad Creole input result



Case 4: Gemini AI Trinidadian Creole Input result

Figure 1. LLM translation results for a common Trinidad English Creole input: "Waz de scene for di weekend"

This work uses high frequency sentences to refer to common phrases, greetings and conversational phrases that are said frequently among Trinidadian speakers. The structural characteristics of TEC, such as zero copula constructions, preverbal tense markers (e.g. done, go), negative concord and non-inflectional plural strategies, present challenges for translation models that are typically trained on languages with inflectional morphology and fixed word order. These features reflect a consistent internal grammar rather than irregularity. Accordingly, high-frequency sentences used in this study were selected to capture a representative range of these features and inform a translation system attuned to TEC's syntactic and semantic patterns (Winford, 1997). Thus, sentences were chosen so as to ensure most unique features of TEC were included, increasing accuracy.

2. Comparable systems in literature

Currently, there are no chatbots designed with the ability to interpret TEC or its dialects. However, there are existing chatbots trained in English and Arabic dialects. General English–based NLP chatbots often use American English or British English datasets. British English and American English are both dialects of the English language. Additionally, British and American English variations also contain their own dialects. The most notable chatbots that utilize English dialect datasets or knowledge bases (KBs) are ALICE, Replika, Dialo-GPT and Meena among many other chatbots. Chatbots like Replika, Meena and Dialo-GPT are conversational chatbots that use neural networks with a modern version of latent semantic analysis trained with a large corpus of conversational and natural questions from social media sites.

Botta and Nabiha are two Arabic dialect chatbots (Ali & Habash, 2016). BOTTA is the first Arabic dialect conversational chatbot that simulates amicable interactions with users by using the widely understood Egyptian Arabic dialect (Ali & Habash, 2016). BOTTA uses the Artificial Intelligence Mark-up Language (AIML) technology which utilizes pattern matching to connect a user input with a KB response. AIMLs are used in forms of objects to simplify conversational modelling in terms of a "stimulus-response" process (Hasal *et al.*, 2021). An AIML object's core components are categories, patterns and templates. For each category, a response template and a set of criteria that give the template meaning, known as context, are constructed (Hasal *et al.*, 2021). Nabiha is a question-answering chatbot that uses the AIML method of implementing a chatbot with a custom KB or dataset gathered from students' questions and complaints (Al-Ghadhban & Al-Twairesh, 2020).

AIML technology is not ideal for chatbot implementation regarding the TEC language since it is not yet standardized and would require a large corpus to achieve nominal results. Additionally, due to the complexity of the TEC language, this makes scalability very limited. However, the use of recurrent neural networks is a viable option in dealing with the complexities of TEC. The implementation of a new Creole chatbot would require a very large corpus of Creole data which is not publicly available unless generated ourselves.

The unique linguistic structures and cultural contexts of low-resource languages and Creoles present challenges in accurate translation, sentiment analysis and overall language modelling (Lent *et al.*, 2021). Additionally, sentiment analysis in low-resource languages often relies on lexicon refining and hybrid models. Ehsan Basiri and Kabiri (2018) demonstrated improvements in Persian sentiment analysis through lexicon refining, while Le *et al.* (2016) explored sentiment analysis in informal Indonesian tweets using data augmentation and transfer learning. Mohammed and Prasad (2023) proposed a lexicon-based approach for low-resource languages, emphasizing the need for culturally contextual sentiment lexicons. Ghafoor *et al.* (2021) showed that translating resource-rich datasets to low-resource languages improves performance in multilingual models. Kumar, Jha, and Sahula (2020) utilized augmented translation techniques for Sanskrit-Hindi pairs, highlighting the potential for similar approaches in Creoles. These studies emphasize that data augmentation is crucial for expanding low-resource language datasets.

