INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Creating NLP resources for Cape Verdean Creole: Challenges and Solutions for Underrepresented Languages

Roberto Carlos Medina

Master in Data Science

Supervisor:
Doctor Fernando Batista, Associate Professor with Habilitation,
Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
Doctor Estanislau Lima, Assistant Professor,
UTA – Atlantic Technical University, Mindelo Cabo Verde

September 2025

BUSINESS SCHOOL

TECHNOLOGY AND ARCHITECTURE

Department of Quantitative Methods for Management and Economics
Department of Information Science and Technology

Creating NLP resources for Cape Verdean Creole: Challenges and Solutions for Underrepresented Languages

Roberto Carlos Medina

Master in Data Science

Supervisor:
Doctor Fernando Batista, Associate Professor with Habilitation,
Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
Doctor Estanislau Lima, Assistant Professor,
UTA – Atlantic Technical University, Mindelo Cabo Verde

September 2025

This work is dedicated to my family, to all the contributors who helped create the dataset, to everyone who will use this research to advance the inclusion of Cape Verdean Creole into the NLP landscape, and, more than anything, to all Cape Verdeans.

Acknowledgment

I would like to express my deepest gratitude to my supervisors, Doctor Fernando Batista, and Doctor Estanislau Lima, for their invaluable guidance, constant support, and encouragement throughout the development of this thesis. Their expertise and mentorship have been fundamental to the completion of this work.

I would also like to thank all those who contributed to this project, both directly and indirectly, for their collaboration and assistance along the way.

Finally, I am profoundly grateful to my family for their unconditional love, patience, and support, which have sustained me throughout this journey.

Roberto Carlos Medina

Resumo

O crioulo cabo-verdiano é a língua mais falada em Cabo Verde, mas tem sido e continua a ser muito pouco trabalhado em processamento de linguagem natural. Esta tese oferece pesquisa em processamento de linguagem natural para o crioulo cabo-verdiano com o Morabeza Corpus, uma versão expandida e melhorada do CrioleSet [1], dois corpora paralelos que incluem traduções entre o crioulo cabo-verdiano, inglês, português e francês, com cobertura dialetal aprimorada e pré-processamento abrangente, o que representa uma das primeiras pesquisas confiáveis em processamento de linguagem natural para o crioulo cabo-verdiano. Com base nesses conjuntos de dados paralelos, o estudo explora múltiplos métodos para Tradução Automática, começando com sistemas de base e buscando aprendizagem por transferência e sistemas multilíngues pré-treinados e aprimorados. Os resultados indicam que o mBART-50 ajustado no Morabeza Corpus apresenta bom desempenho, com pontuações BLEU superiores a 80, pontuação chrF próxima a 90 e baixa taxa de edição de tradução, enquanto que as avaliações manuais apresentaram altos níveis de fluência, idiomaticidade e fidelidade semântica. Além disso, a tese propõe uma estrutura automatizada para classificação de variantes dialetais para o crioulo cabo-verdiano, produzindo níveis aceitáveis de precisão, demonstrando a capacidade dos métodos computacionais de capturar a diversidade linguística codificada internamente em uma língua. A tese também fornece uma autoanálise produtiva dos potenciais desafios que podem surgir, como escassez de dados, ortografia variável e recursos computacionais insuficientes, que podem informar possibilidades de examinar dados e moldar pesquisas futuras. No geral, esta tese demonstra que desenvolver recursos sólidos e alcançar desempenho competitivo em processamento de linguagem natural para o crioulo cabo-verdiano é algo alcançável e significativo, o que tem contribuído significativamente para a inclusão digital do crioulo cabo-verdiano no mundo mais amplo da processamento de linguagem natural.

