Recommendations for Overcoming Linguistic Barriers in Healthcare: Challenges and Innovations in NLP for Haitian Creole

Ludovic Mompelat
Department of Modern Languages and Literatures
University of Miami
Miami, FL, USA
lvm861@miami.edu

Abstract

Haitian Creole, spoken by millions in Haiti and its diaspora, remains underrepresented in Natural Language Processing (NLP) research, limiting the availability of effective translation tools. In Miami, a significant Haitian Creole-speaking population faces healthcare disparities exacerbated by language barriers. Existing translation systems fail to address key challenges such as linguistic variation within the Creole language, frequent code-switching, and the lack of standardized medical terminology. This work proposes a structured methodology for the development of an AI-assisted translation and interpretation tool tailored for patient-provider communication in a medical setting. To achieve this, we propose a hybrid NLP approach that integrates fine-tuned Large Language Models (LLMs) with traditional machine translation methods. This combination ensures accurate, context-sensitive translation that adapts to both formal medical discourse and conversational registers while maintaining linguistic consistency. Additionally, we discuss data collection strategies, annotation challenges, and evaluation metrics necessary for building an ethically designed, scalable NLP system. By addressing these issues, this research provides a foundation for improving healthcare accessibility and linguistic equity for Haitian Creole speakers.

Keywords: Haitian Creole, NLP, Healthcare, Low-resource Languages, LLM, Code-switching, Variation

1 Introduction

Creole languages have historically been underrepresented—and often outright ignored—in Natural Language Processing (NLP) research (Joshi et al., 2020; Lent et al., 2021). Many are classified as low-resource languages due to the scarcity of annotated datasets and corpora. This is largely attributed to several factors: the limited availability of speakers for endangered Creole languages, pervasive negative attitudes and stigmatization that discourage research investment, and the overall lack of theoretical and applied linguistic engagement with these languages (Mompelat, 2023). This neglect is particularly striking given that Creole languages, as a group, are spoken by millions of people worldwide. Their exclusion from NLP research increases linguistic inequalities and can limit access to crucial technologies, including healthcare-related applications.

Despite these challenges, there has been a growing effort to develop NLP solutions for Creole languages, leading to advancements in part-of-speech tagging, syntactic parsing, named-entity recognition, and machine translation (Cortegoso and Viktor, 2021; Ramsurrun et al., 2024; Robinson et al., 2024; Schieferstein, 2018; Dabre and Sukhoo, 2022; Lent et al., 2021; Macaire et al., 2022). Researchers have increasingly adopted hybrid approaches that combine traditional machine learning with more data-intensive neural and large language model (LLM) techniques to address data scarcity (Fekete et al., 2024; Smart et al., 2024). This hybrid approach has proven crucial for advancing NLP capabilities in low-resource contexts like Creole languages. However, most existing models fail to account for linguistic variation within Creoles, code-switching patterns, and domain-specific terminology—three key issues critical for real-world deployment, particularly in healthcare.

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 20–31
March 2, 2025 ©2025 Association for Computational Linguistics

Due to the rapid expansion of NLP and AI research, Creole researchers face a race against time and technological advances. This urgency often leads to an overemphasis on dominant varieties within specific Creoles, while linguistic variation—present in all natural languages, including Creoles—receives insufficient attention. Variation, whether diatopic, diachronic, diastratic, or diaphasic, is frequently overlooked, resulting in general LLMs and machine translation systems failing to account for this diversity (Joshi et al., 2024). This issue, described as *translationese* by Volansky et al. (2015), can negatively impact the very language communities these technologies aim to serve.

A unique challenge shared by most Creole languages stems from their origins and ongoing language contact situations. Creole-speaking communities often exist in environments of constant interaction with another language. This contact leads to significant linguistic interference, manifesting as diglossia in some contexts or bilingualism in others. The propensity for interlingual interference results in phenomena like code-switching, borrowing, and other forms of multilingual restructuring. These dynamics highlight the critical need to incorporate linguistic variation into NLP research for Creole languages, ensuring that technologies reflect their rich diversity and complex sociolinguistic realities.

One domain where language access is critical is healthcare. Haitian Creole, the most widely spoken Creole language, is vastly underrepresented in NLP, contributing to severe healthcare disparities for Haitian Creole-speaking communities in multilingual environments like Miami. In these settings, patients frequently switch between Haitian Creole, French, English, and Spanish, a phenomenon that existing translation systems fail to handle effectively. Additionally, formal medical discourse differs significantly from everyday conversational Haitian Creole, further complicating automatic translation efforts.

The lack of medical translation tools tailored to Haitian Creole leads to miscommunication between healthcare providers and patients, which has been linked to misdiagnoses, non-compliance with treatment plans, and preventable health complications. Addressing this issue requires NLP models that accurately capture Creole linguistic variation, handle multilingual and code-switched text, and integrate standardized medical terminology—none of which are adequately covered by current Haitian Creole language models.