Fine-tuning has emerged as a crucial technique for leveraging large pretrained language models (LMs) in low-resource NLP tasks. For other tasks such as MT and dialogue generation, fine-tuning is also the prevalent paradigm (Zhang, Sax, Zamir, Guibas, & Malik, 2019; Stickland, Li, & Ghazvininejad, 2020; Liu *et al.*, 2020). However, it requires updating and storing all the parameters of the LM, which can be prohibitively expensive due to the large size of modern LMs, such as GPT-2 with 774M parameters (Radford *et al.*, 2019) and GPT-3 with 175B parameters (Brown *et al.*, 2020). There are a diverse number of pre-trained NLP models, such as Text-to-Text Transfer Transformer (T5) and Bidirectional Encoder Representations from Transformers (BERT). T5 optimized for sequence-to-sequence tasks but requires large-scale pretraining. While BERT is primarily designed for understanding tasks, integrating its contextual embeddings into MT systems has shown promise (Wang, Karthikeyan, Mayhew, & Roth, 2020). For low-resource languages, the scarcity of parallel corpora hinders the direct training of robust translation models. To address this, researchers have explored methods such as fine-tuning pre-trained multilingual BERT models on limited data, leveraging cross-lingual transfer learning to enhance translation quality (Wang *et al.*, 2020). In this study, we employ a similar paradigm by tuning a unique dataset and then using it to tune a standard English LLM to adapt it for Trinidadian Creole, ensuring that the model captures the unique linguistic features and cultural nuances effectively.

Transformer-based models have significantly improved the quality of text-to-text MT (Liu *et al.*, 2020; Fan *et al.*, 2020). However, there is a trade-off between the number of languages a model supports and its efficiency. Multilingual models, such as mBERT and XLM-R, are designed to support multiple languages, but their performance is often lower for individual low-resource languages compared to models trained specifically on a single language (Siddhant *et al.*, 2020).

MT approaches for low-resource languages include two main strategies: (i) translating data from a resource-rich language to a low-resource language to train a model specifically for the low-resource language and (ii) training a model in the resource-rich language and translating instances from the low-resource language for evaluation (Conneau *et al.*, 2018; Ghafoor *et al.*, 2021). Often, models trained on low-resource text yield lower accuracy. To address this, researchers suggest training on resource-rich languages, like English, which are better understood by learning algorithms (Araújo, Pereira, & Fabrício, 2020). Barriere and Balahur (2020) proposed a multilingual transformer model that combined translated English datasets with smaller native datasets for French, Spanish, German and Italian. This merged dataset approach outperformed using smaller original corpora alone (Barriere and Balahur, 2020).

This strategy is particularly relevant for low-resource Trinidaian Creole, as cultural nuances can be preserved through hybrid augmentation techniques, as proposed in this framework. Thus, this study also highlights that the dataset is a significant factor to accuracy of the model.

Prompting involves prepending instructions and a few examples to the task input and generating the output from the LM. GPT-3 (Brown *et al.*, 2020) employs manually designed prompts to adapt its generation for different tasks, a framework known as in-context learning. In this aspect, prompting can be employed for existing LMs to tailor datasets and provide a corpus of data which can then be used to train a model for MT on a low resource language such as Creole.

3. Methodology

With consideration of the findings presented, here an approach is developed to improve existing LMs, by using existing pretrained LMs and AI tools that have generally understood Creole. In the proposed framework shown in Figure 2, we manipulate these existing tools, by including unique data and modifying generated data, to retrain and create a better performing model, specific to Trinidad Creole.

In retraining a pretrained English LM to accurately translate into TEC, a multi-step approach was employed to ensure both linguistic fidelity and cultural appropriateness. As discussed, TEC possesses a well-formed but typologically distinct grammar from English, including analytic tense/aspect systems, zero copula constructions and flexible noun phrase plurality; thus, our approach prioritized incremental learning. The process began with the careful selection of an appropriate LM. Multiple candidate models were tested for their translation capabilities, and the Bi-directional and Auto-Regressive Transformer (BART) became the preferred choice. Other LMs trained were BERT and T5. The model trained with T5 gave repeated errors and showed no increase in accuracy with dataset tuning. BART's ability to reconstruct corrupted inputs through bidirectional modelling enhances its performance for Creole text generation and translation, where grammatical inconsistencies and informal orthographies are prevalent. The BART model was selected for its hybrid encoder–decoder architecture, which supports contextualized bidirectional representations necessary to model TEC's non-inflectional morphology and discourse-level structure.

