GuyLingo: The Republic of Guyana Creole Corpora

Anonymous NAACL submission

Abstract

While major languages often enjoy substantial attention and resources, the linguistic diversity across the globe encompasses a multitude of smaller, indigenous, and regional languages that lack the same level of computational support. One such region is the Caribbean. While commonly labeled as "English speaking," the ex-British Caribbean region consists of a myriad of Creole languages and dialects thriving alongside English. In this paper, we present **GuyLingo**: a comprehensive corpus designed for advancing NLP research in the domain of Creolese, the most widely spoken language in the culturally rich nation of Guyana. We first outline our framework for gathering and digitizing this diverse corpus, inclusive of colloquial expressions, idioms, and regional variations. We then demonstrate, alongside discussions with Creolese experts, the challenges of training and evaluating NLP models for machine translation for Creolese. Lastly, we discuss the unique opportunities presented by recent NLP advancements for accelerating the formal adoption of Creole languages in the Caribbean.

1 Introduction

Major languages such as English and Chinese frequently receive considerable attention and resources due to their global prominence and economic influence (Lent et al., 2021, 2022a). The extensive focus on these major languages in natural language processing (NLP) has resulted in the development of sophisticated models, extensive datasets, and digital applications consumed by millions of users today. However, despite this global prominence, the linguistic landscape of the globe extends far beyond these dominant languages, encompassing a plethora of smaller, indigenous, and regional languages that play crucial roles in the cultural heritage and communication of their respective communities (Lent et al., 2022c; Hershcovich et al., 2022). The Caribbean Community (CARICOM) is an example of one such region.

Figure 1: Map of Guyana and its neighboring territories

Within the diverse linguistic tapestry of the Caribbean Community, a rich array of languages thrives, reflecting the historical, cultural, and ethnic diversity of the region (Rickford, 1987; Holbrook and Holbrook, 2001). While English is commonly used as the official language in many CARICOM member states, the linguistic heritage goes beyond just English, encompassing a variety of Creole languages, indigenous languages, and influences from African, Indigenous, European, and Asian languages (Devonish and Thompson, 2013).

Creole languages, often born out of the historical context of slavery and colonialism, hold a significant place in the linguistic mosaic of the Caribbean. These languages, such as Jamaican Patois (Armstrong et al., 2022), Trinidadian Creole (Michaelis et al., 2013), and Haitian Creole (Hewavitharana et al., 2011), have evolved as vibrant means of communication, blending elements of African languages with those of European colonizers (Hagemeijer et al., 2014b).

Creole	English
“It luk laik nof ting cheenj op,” Seera se. “Somtaim mi doz get fraikn.”	“So many things feel like they have changed,” said Sara. “I get scared about it sometimes.”
When me lef’ han’ ’cratch me, money a-come	When my left hand itches, money is coming.
Di leedii prapa nais	The lady is very pretty

Table 1: Example Guyanese Creole from GuyLingo and its English Translation

Despite its prominence as the mother tongue of the majority of the over 700,000 inhabitants of the Republic of Guyana, Creolese is often subject to challenges in terms of recognition and preservation due to its co-existence with British English, the sole official language of the country. Like other Creole languages, this is partially due to the deeply rooted social attitudes and perceptions about language prestige, resulting in limited efforts to digitize, document, and promote Creolese in official and educational contexts (Hershcovich et al., 2022).

In this work, we introduce GuyLingo, a corpus for Guyanese Creole curated for advancing NLP research and development in Creole. Using these sources, we explore the machine translation task between English and Creolese. To aid in this process we design and implement the Guyanese Creole Translation tool[1], a web-based GPT-powered machine translation tool. Lastly, we briefly discuss insights gained from these developments for accelerating the formal adoption of Creole languages in the nation and potentially the Caribbean Region.

2 GuyLingo Corpus

This section describes the curation of GuyLingo, a corpus of Creolese, the primary spoken language of Guyana. The creation of this corpus aims to address the scarcity of resources and attention devoted to indigenous and regional languages within the NLP community.