Palavras Chave: *Crioulo cabo-verdiano*, *Processamento de Linguagem Natural*, *Tradução Automática*, *datasets*

Abstract

Cape Verdean Creole is the most commonly spoken language in Cape Verde, it has been and continues to be, severely under-researched in natural language processing. This thesis provides natural language processing research for Cape Verdean Creole with Morabeza Corpus, an expanded and improved version of CrioleSet [1], both two parallel corpora for that includes translations pairs between Cape Verdean Creole, English, Portuguese and French, with broader dialectal coverage and thorough preprocessing, which represents some of the first reliable research for natural language processing in Cape Verdean Creole. Based on these parallel datasets, the study explores multiple methods for machine translation starting with baseline systems and pursuing transfer learning and fine-tuned multilingual pre-trained systems. The results indicate that mBART-50 fine-tuned on Morabeza Corpus delivered state-of-the-art performance with BLEU scores exceeding 80, chrF score near 90, and low translation edit rate, while manual raters showed high levels of fluency, idiomaticity, and semantic fidelity. In addition, the thesis proposes an automated framework for classification of Cape Verdean Creole dialectal variants yielding acceptable levels of accuracy, showing the capacity for computational methods to capture the linguistic diversity encoded internally in a language. The thesis also provides productive self-examination of potential challenges that may arise such as scarcity of data, variable orthography, and poor computational resources which may inform possibilities to examine data and shape further research. Overall, this thesis shows that developing strong resources and achieving competitive performance in natural language processing for Cape Verdean Creole is both achievable and significant, which has made an important contribution to the digital inclusion of Cape Verdean Creole into the wider world of natural language processing.

KEYWORDS: *Cape Verdean Creole*, *Natural Language Processing*, *Machine Translation*, *Datasets*

Contents

List of Figures

List of Tables

List of Acronyms

AI: Artificial Intelligence

BART: Bidirectional and Auto-Regressive Transformers

BERT: Bidirectional Encoder Representations from Transformers

BPE: Byte-Pair Encoding

CSV: Comma-Separated Values

CVC: Cape Verdean Creole

DL: Deep Learning

GPT: Generative pre-trained Transformer

GRU: Gated Recurrent Unit

INE: National Statistics Institute

LLM: Large Language Model

LRLs: Low-resource Languages

LSTM: Long Short-Term Memory

MLM: Masked Language Modeling

MT: Machine Translation

NER: Named Entity Recognition

NLLB: No Language Left Behind

NLP: Natural Language Processing

NMT: Neural Machine Translation

NSP: Next Sentence Prediction

PII: Personally Identifiable Information

TER: Translation Edit Rate

TTR: Type–Token Ratio

CHAPTER 1

Introduction

Cape Verde is a small island nation with just over half a million people living on it, Portuguese is the official language, and Cape Verdean Creole (CVC) is the first language of the vast majority of the population. Although it is used officially throughout the country, Portuguese is not the language used for most communication, CVC remains the dominant language in everyday communication and is widely spoken. However, despite being vastly used, it is extremely scarce in terms of language resources for Natural Language Processing (NLP), Artificial Intelligence (AI), and Deep Learning (DL).

Cape Verde is also a well-known tourist destination. Based on data released by the Cape Verdean National Statistics Institute (INE) [1], the hotel industry registered a total of 1,177,467 guests in 2024. Out of these, 60,517 were nationals (5.1%) and 1,116,950 were foreigners (94.9%). The largest foreign groups came from the United Kingdom (29.9% of entries, 35.1% of overnight stays), Germany (10.5% of entries, 10.5% of overnight stays), Belgium and the Netherlands combined (9.5% of entries, 9.2% of overnight stays), Portugal (9.4% of entries, 9.4% of overnight stays), and France (9.2% of entries, 6.4% of overnight stays). These figures highlight that the vast majority of hotel guests in Cape Verde are foreigners, implying that CVC is rarely the primary language of communication in tourism-related interactions. Instead, other languages, particularly English are more commonly used.

However, communicating in English is not always straightforward, since few Cape Verdeans have sufficient knowledge of the language. One solution could be bringing people in who are multilingual and training them to be translators, but this takes time and/or is expensive. A more scalable and sustainable solution is to employ Machine Translation (MT), to translate automatically between CVC and English (or another language). This could immediately address the communication barriers with things like tourism, education, public administration, or online content.

Despite the potential of NLP, building high-quality MT systems for CVC is challenging due to the lack of structured datasets, resources, software, and other linguistic tools. This gap puts CVC in the same situation as Low-resource Languages (LRLs), which have faced significant barriers to enter into modern NLP research. In addition to this, filling the gaps will help Cape Verde preserve its linguistic and cultural heritage and be part of a larger effort to help underrepresented languages with AI.