This work proposes a structured methodology for developing an AI-assisted translation and interpretation tool specifically designed for healthcare communication. Our approach prioritizes linguistic variation, code-switching, and domain-specific adaptation to create a culturally and context-sensitive NLP system.

To achieve this, we:

1. Develop strategies to collect domain-specific data, leveraging community engagement, partnerships with local organizations, and web scraping while adhering to ethical and legal guidelines for medical data.

2. Design advanced annotation methods, involving linguists, medical professionals, and native speakers to ensure accurate and culturally appropriate translations.

3. Adopt a hybrid NLP approach, integrating fine-tuned Large Language Models (LLMs) with traditional machine translation methods and Retrieval-Augmented Generation (RAG) to handle complex sentence structures and specialized medical language.

4. Define evaluation metrics that assess linguistic variety, code-switching accuracy, and domain adaptation performance while incorporating human evaluation to measure real-world usability.

By addressing these linguistic and computational challenges, this project contributes to both NLP research and healthcare equity. It also provides a scalable framework for other low-resource languages facing similar issues in medical translation, multilingual communication, and linguistic variation.

2 Background and Related Work

Haitian Creole exhibits significant linguistic variation, including basilectal and mesolectal varieties influenced by French and other languages. The basilect-mesolect-acrolect continuum in Creole-speaking territories describes the range of language varieties, from the most Creole-like variety (the basilect) to the variety most closely resembling the European lexifier language (in this case, French), referred to as the acrolect. The mesolect, or mesolectal zone, serves as an intermediary area encompassing a blend of phonological, lexical, morphosyntactic, and semantic features from both the basilect and acrolect (Bernabé, 1982).

In this context, linguists have characterized the mesolect as containing a Creole-based variety influenced by the acrolect, sometimes described as a "Frenchified Creole." In Haiti, this variety is known as Kréyòl swa (Tezil, 2022). Conversely, the continuum also includes a local French variety influenced by the basilect, often termed "Creolized French." Of particular interest to this work is the relationship between the basilectal variety, known as Krèyòl rèk, and Kréyòl swa within the linguistic continuum.

Krèyòl rèk is predominantly spoken by monolingual Haitian Creole speakers in Haiti and possesses distinctive features that set it apart from Kréyòl swa, which is primarily used by bilingual Haitian Creole-French speakers in Haiti and its diaspora. Due to their numerous structural differences, these two varieties need to be treated as two distinct linguistic units. Krèyòl rèk is often associated with lower prestige and is viewed as the most authentic representation of the basilectal variety, while Kréyòl swa carries higher prestige due to its proximity to French. These dynamics reflect deeper sociolinguistic patterns tied to language, identity, and power in Haitian society (Tezil, 2022; Tézil, 2024).

Among NLP initiatives and LLM developments for Haitian Creole and other Creole languages, several notable contributions stand out. Lent et al. (2024) introduced Creoleval, a multilingual benchmark for Creole languages and Lent et al. (2022) proposed guidelines for developing NLP technologies for Creole languages. Older but equally important initiatives include the Haitian Creole language data by Carnegie Mellon [1], which contains medical domain phrases and sentences; the Universal Dependencies (UD) Haitian Creole Autogramm Treebank (Jagodzińska et al.)[2], with sentences sourced from the Bible, novels, and newspapers; and the Leipzig Corpora Collection[3] a scraped Haitian Creole corpus primarily composed of Wikipedia articles. Additionally, mainstream multilingual LLMs, such as mBERT (Devlin et al., 2018), XLM-R (Conneau et al., 2019), mT5 (Xue et al., 2020), and M2M-100 (Fan et al., 2020), include Haitian Creole as part of their training data.

Despite these contributions, no existing model sufficiently addresses the specific challenges of our task. Variation is a critical component of NLP tasks for Haitian Creole due to the complexity of its linguistic landscape, which spans multiple varieties and contact languages. Current models often fail to encompass the linguistic diversity of Haitian Creole, whether within Haiti or the broader diaspora. While linguistic variation can theoretically be "learned" by systems through extensive data, the scarcity of annotated resources for underrepresented languages like Haitian Creole renders this approach ineffective.

Another significant limitation of existing models is their lack of robust accuracy in handling code-switching effectively for language pairs containing low-resource languages in particular (Çetinoğlu et al., 2016; Sitaram et al., 2019; Winata et al., 2022). This issue is also particularly pronounced in multilingual environments like Miami, where Haitian Creole speakers frequently switch between Haitian Creole, French, English, and Spanish.

Finally, there is a lack of resources and models specifically designed for Haitian Creole in domain-specific contexts, such as healthcare. This gap is compounded by broader challenges in extending Creole languages beyond their traditionally established functions, particularly in scientific and technical domains. A notable effort in this regard is the MIT-Ayiti lab's initiative to create new vocabulary for STEM materials[4]. However, this project faced criticism from linguists for its reliance on the lexifier language (French) to generate new Haitian Creole terms, which sparked debates about linguistic authenticity and community acceptance[5].