Central to the framework was the development of a dataset that authentically captures the morphosyntactic and lexical characteristics of TEC. Data scarcity, orthographic variation and the oral nature of Creole languages made this process particularly challenging. A base corpus of high-frequency words, phrases and constructions was constructed to reflect key features like

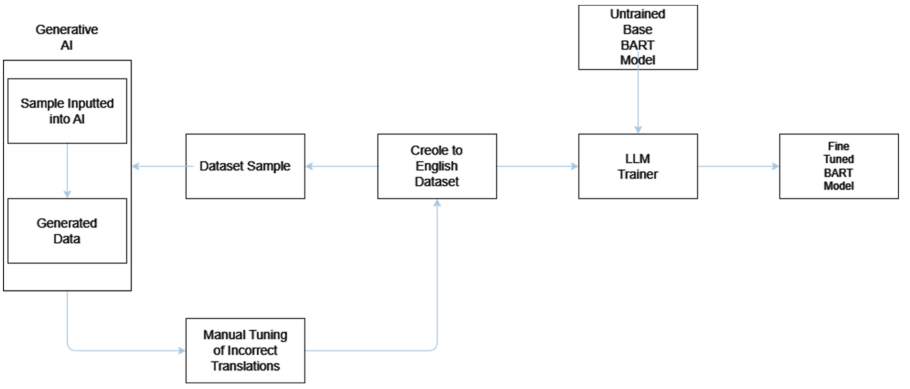


Figure 2. Proposed framework for retraining standard English LLM to Trinidad Creole

preverbal tense markers (go, done), use of dem as an associative plural and Creole-specific interrogatives (which part yuh deh?). Data were augmented with synthetic examples generated via prompt engineering, followed by rigorous manual review from native speakers to preserve naturalistic usage and avoid hyper-Anglicized translations.

One of the main obstacles faced was the lack of large, structured corpora available for training models. Most Creole data are scattered across social media, informal writings and speech recordings, which are often inconsistent in terms of language conventions. The dataset was constructed by aligning pairs of words, phrases and sentences, with each TEC, matched to its corresponding TTSE version. The initial phase involved curating a collection of common words, phrases and expressions unique to TEC. These carefully chosen phrases provided the foundational language structures necessary for the model to learn authentic translation patterns. Collaboration with native speakers to manually translate TEC into TTSE text ensured cultural accuracy while diversifying the dataset with authentic language usage. Once these baseline entries were established, a representative sample was extracted and used as seed data for the expansion process. To augment and add more variety to the dataset further, a generative AI tool was employed to produce additional words, phrases and sentence pairs following the same format. Although this tool was capable of rapidly generating a diverse array of sentences, the process did not end with the rapid generation. Each produced sentence was then reviewed by the trainers. During this manual refinement, inaccuracies and contextual discrepancies were identified and adjusted, "tuned" to reflect the genuine usage and cultural nuances inherent in TEC. This manual intervention was crucial, ensuring that the expanded dataset maintained a high standard of linguistic fidelity and authenticity. After the initial dataset was filled with both curated and generated content, it was organized in a deliberately structured manner. The dataset was structured with the pairs arranged in increasing order of complexity. This incremental design allowed the model to first master simple word structures before gradually tackling more complex phrases and sentence expressions. This approach was key to enabling the model to build a deep understanding of TEC, as without this structured dataset, the model would not be able to translate any inputs due to lack of proper understanding of the nuanced linguistic properties of the Creole.

Training the model involved fine-tuning the base BART model using its default training parameters, followed by iterative cycles of learning and refinement. Each training iteration was designed to optimize the model's ability to generalize from the structured dataset, reinforcing both common usage patterns and the more intricate linguistic features. This methodical training process not only enhanced translation accuracy but also contributed valuable insights into how increasing the complexity of training data can directly impact overall model performance.

When it came to the evaluation of the model, first, a baseline evaluation was performed in which 20 native speakers of TEC and MSE rated a selection of translations. These individuals were sampled based on three primary criteria: (1) self-identified fluency in both TEC and TTSE, (2) Trinidadian nationality and (3) previous experience in language teaching or linguistics. The pool included individuals from both academic and non-academic backgrounds to ensure broader representativeness of real-world usage.

Informants were asked to assess a randomized selection of machine translated sentences. Each sentence was presented alongside its Standard English input, and informants were instructed to rate the Trinidadian Creole output on two main criteria: fluency (how naturally the sentence sounded) and accuracy (how faithfully it preserved the original meaning). These assessments were rated on a 5-point Likert scale and supplemented with open-ended qualitative comments. The sentences chosen covered a range of tense, plurality and interrogative constructions to assess performance across different linguistic features.