2.1 Data Collection

The compilation of GuyLingo involves the collecting and digitizing of a series of linguistic resources encompassing a spectrum of Creolese expressions, idiomatic phrases, and regional variations. To ensure inclusivity and authenticity, we employ a multi-pronged approach:

2.1.1 Expert Collaboration

In collaboration with [removed for anonymity], a collection of original Guyanese Creole sources were curated, digitized, and manually transcribed by a team of researchers. Examples of this include Peirs (1902) a book of Guyanese proverbs, containing over 1k culturally rich proverbs from early British Guiana times still used today, and (Helen Patuck, 2020) a COVID-19 children’s book transcribed by Creolese experts for primary education students. In addition, our team of native Creole experts manually construct a corpus of high-quality common Guyanese Creole sayings and terms. Table 2 shows a full breakdown of all information sources.

2.1.2 Online Resources

While no nationally adopted writing system for Creolese exists, many web-based sources, such as language forums, blogs, educational platforms, etc., contain small excerpts of colloquialisms, everyday conversations, and idiomatic expressions prevalent in the Guyanese Creole. These sources were scraped, cleaned, verified, and added to GuyLingo as shown in 2.

2.2 Dataset Characteristics

GuyLingo encapsulates a diverse array of linguistic data, including but not limited to:

- Conversational dialogues
- Idiomatic expressions and phrases
- Proverbs and folklore
- Regional variations and dialectical nuances

In total, GuyLingo consists of 1969 Guyanese Creole sentences with a vocabulary size of 4177 unique Creole words.

3 GuyLingo for Machine Translation

To investigate the utility of GuyLingo, we conduct experiments in the setting of machine translation to assess the ability of NLP models to facilitate English←→Guyanese Creole translation. As such to enable the training and evaluation of these models GuyLingo was further expanded to include English Creole translation pairs. Of the 1969 sentences, the Common Guyanese Creole Saying

[1] https://github.com

Sources	Type	# Sentences	Vocab Size
Guyanese-Creole-English Vocabulary-Basic words. (Polyglot Club, Accessed 2023)	Corpus	20	71
Guyanese Creole. (Wikipedia, Accessed 2023)	Article	6	28
Gender and Pronominal Variation in an Indo-Guyanese Creole-Speaking (Sidnell, 1999)	Journal Article	20	82
Review of Guyanese Creole English (Guy, Accessed 2023)	Presentation	9	96
Guyanese Creole Survey Report. (Holbrook and Holbrook, 2001)	Language Survey	7	45
APiCS Online -Structure dataset. (Michaelis et al., 2013)	Report	176	351
Creolese. (Devonish and Thompson, 2013)	Journal Article	49	112
Habitual and Imperfective in Guyanese Creole. (Sidnell, 2002)	Journal Article	50	103
Tense and aspect in Guyanese Creole: A syntactic, semantic and pragmatic analysis (Gibson, 1982)	PhD Thesis	191	374
Two areas of Guyanese Grammar (Guyanese Languages Unit, 2016)	Article	14	26
Me Na Able: Creolese 101 (Letters from Guyana, 2017)	Blog	8	25
Travel Phrases - Guyanese Creole (Travel Phrases)	Blog	4	9
My Hero is you (Helen Patuck, 2020)	Educational	114	831
The Proverbs of British Guiana (Peirs, 1902)	Book	1060	2054
Common Guyanese Creole Sayings (Manually created by experts)	Corpus	302	712
Total		**1969**	**4177**

Table 2: Compilation of Guyanese Creole Language Resources: Sources, Type, Sentences, and Vocabulary Size

corpus was manually transcribed into English. In addition, 339 common Creole terms from (Peirs, 1902) alongside their English pairs we extracted and verified. Using these initial translation pairs, the Guyanese Creole Translation Tool was built to allow the initial translation of remaining sentences in GuyLingo.

3.1 Guyanese Creole Translation Tool

The Guyanese Creole Translation tool, as shown in figure 4, is a web-based application built using Django+React to facilitate easily storing, editing, and iterative testing of English Creole translations. The UI allows Creolese experts to easily enter text in English or Creole and get a sample translation. We utilize GPT-4 (OpenAI, 2023) to automatically perform these translations. The advanced prompt includes a subset of example verified translations from GuyLingo as in-context examples for generation. Once prompted, the user can modify the generated output before saving it to the database. Users also have the option to modify the advanced prompts as well as provide more seed examples for greater control over the translation process. For instance, users can provide a Guyanese proverb and instruct GPT4 to consider the nuances of the Guyanese culture while translating the text. The remaining translation pairs from GuyLingo were generated and verified using this tool, resulting in 1969 total translation pairs. For training and evaluation, we use our 302 manually curated translation pairs for testing and the remaining translation pairs for model training.