This project therefore requires a series of targeted steps to address the multifaceted challenges of developing an accurate and culturally sensitive machine translation (MT) model for Haitian Creole speakers in Miami's healthcare context.

[1] http://www.speech.cs.cmu.edu/haitian/

[2] https://github.com/UniversalDependencies/UD_Haitian_Creole-Autogramm - Accessed:2024-12-07

[3] https://corpora.uni-leipzig.de?corpusId=hat_community_2017 - Accessed:2024-12-07

[4] https://haiti.mit.edu/glossaryglose/

[5] https://rezonodwes.com/?p=314768

By leveraging existing research and creating new task-specific NLP resources, we aim to tackle the following critical issues:

1. Addressing linguistic variation in Haitian Creole to ensure the model can encode and decode language reflective of the target community's usage. This includes accurately representing Kreyòl swa and Kreyòl rèk varieties.

2. Managing code-switching in this multilingual environment. The model must handle (a) frequent switching between Haitian Creole and French, and (b) more complex switching among Haitian Creole, English, and Spanish, which is common in Miami's diverse linguistic landscape.

3. Translating specialized medical terminology accurately, which is crucial to facilitating effective communication between patients and healthcare providers. This requires not only linguistic precision but also cultural sensitivity.

4. Involving Haitian Creole-speaking communities in Miami throughout the development process is key to ensuring cultural relevance and linguistic authenticity. This includes collaboration with healthcare professionals, linguists, and native speakers to guide resource creation, annotation, and evaluation.

5. At a later stage, prioritizing ethical issues, such as ensuring patient confidentiality and addressing potential biases in the MT model. Additionally, practical concerns, such as deploying the model in real-time healthcare settings, should be addressed to ensure usability.

To achieve these objectives, we propose evaluating existing and new models across the following tasks:

- Task 1: Classification and identification of Kreyòl swa and Kreyòl rèk varieties.

- Task 2: Accuracy in producing texts in Kreyòl swa and Kreyòl rèk.

- Task 3: Language identification for code-switched texts, specifically Haitian Creole-French and Haitian Creole-English-Spanish.

- Task 4: Domain-specific machine translation for Haitian Creole-English and Haitian Creole-Spanish.

- Task 5 : Context-aware evaluation to ensure that translations align with cultural norms and healthcare-specific needs in real-world situations.

By integrating these steps, the project aims to address linguistic, cultural, and practical challenges, ensuring the resulting MT model is not only accurate but also relevant and beneficial to the Haitian Creole-speaking community in Miami's healthcare system.

3 Methodology and Guidelines

3.1 Linguistic Variation and Code-Switching

Addressing linguistic variation requires collecting data that represent the diverse varieties of Haitian Creole for classification tasks. Table 1 outlines existing corpora that form the basis of our investigation.

Corpora	Genre	Quantity
(Munro, 2010)	SMS	80k messages
CMU (1997-1998)	multi	2k sentences, 33k tokens, 1.2m. words
UD-HC	multi	144 sentences, 3k tokens
Leipzig	Wikipedia	23k sentences, 32k tokens, 290k words

Table 1: Corpora for Haitian Creole

While these corpora represent valuable resources, they provide limited coverage of Haitian Creole's linguistic diversity. For example, Munro (2010) compiled an 80k SMS corpus translated into English, offering insights into informal, casual Creole. However, its spelling has been normalized by the authors, diverging from Haiti's standard orthographic norms. For instance, Lewis (2010) noted the alternation between the personal pronouns *mwen* and *m* as reflecting high and low registers, respectively. The corpus homogenized this feature by replacing all instances of *m* with *mwen*, thereby prioritizing the high register. However,Valdman (2015) attributes this variation to phonological processes or free variation rather than solely register distinctions. This demonstrates the need to include linguistics research and developments in NLP.

The CMU corpus encompasses multiple genres, including novels, political speeches, and training manuals, and provides parallel Haitian Creole-English texts, including a medical domain subset. It also includes audio recordings of 150 Haitian Creole speakers from diverse locations (Pittsburgh, New York City, and Paris) recorded in 1997-1998 while reading various texts. However, it does not offer authentic oral data that can adequately represent diaphasic and diastratic variation.

The UD-HC treebank contains annotated data from literature and newspapers, providing part-of-speech, lemma, and dependency information. However, its limited size —144 sentences and 3k tokens—limits its scalability to more genres or everyday language use.

Lastly, the Leipzig corpus consists of 290k words scraped from Haitian Creole Wikipedia articles. While useful for understanding formal and encyclopedic language, it lacks representation of informal or spoken varieties.

Overall, the existing freely available corpora each have their strengths and limitations, but none explicitly represent the distinctions between Kréyòl rèk and Kréyòl swa—whether in written or spoken form—or include instances of code-switching. Despite these limitations, these corpora will provide a valuable and foundational baseline for training and fine-tuning multilingual language models to account for linguistic variation and code-switching in Haitian Creole.