In parallel, the Bi-Lingual Evaluation Understudy (BLEU) score was employed as the primary quantitative metric. The BLEU score serves as a standard evaluation metric, enabling a consistent comparison of translation outputs across different models and approaches. By utilizing BLEU, we ensure that results are comparable within the broader research

community, facilitating the assessment of improvements and the effectiveness of various techniques in low-resource language translation (Bansal, Kamper, Livescu, Lopez, & Goldwater, 2018). The BLEU score compared machine-generated translations with one or more reference translations, assigning a score from 0 to 1, where a score of 1 represents a perfect match (Papineni, Roukos, Ward, & Zhu, 2002).

This metric allowed for an objective measurement of translation quality and provided a standard against which to gauge improvements. To further contextualize the performance of the retrained model, the results from both the human evaluations and BLEU score assessments were compared against those of existing LMs. This comparative analysis was necessary in identifying the strengths and limitations of the newly developed translator, and it established a benchmark for future enhancements. The insights gleaned from this evaluation phase not only confirmed the model's effectiveness but also highlighted areas where further refinements could be made.

These findings mattered for how we build language tech for Creole-speaking communities. First, they showed why native speakers needed to be involved from the start, not just to evaluate results but to help build datasets and check grammatical judgements. Second, they exposed a gap in today's NLP models: handling underrepresented languages like TEC often takes heavy fine-tuning or extra rule-based fixes. And third, the strong performance of the tuned system signals that Creole-specific translation tools are both doable and genuinely useful, especially for education and cultural preservation.

Beyond the technical win, this translator serves a bigger sociolinguistic goal; it helps legitimize TEC in spaces where it's long been shut out. By giving TEC a formal computational footprint, the project pushes digital language technologies in a decolonizing direction. It also nurtures linguistic pride among speakers who've been told that Creole is "uneducated" or "improper." Treating TEC as a structured, valid system doesn't just chip away at stigma, it creates room for real inclusion in schools, government services and AI-driven communication. It also opens a path to future tools: TEC-inclusive learning materials, official translations and even digital assistants that understand and respond to Creole input.

4. Results

The framework used required experimentation to validate results. In preparation, the trained model and tokenizer was stored locally. The dataset used to train the model was gathered firstly through online dictionaries, for words. Over 600 words were gathered, including manually added words relevant to Trinidad's modern culture. Using these direct translations of words, a sample of 50 lines of data, containing words and sentences was input into generative AI with didactic prompts. This was done repeatedly with all words; up to 1800 lines of data were gathered. Further data augmentation included manually creating unique samples of 20 sentences each organized around conversational theme, to improve accuracy. The final dataset resulted in 3,400 lines arranged in an ascending order, with simpler sentences at the top and more complex sentences at the lower half. The data contained duplicates and errors in translations of Trinidadian Creole, resulting in a poor performing model as seen in Figure 3.

To enhance model performance, iterative refinement of the dataset was conducted. Duplicate entries and mistranslations that misrepresented the grammar of TEC were removed or corrected. A significant portion of errors stemmed from generative tools imposing English syntactic rules on TEC, resulting in ungrammatical or contextually inappropriate outputs. Manual review helped ensure proper handling of grammatical features like null copula, preverbal tense/aspect marking and consistent pronoun reference. The final corpus contained 3,000 validated pairs structured by syntactic complexity. Sentences generated by AI were combed through to identify any incorrect translations and were tuned so that translations were only specific to Trinidadian Creole. The tuning resulted in 3,000 lines of data; however, the performance of the model increased as seen in Figure 4. BLEU score was the main metric used to evaluate the performance before and after the tuning phase. Human evaluation was also used; native speakers of Trinidadian Creole were asked to score the accuracy of translations on

```
Trinidad Creole: Ah hear 'bout de Paramin Parang Festival, festive tunes
Reference English: I heard about the Paramin Parang Festival, festive tunes
Tuned Generated English: I heard about the Paramin Parang Festival, festive tunes
Untuned Generated English: I heard 'bout the Paramin Parang Festival, festive tunes


Trinidad Creole: Yuh ever try de local pepper sauce, spicy delight
Reference English: Have you ever tried the local pepper sauce? Spicy delight
Tuned Generated English: Have you ever tried the local pepper sauce? Spicy delight
Untuned Generated English: You'll ever try the local pepper sauce, spicy delight


Trinidad Creole: Waz di scene for the weekend?
Reference English: What's the plan for the weekend?
Tuned Generated English: What's the scene for the weekend?
Untuned Generated English: What's the plan for the weekend?
```