[1] INE - National Statistics Institute is the Cape Verdean national institute responsible for collecting national statistics in various domains. Tourism statistics for 2024 are available at `https://ine.cv/wp-content/uploads/2025/04/estatisticas-do-turismo-movimentacao-de-hospedes-ano-2024.pdf`

A recently peer-reviewed article [1] was published that introduced CrioleSet, a parallel CVC–English dataset with more than 6000 sentence pairs that can be utilized for MT and other NLP tasks. This thesis expands upon that work by supplementing the dataset, increasing linguistic diversity, and conducting NLP experiments such as MT and classification, as well as trying to address the barriers to processing LRLs.

1.1. Research Goals, Motivation, and Relevance

As previously stated, CVC serves as the primary mode of engagement for day-to-day communications throughout Cape Verde, but it is also seriously underrepresented in the realm of digital and computational communications. These imbalances demonstrate significant and pragmatic socio-economic and technological challenges. In relation to economic and cultural outcomes, the lack of development or delivery of automated translation between CVC and language systems that provide socioeconomic leverage, such as English, restricts opportunities for visitors to Cape Verde to access information easily, while also reducing opportunities to communicate across cultures.

From a NLP perspective, CVC is also an interesting case study and allows for a broader conversation about research considering LRLs, where the development and evaluation of datasets and models may assist in tackling urgent, world-wide global issues related to linguistic diversity in AI-based systems.

The primary objective of this work is to strengthen the computational presence of CVC by creating high-quality linguistic resources and applying modern NLP methodologies, expanding on the author's prior contribution.

The research is guided by the following questions:

- **RQ1:** Is it possible to collect a reliable and representative parallel corpus for CVC–English, which can be cleaned, and structured to support modern NLP applications?
- **RQ2:** What is the best performing MT approaches for CVC–English translation?
- **RQ3:** Can automatic classification methods be used to identify and differentiate between the two most relevant CVC variants within the dataset? How accurate are they?
- **RQ4:** What challenges and limitations arise when building NLP resources for CVC, and how can these be addressed to improve future research?

This work seeks to contribute to the preservation and modernization of CVC and the global conversation about inclusive, multilingual AI systems for minority languages by responding to these questions. Other initiatives, such as Orife *et al.* [2] collaborative NLP approaches for African languages and transformer-based methods for LRLs processing and cross-lingual generalization [3], [4] have shown the necessity and viability of developing scalable, high-quality resources for languages that are underrepresented, neglected, or excluded from NLP research.

This research is also a personal contribution toward ensuring that CVC is not left behind in the current wave of AI development. By integrating CVC into cutting-edge NLP workflows, this work seeks to place the language on equal footing with more resourced languages, fostering its visibility, usability, and preservation in digital spaces for future generations.

1.2. Document Structure

This document is organized as follows:

- **Chapter 1 – Introduction:** Presents the research context, questions, goals, motivation, and structure of the work.
- **Chapter 2 – Deep Learning Models:** Introduces the deep learning architectures employed to this work, outlining their relevance to NLP, MT, and classification tasks in low-resource language settings.
- **Chapter 3 – Related Work:** Reviews key concepts in NLP, MT, and LRLs processing applied to this work, and provides linguistic and cultural context for CVC.
- **Chapter 4 – The Morabeza Corpus:** Dataset Creation: Details the collection, preprocessing, and structuring of the Morabeza Corpus, along with associated challenges and ethical considerations.
- **Chapter 5 – Methodology:** Describes the methodologies used to implement and test the classification and MT tasks, including model architectures, training strategies, and evaluation methods.
- **Chapter 6 – Experiments and Results:** Reports experimental results for both classification and MT, compares results with different evaluations metrics, and includes analysis and discussion.
- **Chapter 7 – Conclusion:** Summarizes the main findings, discusses limitations, and outlines directions for future work.

CHAPTER 2

Deep Learning Models

Many powerful models have been introduced in recent years, since the introduction of the transformer [5], that perform well across many languages (including LRLs). The subsequent subsections will introduce the main models examined in this thesis, and speaking to their design, training, and application in the areas of classification and MT.