Now looking specifically at the medical field, we collected pedagogical and instructional materials meant to facilitate patient-provider communication and information sharing (see examples in Figures 1 and 2 from EMSC (2023) and USSAAC (2023)). These documents provide, most of the time, a medical term in English and its equivalent in Haitian Creole, with or without the support of pictures. These resources have clear limitations as they show very limited and simplified medical terminology and therefore fail to represent the vast diversity of communicative situations a patient and a provider might find themselves in.

3.2 Data Collection and Augmentation Methods

Due to the limitations of the existing corpora of Haitian Creole, data collection and augmentation will be a necessary step to this project. For this, we will engage linguists, educators, healthcare professionals, and community leaders to help with data collection, ensuring ethical representation, and aligning linguistic standardization efforts with community needs. Their expertise may also help distinguish Kréyòl rèk from Kréyòl swa and refine domain-specific terminology. To address the specific needs of this project, we outline four key strategies for augmenting the existing corpora.

3.2.1 Community Engagement

Engaging with the Haitian Creole-speaking community and the linguistics community is essential for ensuring the cultural relevance and linguistic authenticity of the collected data. Community-driven initiatives such as focus groups, surveys, and storytelling workshops can help capture linguistic nuances that might otherwise go undocumented. For instance, via oral interviews, we propose collecting data on regional phonological and syntactic variations and documenting informal language use and code-switching patterns in real-life scenarios. Via crowdsourcing, we aim to draw on successful methods like those from Abraham et al. (2020), where mobile applications and community events can be employed to gather diverse speech samples, particularly from underrepresented speakers. Collaborating with the community also fosters trust and ensures that the data collected reflect the use of the language in the real world.

3.2.2 Web Scraping

Web scraping serves as a complementary strategy to gather written data from online sources such as blogs, forums, social media, and news websites. The newly collected data shall update language use as of today to augment the data collected 10 to over 20 years ago. Platforms popular within the Haitian diaspora, especially those catering to Miami's multilingual community, are particularly valuable. These sources can provide insights into both formal registers, such as news articles, and informal registers, such as casual online discussions.

This will allow us to develop a model that is sensitive to spelling variations in everyday communications.

3.2.3 Collaborations with Local Organizations

Partnering with local organizations offers a practical and impactful avenue for gathering and evaluating domain-specific data. To train our model, rather than using direct patient-provider interactions, we will rely on publicly available health resources, including patient education materials, public health campaign documents, and instructional content developed specifically for the Haitian Creole-speaking community. We will collaborate with medical professionals and interpreters to validate terminology, ensuring that translated materials reflect the nuances of real-world medical discourse. This expert-validated data will also serve as feedback for reinforcement learning, allowing us to fine-tune Large Language Models (LLMs) by iteratively improving translations based on linguistic accuracy and domain relevance.

Beyond medical content, linguistic diversity will be reinforced by incorporating educational materials, children's literature, and oral narratives from schools and cultural institutions. These additional sources will provide valuable insights into age-specific language use, different speech registers, and regional variations within Haitian Creole.

3.2.4 Machine Learning Methods for Augmentation

The UD-HC treebank provides valuable syntactic insights into Haitian Creole and is a key resource for improving NLP models. However, its small size limits the ability of language models to generalize effectively, making data augmentation necessary for robust parsing. One effective method is to leverage structurally similar languages with larger datasets to enhance the parsing performance of a Creole-specific syntactic parser. This method was previously explored in Mompelat et al. (2022) for parsing Martinican Creole (MC), another French-based Creole closely related to Haitian Creole. The approach involved using UD-French treebanks in Fine-tuning and Multitask Learning methods to compensate for the lack of annotated Martinican Creole data, resulting in promising improvements in parsing performance.

For Haitian Creole, we propose a similar strategy, combining the UD-HC treebank with our UD-formatted Martinican Creole treebank while also leveraging existing UD-French treebanks. By applying multitask learning and fine-tuning techniques, we aim to enhance syntactic parsing accuracy, ensuring that models trained on Haitian Creole can generalize more effectively across diverse linguistic structures.

3.3 Annotation and Data Curation

Accurate and consistent annotation is fundamental for training effective NLP models, especially for low-resource languages like Haitian Creole. Given Haitian Creole's linguistic complexity—including its regional variations, code-switching phenomena, and diverse registers, careful curation and processing of available corpora are essential. This process involves cleaning, annotating, and standardizing data to ensure it can be effectively used for both training and evaluation tasks.

3.3.1 Annotation Process

Capturing linguistic nuances such as phonological variation, syntactic structures, and lexical distinctions requires the involvement of both linguists and native speakers for both written and spoken forms of Haitian Creole. To ensure consistency and reliability across datasets, we will develop annotation guidelines tailored specifically to Haitian Creole. These guidelines will integrate feedback from linguistic experts, native speakers, and community stakeholders to address the diversity and sociolinguistic dynamics of the language.