Figure 3. Comparing results of recognition without tuning phase and with tuning (AI generated sentences)

```
Trinidad Creole: Ah fed up, dis sun too hot
Reference English: I'm fed up, this sun is too hot
Tuned Generated English: I'm fed up, this sun is too hot
Untuned Generated English: I'm fed up, this sun is too hot


Trinidad Creole: Ah cah wait fuh Carnival, iz rhell vibes
Reference English: I can't wait for Carnival, it's good vibes
Tuned Generated English: I can't wait for Carnival, I'm craving good vibes
Untuned Generated English: I can't wait for Carnival, I'm excited


Trinidad Creole: Yuh see de cricket game last nite owa
Reference English: Did you see that cricket game last night?
Tuned Generated English: Did you see the cricket game last night or what?
Untuned Generated English: You saw the cricket game last week owa
```

Figure 4. Comparing results of recognition without tuning phase and with tuning (native suggested sentences)

a scale of 1–5. Both the tuned and untuned models were evaluated on two separate sets; the former is a small sample of 20 sentences generated by AI and unmodified, the latter was 20 curated sentences suggested by native Trinidadian Creole speakers.

Table 4 shows that, before tuning of the dataset, the model performed inadequately compared to the performance with the tuned dataset. The primary metric is the BLEU score; higher percentage corresponds to higher accuracy of MTs vice versa. Native Trinidadian Creole speakers also rated the model after tuning with a better average score of accuracy than the model before tuning. Additionally, the comparison of results shows that the model after the tuning outperformed the model before tuning and maintaining an accuracy above 75%. The untuned model failed to

Table 4. Recognition rates without tuning phase and with tuning

Dataset description	BLEU score (with AI sentences) (%)	BLEU score (with native sentences) (%)	Average score of native Trinidadian Creole speakers
Before tuning	55.91	69.26	3.45
After tuning	76.36	75.16	4.50

adequately capture the grammatical and pragmatic conventions of TEC, frequently defaulting to gloss, English-influenced constructions that ignored Creole-specific features such as preverbal TMA markers, unmarked plurality and question formation. In contrast, the tuned model demonstrated improved linguistic coherence and sociocultural appropriateness, reflecting the value of expert-curated data for underrepresented languages with distinct typologies.

5. Conclusion

This study demonstrates that effective translation of TEC using LLMs is achievable when sociolinguistic realities are incorporated into the design and training pipeline. By treating TEC as a rule-governed, structurally rich language, rather than a deviation from English, the framework challenges dominant ideologies that have long marginalized Caribbean Creoles in both academic and technological domains. Beyond the 75% BLEU accuracy achieved, the project underscores the importance of creating NLP tools that serve linguistically minority communities. Future work can expand this approach to include speech-based systems, code-switched data and regional Creole variants, all while continuing to engage native speaker communities in a participatory design process.

About the authors

Christon Nunes: From Rousillac, Trinidad, pursuing BSc Electrical and Computer Engineering. With a specific interest in generative artificial intelligence (AI), and engagement in projects that leveraged AI technologies. Interested in pursuing a career related to control systems as well as generative AI.

Matthew Mohammed: From Port-of-Spain, Trinidad, studying BSc. in Electrical and Computer Engineering. Interested in a career related to degree of study as well as research on the applicable use of artificial intelligence.

Keston Smith: Keston is a graduate of the Department of Electrical and Computer Engineering at the University of the West Indies, St. Augustine. His research interests include natural language processing, cyber security and robotics.

Akash Pooransingh: Dr Akash Pooransingh is currently a lecturer in computer systems at the Department of Electrical and Computer Engineering, the University of the West Indies, St. Augustine Campus. His research interests include signal processing and audio, image and video processing as it relates to consumer applications and sports.

Daniel Joseph Ringis: Dr Daniel Joseph Ringis previously studied Electrical and Computer Engineering from 2008 to 2012 and 2016 to 2017 at the University of the West Indies. After which he completed a PhD at Trinity College Dublin, Ireland, in video processing. His current research area is in video processing, engineering education and sports engineering.

References

Agić, Ž., & Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3204–3210. doi: 10.18653/v1/P19-1310.

Al-Ghadhban, D., & Al-Twairesh, N. (2020). Nabiha: An Arabic dialect chatbot. *International Journal of Advanced Computer Science and Applications*, *11*(3), doi: 10.14569/ijacsa.2020.0110357.