2.1. BERT Base

Bidirectional Encoder Representations from Transformers (BERT) [6] was a watershed moment in the evolution of NLP as it introduced the use of the transformer encoder architecture [5] to generate deep bidirectional context representations. BERT developed from bidirectional pre-training for attention models, or unidirectional use in left to right or right to left contexts, whereas this powerful model jointly conditions on both the left and right contexts at every layer. Consistently combining both left and right contexts in training allows this approach to discover richer semantic and syntactic dependencies.

The base configuration of BERT is composed of 12 transformer encoder layers with 12 self-attention heads, a hidden size of 768, and about 110M parameters. BERT was pretrained using large-scale English corpora (Wikipedia and the BookCorpus) using two self-supervised objectives:

- Masked Language Modeling (MLM): random 15% of tokens are masked and the model is trained to predict the masked words, thus biasing it to capture bidirectional context from both directions.
- Next Sentence Prediction (NSP): it trains the model to predict if one sentence follows another which enable tasks that require discourse-level understanding.

BERT quickly achieved state-of-the-art performance on a wide range of downstream NLP tasks, including text classification, Named Entity Recognition (NER), question answering, and natural language inference, evidenced in benchmarks such as GLUE and SQuAD [7], [8]. BERT's release signified a move away from task-specific architectures towards fine-tuning large pretrained models.

The multilingual variant, *bert-base-multilingual-cased*, scales BERT from 1 to 104 languages by training on their collected Wikipedia dumps, and a joint 110k WordPiece vocabulary. It is case-sensitive and was built to be a single universal encoder that supports multilingual tasks. Although it is not specifically formulated for cross-lingual transfer, later studies revealed its usefulness in zero-shot and cross-lingual transfer instance learning tasks across languages, especially in classification and sequence labeling tasks [9].

2.2. DistilBERT Base

DistilBERT [10] is a smaller and faster version of BERT, obtained via knowledge distillation [11] and will guide a smaller student network to reproduce the behavior of the larger teacher model. The main goal was to keep the majority of BERT performance with a much lower computational cost.

The DistilBERT architecture reduces the number of transformer encoder layers from 12 (in BERT base) to 6, while the hidden size (768) and the number of attention heads (12) is unchanged. The reduction in size results in a model with approximately 66 million parameters—roughly 40% fewer than BERT base—while preserving about 95% of the performance of BERT on standardized NLP benchmarks like GLUE. Hence, DistilBERT is smaller, faster (approximately 60% faster inference), uses less memory and is very useful when resources are constrained [10].

The multilingual model, *distilbert-base-multilingual-cased*, was trained using the same procedures and corpora as multilingual BERT, and it is optimized for 104 languages which share a subword vocabulary. It is smaller, but it still performs competitively on multilingual classification and cross-lingual transfer tasks. Research has found that DistilBERT offers a good trade-off between efficiency and accuracy, making it an interesting option for tasks requiring large scale or real-time performance [9], [10].

2.3. XLM-RoBERTa Base

XLM-RoBERTa [12] is a multilingual extension of RoBERTa [13] that was built to overcome some limitations of previous multilingual models like mBERT [9]. Although the architecture of XLM-RoBERTa is identical to that of BERT and RoBERTa (12 transformer encoder blocks, hidden size 768, and 12 self-attention heads), the source of its training data and the tokenization method are quite different.

XLM-RoBERTa was trained on 2.5TB of filtered CommonCrawl data in 100 languages, and is one of the largest multilingual corpora at the time. Rather than language-specific segmentation, XLM-RoBERTa uses a SentencePiece tokenizer that uses a joint shared subword vocabulary of 250k tokens. This approach also helps it model low-resource and morphologically rich languages better, by allowing different languages to share representations from overlapping subword units [12].

Over a variety of evaluations, XLM-RoBERTa achieved state-of-the-art results on multiple cross-lingual understanding benchmarks (XNLI [14], MLQA [15], TyDiQA [16]), consistently outperformed multilingual BERT, and excelled in typologically diverse and under-represented languages [12]. Overall, this performance advantage may be partly due to the fact that the XLM-RoBERTa was trained on a corpus that is larger and better balanced overall (e.g., multilingual vocabulary) and partly due to greater degrees of generalization across families of languages.