3.3.2 Data Cleaning and Preprocessing

To prepare the data for model training and fine-tuning, we will implement a multi-step cleaning and preprocessing pipeline that will include a normalization stage to resolve inconsistencies in spelling, punctuation, and capitalization across datasets. This step is particularly important for reconciling informal and formal as well as interlectal variations in the language. Data will also be annotated for Part-of-Speech (POS) and dependency Parsing. This will help leverage existing tools in performing tasks that require detailed syntactic understanding, such as those involved in code-switching decoding and domain-specific utterance decoding and encoding.

With the current datasets available, this will include the augmentation of the UD treebank corpus, the normalization of the parallel text corpus Haitian-English by the CMU, and the transcription of the oral corpus to be collected within the community.

To ensure the reliability and utility of the preprocessed data, annotators will undergo rigorous training to minimize errors and adhere to standardized annotation guidelines. We will regularly use calculated metrics to assess consistency among annotators and identify areas needing further clarification or refinement. Finally, a continuous feedback system between linguists and annotators will address ambiguities in the data and refine annotation practices over time.

4 Modeling Approach

The modeling approach for this project builds upon recommendations from Zampieri et al. (2020), combining traditional machine learning methods with modern transformer-based techniques. They point out that traditional classifiers, such as support vector machines (SVMs), have proven effective in distinguishing closely related languages but that advancements in contextual embedding models, particularly BERT, have outperformed traditional methods in tasks requiring nuanced language understanding.

For Haitian Creole, multilingual transformer-based models (e.g., mBERT, XLM-R) offer significant potential to handle linguistic complexity and code-switching. This section outlines strategies to adapt and fine-tune these models to the unique challenges of Haitian Creole in healthcare contexts.

4.1 Leveraging Large Language Models

LLMs based on architectures like GPT have proven particularly effective for language generation tasks driven by a given query or prompt. These models, trained on vast datasets, excel in language understanding, generation, and even reasoning and have been used to create synthetic data used to fine tune models (Long et al., 2024). Advances in Retrieval-Augmented Generation (RAG) have further enhanced their utility, allowing general-purpose models to be specialized for specific tasks and domains, such as those encountered in the medical sphere (Amugongo et al., 2024; Yu et al., 2024; Anandavally, 2024). For instance, Wang et al. (2023) propose a framework to align LLMs with conversational patterns characteristic of medical consultations, enabling models to generate domain-specific, context-aware responses. This strategy provides a pathway for designing models that are not only accurate in their domain knowledge but also culturally sensitive in their output.

The goal of the Haitian Creole translator/interpreter model is to deliver contextually appropriate and linguistically accurate responses to queries, particularly when bridging Haitian Creole and other languages like English and Spanish in healthcare communication scenarios. This requires a model capable of translating and interpreting domain-specific content accurately while addressing linguistic nuances and sociolinguistic dynamics.

The application of LLMs in medical contexts has already demonstrated promising results across various use cases. For example, models have been employed to assist in diagnostics and provide clinical decision support, yielding improved outcomes in patient care (Nazary et al., 2024). LLMs have also been fine-tuned to offer medical diagnostic advice and personalized patient information (Panagoulias et al., 2024). Finally, frameworks aligning LLMs with medical consultation scenarios have successfully captured the nuances of patient-provider interactions, enhancing the relevance and accuracy of generated responses (Wang et al., 2023).

While these advancements lay a strong foundation, the specific linguistic and sociolinguistic characteristics of Haitian Creole require specialized adaptations of LLMs. Pre-trained multilingual models such as BERT, GPT, XLM-R, and mT5 will be fine-tuned on Haitian Creole corpora. This adaptation allows the models to capture unique linguistic features, including morphosyntactic patterns, phonological distinctions, and lexical variations inherent to Haitian Creole. By combining the generative capabilities of LLMs with retrieval mechanisms, the model will integrate external domain-specific knowledge. This includes medical terminology, patient-provider communication conventions, and sociolinguistic context, ensuring responses are both accurate and culturally appropriate. To address the challenges of healthcare communication, the model will be trained on authentic and synthetically generated scenarios requiring high precision in translation between Haitian Creole and English or Spanish. This includes translating specialized medical terms and interpreting patient narratives or provider instructions.

4.2 Handling Code-Switching

Handling code-switching effectively is essential for building a translator and interpreter model that aligns with the linguistic realities of Haitian Creole speakers. This capability is particularly important in multilingual healthcare settings, where accurate understanding and translation of mixed-language input can directly impact patient outcomes.

By addressing code-switching through tailored datasets and fine-tuned multilingual architectures, this project not only advances NLP for Haitian Creole but also contributes to the broader field of multilingual NLP by providing scalable solutions for similar low-resource languages and mixed-language contexts. Strategies involve incorporating loss functions that emphasize language boundary detection and coherence, ensuring that the embeddings capture the relationships between languages, particularly between Haitian Creole and French, and including examples from healthcare and other formal domains to improve the model's performance in professional contexts.