3.2 Experiment Setup

Training and Models We consider the models of T5 (Raffel et al., 2023), BART (Lewis et al., 2020) and Pegasus (Zhang et al., 2020) for their demonstrated performance on several machine translation tasks. All models were implemented with PyTorch and Hugging Face Transformers. We train all models with AdamW (Loshchilov and Hutter, 2019) and a weight decay of 0.01. We use a learning rate of 2e-5, batch size of 4, and a linear learning rate warmup over the first 10% steps with a cosine schedule. We pre-process the data and train all models with varying random seeds over multiple runs for 10 epochs. Approximately 200 GPU hours were required to train all hyperparameter variations across all tasks.

Evaluation For automatic evaluation metrics, we adopted the common methods used for language generation based on n-gram overlap: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005).

3.3 Results

Table 3 and 4 summarize our evaluation results on automated metrics. For en-creole translation, our

Model	Bleu	Rouge1	Rouge2	RougeL	Meteor
T5-Large	09.74	37.44	13.74	36.63	28.19
MBart-Large	12.02	38.39	17.02	37.58	31.33
Bart-Large	**12.11**	**40.56**	**18.47**	**39.64**	**32.77**
Bart-Base	10.17	37.49	16.08	36.59	29.54
Pegasus-Large	02.67	24.15	05.30	23.16	16.38

Table 3: Performance of MT Models on English-Creole Translation

Model	Bleu	Rouge1	Rouge2	RougeL	Meteor
T5-Large	**19.70**	**47.71**	**26.99**	**46.47**	**42.45**
Bart-Large	17.70	45.74	24.41	39.75	32.77
Bart-Base	14.20	41.68	20.04	40.40	35.95
Pegasus-Large	6.10	28.96	09.88	27.91	22.53

Table 4: Performance of MT Models on Creole-English Translation

results show that the Bart-Large model achieves the best performance amongst all models with a BLEU score of 12.11, ROUGE-1 score of 0.41, ROUGE-2 score of 0.18, ROUGE-L score of 0.40, and METEOR score of 0.33. whereas on creole-eng translation, T5-Large achieves the best performance with a BLEU score of 19.70, ROUGE-1 score of 0.48, ROUGE-2 score of 0.27, ROUGE-L score of 0.46, and METEOR score of 0.42. From manual inspection of both settings, we find that the models struggle mainly with en-creole translation due to incoherent mapping of certain key creole terms to English.

4 Discussion

In this section, we briefly discuss the unique opportunity presented by recent NLP advancements for accelerating the formal adoption of Creole languages in the Caribbean.

4.1 AI-Driven Applications for Native Languages

One of the major issues affecting the formal adoption is Creolese despite its prominence as a spoken language is its lack of use in formal communication outlets such as literature, news, and written texts. AI-driven applications fueled by rich data sources such as GuyLingo present a major opportunity for enabling the development of educational content, legal documents, and official communications in Creolese. Figure 2 showcases a conversational AI Assistant named [removed for anonymity] deployed to citizens of Guyana speaking in Creolese fueled by GuyLingo. Such applications present the ability to make Creolese more accessible and applicable in various formal contexts further allowing citizens to feel empowered and foster a sense of national pride.

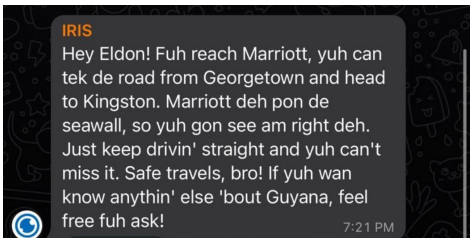

Figure 2: Conversational Agent in Whatsapp speaking in Guyanese Creole.