Ali, D. A., & Habash, N. (2016). Botta: An Arabic dialect chatbot. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations* (pp. 208-212).

Alleyne, M. (1971). *Acculturation and the cultural matrix of creolization* (pp. 169–186). Hymes.

Araújo, M., Pereira, A., & Fabrício, B. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, *512*(February), 1078–1102. doi: 10.1016/j.ins.2019.10.031.

Bakker, P., Daval-Markussen, A., Parkvall, M., & Plag, I. (2011). Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole Languages*, *26*(1), 5-42. doi: 10.1075/jpcl.26.1.02bak.

Bancu, A., Peltier, J. P., Bisnath, F., Burgess, D., Eakins, S., Gonzales, W. D. W., … Baptista, M. (2024). Revitalizing attitudes toward Creole languages. *Decolonizing Linguistics*, *293*.

Bansal, S., Kamper, H., Livescu, K., Lopez, A. and Goldwater, S. (2018). Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. September.

Barriere, V., & Balahur, A. (2020). Improving sentiment analysis over non-English tweets using multilingual transformers and automatic translation for data-augmentation. In *Proceedings of the 28th International Conference on Computational Linguistics*, Stroudsburg, PA (pp. 266–71). International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.23.

Basiri, M. E., & Kabiri, A. (2018). Words are important: Improving sentiment analysis in the Persian language by lexicon refining. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *17*(4), 1–18. doi: 10.1145/3195633.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., …, & Amodei, D. (2020). Language models are few-shot learners. doi: 10.48550/arXiv.2005.14165.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations.

DeGraff, M. (2003). Against creole exceptionalism. *Language*, *79*(2), 391–410, doi: 10.1353/lan.2003.0114.

Ebrahimi, A., Mager, M., Oncevay, A., Chaudhary, V., Chiruzzo, L., Fan, A., … Ortega, J. (2021). AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., … Edunov, S. (2020). Beyond English-centric multilingual machine translation.

Farquharson, J. T. (2013), "Jamaican structure dataset", in Michaelis, S.M., Maurer, P., Haspelmath, M. and Huber, M. (Eds.), *Atlas of Pidgin and Creole Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, available at: http://apics-online.info/contributions/8 (accessed 21 February 2025).

Ferreira, Jo-A. (1997). (A brief overview of) the sociolinguistic history of Trinidad & Tobago. doi: 10.13140/RG.2.1.2889.6482.

Ghafoor, A., Ali, S. I., Daudpota, S. M., Kastrati, Z., Abdullah, R. B., & Mudasir, A. W. (2021). The impact of translating resource-rich datasets to low-resource languages through multi-Lingual text processing. *IEEE Access*, *9*, 124478. doi: 10.1109/ACCESS.2021.3110285.

Hasal, M., Nowaková, J., Ahmed Saghair, K., Abdulla, H., Snášel, V., & Ogiela, L. (2021). Chatbots: Security, privacy, data protection, and social aspects. *Concurrency and Computation: Practice and Experience*, *33*(19), e6426, doi: 10.1002/cpe.6426.

Inuwa-Dutse, I. (2021). The first large scale collection of Diverse Hausa language datasets. February.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. arXiv preprint arXiv:2004.09095.

Kouwenberg, S., & Singler, J. V. (2008). *The handbook of Pidgin and Creole studies*. Wiley-Blackwell.

Kumar, R., Jha, P., & Sahula, V. (2020). An augmented translation technique for low resource language pair: Sanskrit to Hindi translation, 377–383. doi: 10.1145/3377713.3377774.

Le, T. A., Moeljadi, D., Miura, Y., & Ohkuma, T. (2016). Sentiment analysis for low resource languages: A study on informal Indonesian tweets. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*.

Lent, H., Bugliarello, E., de Lhoneux, M., Qiu, C., & Søgaard, A. (2021). On Language models for creoles. In *CoNLL 2021 - 25th Conference on Computational Natural Language Learning, Proceedings* (pp. 58–71). doi: 10.18653/v1/2021.conll-1.5.

Lent, H., Ogueji, K., de Lhoneux, M., Ahia, O., & Søgaard, A. (2022). What a creole wants, what a creole needs. *arXiv preprint* arXiv:2206.00437.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 726–742. doi: 10.1162/tacl_a_00343.

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. arXiv preprint arXiv:2006.07264.