5 Evaluation and Mitigation of Bias in Domain-Specific Tasks

Ensuring fairness and accuracy in NLP models tailored for Haitian Creole, particularly in domain-specific tasks like healthcare communication, requires a comprehensive evaluation framework. This framework must address linguistic variation, code-switching, and the unique demands of domain-specific applications. By carefully designing evaluation criteria and incorporating iterative improvements, this section outlines a strategy to assess model performance while identifying and mitigating biases that may affect the utility and inclusivity of the system.

5.1 Evaluation Metrics

To evaluate linguistic variety, the dataset must include basilectal forms such as Kréyòl rèk and mesolectal forms like Kréyòl swa. Evaluation metrics in this context should assess the model's ability to accurately recognize and process these distinct varieties. Precision and recall metrics should be employed to determine how well the model identifies key linguistic features unique to each variety, while qualitative assessments should gauge the naturalness and cultural appropriateness of outputs.

For code-switching contexts, the dataset must incorporate authentic instances of language switching between Haitian Creole and other languages, particularly French, English, and Spanish, as these are the most commonly intertwined in multilingual settings like Miami. Evaluation metrics here should measure the coherence and fluency of the model's outputs when processing mixed-language inputs. BLEU and METEOR scores can quantify translation quality in these contexts, while human evaluators can provide insights into the semantic and syntactic coherence of the outputs.

In addressing registers, the dataset should span a range of formal and informal language uses. Formal registers may include medical documents or professional communications, while informal registers could consist of conversational Haitian Creole found in social interactions or casual settings. The evaluation for registers should measure the model's ability to align outputs with the expected level of formality or informality. Metrics like domain-specific accuracy and register appropriateness scores can help quantify the model's adaptability across varying communication styles.

5.2 Mitigation bias

Haitian Creole speakers, particularly those from diverse sociolinguistic backgrounds, will play a central role in the evaluation process. Regular consultations with community members will help identify biases that may not be apparent through automated metrics alone. For example, the model's treatment of linguistic variation, such as its handling of Kréyòl rèk versus Kréyòl swa, will be closely examined for equitable representation. Healthcare professionals, linguists, and cultural experts will provide critical insights to ensure that the model aligns with real-world usage patterns, particularly in sensitive contexts like medical communication. Their feedback will help refine the system to avoid potentially harmful inaccuracies or cultural missteps.

6 Scalability and Generalization to NLP Field

To ensure the scalability and adaptability of the methodologies developed, the model must be tested with Haitian Creole-speaking populations in various contexts, including Haiti, the wider Caribbean, and diaspora communities across North America and beyond.

Testing across these diverse linguistic and cultural environments will help validate the tool's flexibility in capturing regional and sociolinguistic nuances. The outcomes of this testing will also provide valuable insights into the scalability of the framework to other under-resourced languages. Many such languages share challenges similar to those faced by Haitian Creole, such as limited availability of annotated datasets, significant regional variation, and a lack of domain-specific corpora.

7 Conclusion

7.1 Contributions to NLP

By demonstrating the effectiveness of these approaches for Haitian Creole, this research shows the blueprints for a replicable framework for addressing these issues in other low-resource languages. This generalization is particularly important for the global NLP field, as it paves the way for scalable solutions that can address linguistic diversity and underrepresentation on a larger scale. By prioritizing inclusivity and contextual accuracy, this project seeks to inspire advancements in multilingual NLP, empowering researchers and communities worldwide.

7.2 Future Work

To achieve the large-scale objectives of this project, the next phases will focus on parallel priorities: (1) expanding data collection and (2) developing hybrid NLP experiments to determine the most effective methods given the available data. Running these two priorities simultaneously will allow for progressive model refinement and scalable dataset expansion, ensuring that each iteration improves both real-world and synthetic data quality. Our proposed timeline is as follows:

1. Short-term (0-12 months):
    - Collect additional data through community engagement, web-based sources, and expert annotation to increase linguistic coverage across Haitian Creole varieties.
    - Develop and evaluate hybrid NLP models, comparing traditional machine learning approaches with fine-tuned LLMs
    - Generate initial synthetic data to augment low-resource datasets, using real-world data to fine-tune LLMs and mitigate biases in synthetic outputs.
2. Mid-term (12-24 months):
    - Scale up the dataset by integrating validated synthetic data and iteratively improve data augmentation pipelines
    - Conduct user studies with Haitian Creole speakers and healthcare professionals to assess usability, cultural appropriateness, and translation accuracy.
3. Long-term (24+ months)
    - Deploy the AI-assisted translation tool in clinical settings, community health programs, and mobile applications.
    - Refine real-time translation capabilities, integrating adaptive learning mechanisms to continuously improve model accuracy based on new data and real-world usage.
    - Expand research to other Creole languages, applying the methodology to support low-resource language NLP beyond Haitian Creole.

To foster open research and collaboration, all datasets, fine-tuned models, and evaluation frameworks will be made publicly available, supporting ongoing advancements in NLP for Haitian Creole and other under-resourced languages.