2.4. BART and mBART

Bidirectional and Auto-Regressive Transformers (BART) [17] is a sequence-to-sequence model that incorporates the best aspects of BERT style bidirectional encoders with Generative pre-trained Transformer (GPT) style autoregressive decoder. BART's architecture employs a basic encoder-decoder Transformer architecture where the encoder reads "noised" sequences and the decoder autoregressively reconstructs the original text. The objective function for BART training employs a denoising autoencoding task that applies noise functions such as token masking, token deletion, token permuting, and text infilling, thus enabling BART to learn comprehensive representations for understanding and generation tasks.

BART's flexible architecture allows it to perform exceptionally well on a number of NLP tasks including abstractive summarization, dialogue generation, and especially MT. Compared to purely encoder-based models like BERT, BART is inherently better equipped to handle sequence-to-sequences tasks because of the generative nature of its decoder.

mBART [18] builds on this architecture by extending to the multilingual regime. mBART was trained on huge monolingual corpora across 25 different languages, using one shared and subword vocabulary. It performs the same denoising pretraining objective but in the multilingual domain. A main feature of mBART is that it was purposefully created for many-to-many translation, meaning a single model can translate between multiple source and target languages without language-specific parameters.

In MT benchmarks, mBART has yielded competitive performance even in low-resource settings. Unique to mBART is that it leverages transfer learning by learning related languages and generalizing better when being fine-tuned on very low amounts of parallel data [18], [19].

2.5. MarianNMT

MarianNMT [20] is a fast and flexible open-source Neural Machine Translation (NMT) framework for training and inference. MarianNMT is implemented in C++, and has few dependencies, meaning it is lightweight and can run as needed in both research and production settings. MarianNMT has implemented Transformer-based architectures [5], which allow it to train NMT systems at scale with lower computational compared to other frameworks.

MarianNMT is also more than a framework, as it has also been a platform for multilingual translation projects. Most notably, the Helsinki-NLP group has released hundreds of pretrained translation models through the OPUS-MT project [21], which is also based on MarianNMT. The OPUS-MT models were trained using the large OPUS parallel corpora, with a variety of domains and language pairs.

2.6. The No Language Left Behind (NLLB) initiative

The No Language Left Behind (NLLB) initiative [22] by Meta AI is a landmark initiative for massively multilingual MT. The NLLB models are trained to cover over 200 languages, and cover many languages that have been severely neglected in previous research in NLP. The major advances of NLLB are due to systematic improvements in data curation, filtering and tokenization techniques, which allow a generalization across languages that are typologically diverse and of lower-resource.

The architecture is built upon large-scale Transformer architectures [5], yet NLLB is more focused on fairness and coverage across languages; simple linguistic models for smaller languages cannot constantly be outperformed by high-resource languages during training.

To enhance the accessibility and usability of the NLLB models Meta released models in several different parameter sizes. From small, efficient models to best-performing large ones, the model size allows researchers and developers with low computational power to experiment and deploy high-quality multilingual translation in models like the 600M parameter model while still providing the largest models to perform maximum translation performance in a high-resource environment. This tiered approach creates opportunities for a larger community of people to benefit from the NLLB models, including academic researchers working in low-resource settings and industry applications that rely on state-of-the-art performance.

CHAPTER 3

Related Work

This chapter reviews the foundational concepts and prior research underpinning this study. It describes relevant areas of NLP specifically the challenges for LRLs and the methodologies available to circumvent these challenges. It considers transfer learning, multilingual models, and their comparisons. The chapter also provides a summary of CVC and what studies have noted about the language itself with respect to literature that is open to public access. Finally, it presents related work focusing on datasets and NLP methods for under-resourced languages, with an emphasis on translation and classification tasks.

3.1. Natural Language Processing (NLP)

NLP is an interdisciplinary field at the intersection of computer science, linguistics, and artificial intelligence, and aims for computers to work with human language [23], [24]. After decades of development, NLP methods evolved from early rule based models to data methods that take advantage of statistical methods [25], and recently, DL architects [26]. Along this learning path, two architectures, neural sequence-to-sequence (seq2seq) models [27] and attention based architectures [28], demonstrated extensive capabilities in the areas of MT, sentiment analysis, text classification, etc. The introduction of the Transformer architecture [29], introduced self-attention which lead to parallel computation advances and acceptance of longer dependencies providing great capabilities for developing new state of the art multilingual models and increasingly Large Language Model (LLM).