Mohammed, I., & Prasad, R. (2023). Building Lexicon-based sentiment analysis model for low-resource languages. *MethodsX*, *11*, 102460. doi: 10.1016/j.mex.2023.102460.

Mühleisen, S. (2013), "Trinidad English Creole structure dataset", in Michaelis, S.M., Maurer, P., Haspelmath, M. and Huber, M. (Eds.), *Atlas of Pidgin and Creole Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, available at: http://apics-online.info/contributions/6 (accessed 11 May 2024).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318).

Prescod, P. (2013), "Vincentian Creole structure dataset", in Michaelis, S.M., Maurer, P., Haspelmath, M. and Huber, M. (Eds.), *Atlas of Pidgin and Creole Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, available at: http://apics-online.info/contributions/7 (accessed 11 May 2024).

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever. (2019). "Language models are unsupervised multitask learners".

Richards, H. (1970). Trinidadian folk usage and Standard English: A contrastive study. *Word*, *26*(1), 79–87, doi: 10.1080/00437956.1970.11435582.

Siddhant, A., Johnson, M., Tsai, H., Ari, N., Riesa, J., Bapna, A., ... Raman, K. (2020). Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence* (Vol. 34(05), 8854–8861). doi: 10.1609/aaai.v34i05.6414.

Singh, S., Singh, N. and Singh, G. (2025), "GPT-3.5 vs. GPT-4, unveiling OpenAI's latest breakthrough in language models", in *2025 IEEE International Symposium on Smart Electronic Systems (iSES)*, IEEE, pp. 261-263.

Stickland, A. C., Li, X., & Ghazvininejad, M. (2020). Recipes for adapting pre-trained monolingual and multilingual models to machine translation, April. doi: 10.18653/v1/2021.eacl-main.301.

Tatariya, K., Lent, H., & de Lhoneux, M. (2023). Transfer learning for code-mixed data: Do pretraining languages matter?. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis*, Stroudsburg, PA (pp. 365–78). Association for Computational Linguistics. doi: 10.18653/v1/2023.wassa-1.32.

Wang, Z., Karthikeyan, K., Mayhew, S., & Roth, D. (2020). Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*. doi: 10.18653/v1/2020.findings-emnlp.240.

Winford, D. (1997), "Creole Formation in the context of contact linguistics", *Journal of Pidgin and Creole Languages*, Vol. 12 No. 1, pp. 131-151.

Zhang, J. O., Sax, A., Zamir, A., Guibas, L., & Malik, J. (2019). Side-tuning: A baseline for network adaptation via additive side networks. December.

Zhang, S., Benjamin, E. F., & Bansal, M. (2022). How can NLP help revitalize endangered languages? A case study and roadmap for the Cherokee language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (Vol. 1). doi: 10.18653/v1/2022.acl-long.108.

Further reading

Briakou, E., Wang, S. I., Zettlemoyer, L., & Ghazvininejad, M. (2022). BITEXTEDIT: Automatic bitext editing for improved low-resource machine translation. *Findings of the Association for*

Computational Linguistics: NAACL 2022, 1469–1485. doi: 10.18653/v1/2022.findings-naacl.110.

Ferreira (2012), "Caribbean languages and Caribbean linguistics," in *Caribbean Heritage*, University of West Indies Press.

Henry, W. (2000). The use of Creole alongside standard English to stimulate students' learning. In *Forum for promoting comprehensive education*. Triangle Journals, *42*(1), 23–27.

Jackson, S., Beekhuizen, B., Zhao, Z., & McEwen, R. (2025). GPT-4-Trinis: Assessing GPT-4's communicative competence in the English-speaking majority world. *AI and Society, 40*(3), 1785–1801. doi: 10.1007/s00146-024-01945-9.

Kumar, G. K., A. Singh Gehlot, S. Shaji Mullappilly, & K. Nandakumar (2022). MuCoT: Multilingual contrastive training for question-answering in low-resource languages. April.

Lent, H., Tatariya, K., Dabre, R., Chen, Y., Fekete, M., Ploeger, E., … Bjerva, J. (2024). CreoleVal: Multilingual multitask benchmarks for Creoles. *Transactions of the Association for Computational Linguistics*, *12*, 950–978. doi: 10.1162/tacl_a_00682.

Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA (pp. 3728–38). Association for Computational Linguistics. doi: 10.18653/v1/D19-1387.

Corresponding author
Daniel Joseph Ringis can be contacted at: daniel.ringis@uwi.edu