Acknowledgments

I would like to express my sincere gratitude to Dr. David Tézil and Dr. Marie Denise Gervais for their invaluable contribution and shared dedication to this project. Their expertise and commitment to addressing linguistic issues and healthcare disparities are instrumental in shaping this work. I deeply appreciate their ongoing collaboration in advancing this important initiative.

References

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2819–2826.

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Geoffrey Brooks, Stefan Doering, and Jan Seidel. 2024. Retrieval augmented generation for large language models in healthcare: A systematic review.

Biju Baburajan Anandavally. 2024. Improving clinical support through retrieval-augmented generation powered virtual health assistants. *Journal of Computer and Communications*, 12(11):86–94.

Jean Bernabé. 1982. Contribution à une approche glottocritique de l'espace littéraire antillais. *La linguistique*, 18(Fasc. 1):85–109.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. *arXiv preprint arXiv:1610.02213*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Vissio Nicolás Cortegoso and Zakharov Viktor. 2021. Towards a part-of-speech tagger for sranan tongo. *International Journal of Open Information Technologies*, 9(12):99–103.

Raj Dabre and Aneerav Sukhoo. 2022. Kreolmorisienmt: A dataset for mauritian creole machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 22–29.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Florida EMSC. 2023. Medical Communication Cards 2023 Haitian Creole and English Version.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Marcell Fekete, Ernests Lavrinovics, Nathaniel Robinson, Heather Lent, Raj Dabre, and Johannes Bjerva. 2024. Leveraging adapters for improved cross-lingual transfer for low-resource creole mt. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 212–215.

Sandra Jagodzińska, Claudel Pierre-Louis, Sylvain Kahne, Agata Savary, and Emmanuel Schang. Le premier corpus arboré en créole haïtien. Accessed: 2024-12-05.

Aditya Joshi, Diptesh Kanojia, Heather Lent, Hour Kaing, and Haiyue Song. 2024. Connecting ideas in'lower-resource'scenarios: Nlp for national varieties, creoles and other low-resource scenarios. *arXiv preprint arXiv:2409.12683*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Heather Lent, Emanuele Bugliarello, Miryam De Lhoneux, Chen Qiu, and Anders Søgaard. 2021. On language models for creoles. *arXiv preprint arXiv:2109.06074*.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. *arXiv preprint arXiv:2206.00437*.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, et al. 2024. Creoleval: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

William Lewis. 2010. Haitian creole: How to build and ship an mt engine from scratch in 4 days, 17 hours, & 30 minutes. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Cécile Macaire, Didier Schwab, Benjamin Lecouteux, and Emmanuel Schang. 2022. Automatic speech recognition and query by example for creole languages documentation. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Ludovic Mompelat, Daniel Dakota, and Sandra Kübler. 2022. How to parse a creole: When martinican creole meets french. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4397–4406.

Ludovic Vetea Mompelat. 2023. *To Infinitive and Beyond, or Revisiting Finiteness in Creoles: A Contrastive Study of the Complementation Systems of Martinican Creole and Haitian Creole*. Ph.D. thesis, Indiana University.

Robert Munro. 2010. Crowdsourced translation for emergency response in haiti: the global collaboration of local knowledge. In *Proceedings of the Workshop on Collaborative Translation: technology, crowdsourcing, and the translator perspective*.

Fatemeh Nazary, Yashar Deldjoo, Tommaso Di Noia, and Eugenio di Sciascio. 2024. Xai4llm. let machine learning models and llms collaborate for enhanced in-context learning in healthcare. *arXiv preprint arXiv:2405.06270*.

Dimitrios P Panagoulias, Maria Virvou, and George A Tsihrintzis. 2024. Augmenting large language models with rules for enhanced domain-specific interactions: The case of medical diagnosis. *Electronics*, 13(2):320.

Neha Ramsurrun, Rolando Coto-Solano, and Michael Gonzalez. 2024. Parsing for mauritian creole using universal dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12622–12632.

Nathaniel R Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Bizon Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome A Etori, et al. 2024. Krey\ol-mt: Building mt for latin american, caribbean and colonial african creole languages. *arXiv preprint arXiv:2405.05376*.

Sarah Schieferstein. 2018. *Improving neural language models on low-resource creole languages*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Andrew Smart, Ben Hutchinson, Lameck Mbangula Amugongo, Suzanne Dikker, Alex Zito, Amber Ebinama, Zara Wudiri, Ding Wang, Erin van Liemt, João Sedoc, et al. 2024. Socially responsible data for large multilingual language models. *arXiv preprint arXiv:2409.05247*.

David Tezil. 2022. On the influence of kreyòl swa: Evidence from the nasalization of the haitian creole determiner/la/in non-nasal environments. *Journal of Pidgin and Creole Languages*, 37(2):291–320.

David Tézil. 2024. Sociolinguistic challenges and new perspectives on determining french speakers in creole communities: the case of haiti. *International Journal of the Sociology of Language*, 2024(288):177–207.

USSAAC. 2023. Patient-Provider Bilingual Tools Haitian Creole and English Version.