5 Related Works

In the context of linguistic diversity, prior works (Hershcovich et al., 2022; Lent et al., 2021, 2022a) have highlighted the challenges faced by lesser-known languages, emphasizing the importance of recognition and preservation. Works such as Dabre and Sukhoo (2022), Hagemeijer et al. (2014a), and Liu et al. (2022) have contributed to advancing NLP research in Creole languages by building a corpus of text for various Creole languages, fostering machine translation, and enhancing language modeling techniques specific to these linguistic varieties. Our work falls into this category. On the other hand, works such as Lent et al. (2022b) and Lent et al. (2022c) emphasize the importance of linguistic diversity by documenting the challenges and exploring the complexities of language modeling for underrepresented languages. The juxtaposition of these studies with the dominance of major languages in NLP underscores the need for more inclusive research efforts that consider the linguistic richness and cultural significance of smaller, indigenous languages within global technological advancements.

6 Conclusion

In this paper, we introduce GuyLingo, a corpus of Guyanese Creolese designed to facilitate advancements in NLP research. We discuss the process of gathering and digitizing this diverse corpus while highlighting the unique opportunities presented by recent NLP advancements for accelerating the formal adoption of Creole languages in the Caribbean. By providing access to a rich collection of colloquial language expressions, idioms, and regional variations, we hope to encourage further research in this field and improve the representation and understanding of Creole languages in NLP.

7 Limitations

While our work aims to contribute to the advancement of NLP for Creole, several limitations arise:

Limited Representation: Guyana is home to many languages outside of Creolese such as Wapichan, Makushi, Wai Wai, Akawaio, Arekuna, Patamuna, Kalina (Carib), Warrau, and Lokono to name a few. Given the cultural significance of these languages, future research should prioritize their inclusion to ensure a more inclusive and representative dataset. Additionally, The rich tapestry of languages in the region extends beyond Guyanese Creole, and efforts should be made to include additional Creole languages and dialects for a more comprehensive understanding.

Limited Generalizability: The findings and insights gained from our work, particularly regarding the formal adoption of Creole languages, may have limited generalizability to other regions or linguistic contexts.

Language Evolution: Creole languages, by their nature, are dynamic and subject to continuous evolution. The static nature of a curated corpus and machine translation models may not fully capture the evolving linguistic landscape, necessitating regular updates and adaptations to reflect current linguistic usage.

References

Accessed 2023. Guyanese creole english. Accessed on December 14, 2023.

Ruth-Ann Armstrong, John Hewitt, and Christopher Manning. 2022. JamPatoisNLI: A jamaican patois natural language inference dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5307–5320, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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.

Hubert Devonish and Dahlia Thompson. 2013. Creolese. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors, *The Survey of Pidgin and Creole Languages, Vol. I: English-based and Dutch-based languages*, pages 49–60. Oxford University Press, Oxford.

Kean Amelia Gibson. 1982. *Tense and aspect in Guyanese Creole: A syntactic, semantic and pragmatic analysis*. Ph.D. thesis, University of York.

Guyanese Languages Unit. 2016. Two areas of guyanese grammar. https://guyaneselanguagesunit.com/2016/07/12/two-areas-of-guyanese-grammar/. Accessed on December 14, 2023.

Tjerk Hagemeijer, Michel Généreux, IHE Hendrickx, Amália Mendes, Abigail Tiny, and Armando Zamora. 2014a. The gulf of guinea creole corpora.

Tjerk Hagemeijer, Michel Généreux, Iris Hendrickx, Amália Mendes, Abigail Tiny, and Armando Zamora. 2014b. The Gulf of Guinea creole corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 523–529, Reykjavik, Iceland. European Language Resources Association (ELRA).

Helen Patuck. 2020. My Hero is You: How Kids Can Fight COVID-19! https://www.unicef.org/coronavirus/my-hero-you. Accessed on: Insert Date Accessed.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural nlp.

Sanjika Hewavitharana, Nguyen Bach, Qin Gao, Vamshi Ambati, and Stephan Vogel. 2011. CMU Haitian Creole-English translation system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 386–392, Edinburgh, Scotland. Association for Computational Linguistics.

David J Holbrook and Holly A Holbrook. 2001. Guyanese creole survey report. *https://www.sil.org/resources/archives/9001*.

Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. 2021. On language models for creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 58–71, Online. Association for Computational Linguistics.

Heather Lent, Emanuele Bugliarello, and Anders Søgaard. 2022a. Ancestor-to-creole transfer is not a walk in the park. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 68–74, Dublin, Ireland. Association for Computational Linguistics.

Heather Lent, Emanuele Bugliarello, and Anders Søgaard. 2022b. Ancestor-to-creole transfer is not a walk in the park. *arXiv preprint arXiv:2206.04371*.

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

Letters from Guyana. 2017. Me na able - creolese 101. https://lettersfromguyana.wordpress.com/2017/01/29/me-na-able-creolese-101/. Accessed on December 14, 2023.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhengyuan Liu, Shikang Ni, Aiti Aw, and Nancy Chen. 2022. Singlish message paraphrasing: A joint task of creole translation and text normalization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3924–3936.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013. *APiCS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

James Peirs. 1902. *The Proverbs of British Guiana. With an Index of Principal Words, an Index of Subjects, and a Glossary*. The Argosy Company, Demerara.

Polyglot Club. Accessed 2023. Guyanese creole english vocabulary - basic words. https://polyglotclub.com/wiki/Language/Guyanese-creole-english/Vocabulary/Basic-words. Accessed on December 14, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

John R Rickford. 1987. *Dimensions of a creole continuum: History, texts, and linguistic analysis of Guyanese Creole*. Stanford University Press.

Jack Sidnell. 1999. Gender and pronominal variation in an indo-guyanese creole-speaking community. *Language in Society*, 28(3):367–399.

Jack Sidnell. 2002. Habitual and imperfective in guyanese creole. *Journal of pidgin and creole languages*, 17(2):151–189.

Travel Phrases. Guyanese phrases and basics. http://www.travelphrases.info/languages/guyanese.htm. Accessed on December 14, 2023.

Wikipedia. Accessed 2023. Guyanese creole. https://en.wikipedia.org/wiki/Guyanese_Creole. Accessed on December 14, 2023.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

A Guyanese Creole Translation Tool

In this section, we further showcase the Guyanese Creole Translation tool detailing our prompts and user interface.

```
Translate the following Guyanese Creole text
and provide the resulting English translation.
Please ensure that the translation is clear and
accurate. Guyanese Creole is spoken in Guyana
and may include unique vocabulary and grammar.
Try to capture the original meaning while
 making it comprehensible in English.

 Glossary:
   English: Swallow
   Creole: Swalla

   English: Stagger
   Creole: 'Taggah

   English: Stop-off
   Creole: "Taff-aff
   ...
 Translations
 Translation 1: The beef cooked until it was soft
 Text 1: Di biif kuk kuk kuk til ii saaf

 Translation 2: But my grandfather had a boat
 Text 2: Bo mi granfaada bin ga wan boot
```

Figure 3: Example GPT-4 Prompt with translation examples from Peirs (1902).

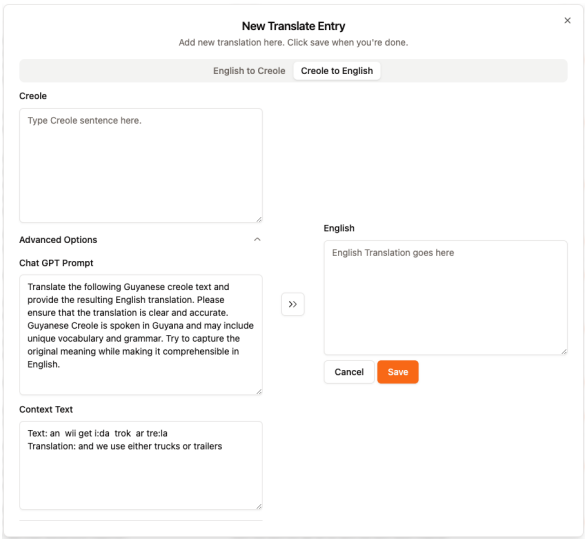

Figure 4: User Interface of Guyanese Creole Translation Tool. This tool allows experts to rapidly and iteratively create translation pairs using GPT-4 (OpenAI, 2023) as a generator.