Albert Valdman. 2015. *Haitian Creole: structure, variation, status, origin*. Equinox Sheffield.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Wen Wang, Zhenyue Zhao, and Tianshu Sun. 2023. Gpt-doctor: Customizing large language models for medical consultation. *arXiv preprint arXiv:2312.10225*.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2022. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. *arXiv preprint arXiv:2212.09660*.

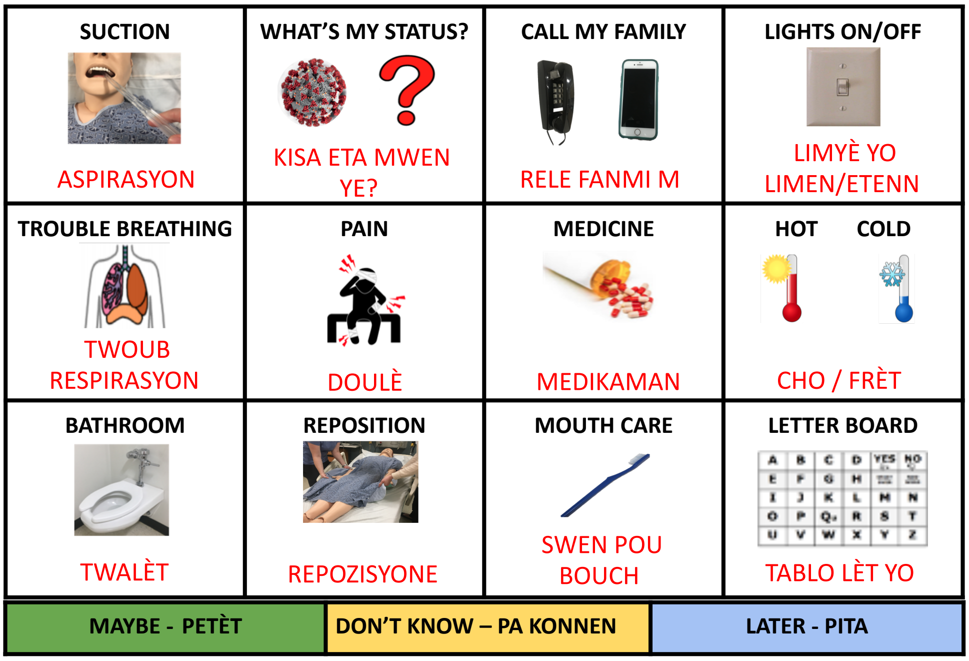
Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Xiang Li, Wenyue Hua, Mingyu Jin, Guang Chen, et al. 2024. Aipatient: Simulating patients with ehrs and llm powered agentic workflow. *arXiv preprint arXiv:2409.18924*.

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

A Appendix A: Example Patient-Provider Communication Cards

To illustrate the challenges of medical translation and register adaptation in Haitian Creole, we provide sample patient-provider communication cards below.



SUCTION ASPIRASYON	WHAT'S MY STATUS? KISA ETA MWEN YE?	CALL MY FAMILY RELE FANMI M	LIGHTS ON/OFF LIMYÈ YO LIMEN/ETENN
TROUBLE BREATHING TWOUB RESPIRASYON	PAIN DOULÈ	MEDICINE MEDIKAMAN	HOT COLD CHO / FRÈT
BATHROOM TWALÈT	REPOSITION REPOZISYONE	MOUTH CARE SWEN POU BOUCH	LETTER BOARD TABLO LÈT YO

MAYBE - PETÈT	DON'T KNOW – PA KONNEN	LATER - PITA

Haitian Creole General needs – 12+ target – photos & text

Figure 1: Example patient-provider communication cards in Haitian Creole and English for Adults.

For a more comprehensive set of Haitian-English medical communication cards, see USSAAC (2023).

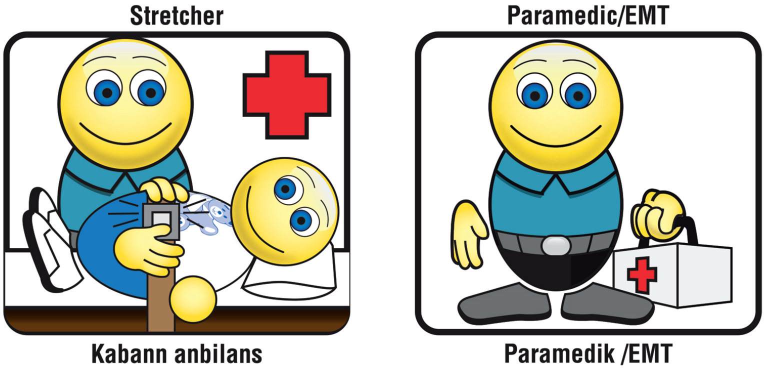


Figure 2: Example patient-provider communication cards in Haitian Creole and English for Children and Families.

For a more comprehensive set of Haitian-English medical communication cards, see EMSC (2023).