For morphologically rich and underrepresented languages, Zhao *et al.* [30] and Rios, Aziz, and Sima'an [31] provide evidence that complex linguistic features such as code-switching, agglutination, and orthographic inconsistency necessitate model architectures and preprocessing pipelines that are uniquely equipped to their challenges. These improvements benefit not just the advancement of individual languages, but contribute to a broader goal of increasing linguistic equity in AI systems and digital tools, which should accurately represent global linguistic diversity instead of prioritising a small set of dominating languages.

While a lot of the developments in NLP have resulted from high-resource languages (English, French, and Chinese, etc.), there is also an increasing body of work dealing with a specific set of LRLs, languages that do not have large annotated datasets, pretrained embeddings, or linguistic resource tools available to build off. Notably, community-driven projects like Masakhane [1] [2] have shown that with enough collaborative community

[1]Masakhane is a grassroots initiative that is trying to increase NLP research on the 2000 African languages. We risk losing these languages through colonialism and the technologies underpinning these

support, translations can take place for African languages, as well as LRLs generally, without institutional funding or LRLs experience. Recent work has also experimented with different paradigms of training in these contexts, like adaptive pretraining [4] and language selection and multilingual fine-tuning [3] to facilitate optimal model training when there is no data directly available. Other proposed methods include transfer learning from high-resource languages [32], very large multilingual pretraining [33], and synthetic corpus generation [18] to alleviate data scarcity.

3.2. NLP for Low-Resource Languages

Although there has been considerable progress in multilingual NLP, LRLs are still very seriously underrepresented both in terms of research and applications [34]. These languages often lack large-scale parallel corpora, pretrained embedding, and instead, rely on the existence of suitable computational infrastructure, which leads state-of-the-art methods to still be mixing ideas developed for English and applying them to languages that can be very different, leading to poor performance, and perpetuating the digital divide. Addressing this challenge will require a mixture of technical, linguistic, and socio-cultural approaches.

Transfer learning from high-resource languages has, so far, topped the leader board in terms of effective methods, because models could exploit shared linguistic information using multilingual architectures such as XLM-R [35] and mBERT [6]. Multilingual NMT systems [33] and adaptive fine-tuning [4] have also contributed to the potential for generalization to unseen languages. Community-driven data creation activities [2] show how local orthographies, dialectal variation and cultural contexts can be catered for through inclusive dataset curation.

For creole like languages, hurdles are magnified with high linguistic diversity, unclear orthographic systems, and an overall lower written tradition, noted by Rios, Aziz, and Sima'an [31] and Zhao *et al.* [30]. These issues will ultimately impact the possibilities for standard tokenization, alignment, and language modeling, and will in effect imply extended use of pre-processing or annotation guidelines with the proposed application. Some research on available resources constrained languages with populations in the millions has turned to methods like creating a synthetic parallel corpus [18], some have used a pivot-based translation with an intermediary language [32], and it is becoming more frequent to use cross-lingual embeddings [36] as a work around for the resource gap here. CVC is an interesting but complicated situation: it serves as an example of an LRLs, highly constrained technically, but offers a chance to use transfer learning and a multilingual aspect with Portuguese as well as with other creoles.

languages are not intentionally neglected but rather remain in the shadows causing a gap where culture is not recognized. More information on `https://www.masakhane.io/`

3.3. Cape Verdean Creole (CVC)

CVC is classified within the Portuguese-based creole language family and has substantial lexical, phonological, and morphosyntactic variation not only from its lexifier, Portuguese, but also between its regional varieties [37], [38]. The two major branches, Sotavento (Southern islands) and Barlavento (Northern Islands), show systematic variation with respect to vowel reduction, consonant cluster reduction, pronoun paradigms, and tense-aspect-mood marking [39]. The stratification of CVC, along with inconsistent orthography, creates special problems for model building NLP system, since system trained on one variant and performs well, may often NLP perform poorly an another variant.

These language-specific features affect tokenization, language modeling, and other developments in NLP, like MT and classifying. Inconsistent orthographic form can increase vocabulary fragmentation and create a bunch of sparse representations and higher out-of-vocabulary rates. Creole morphosyntactic features are rare to represent in pretrained multilingual embeddings (like reduplication for emphasis or aspect markers), thus making transfer learning less effective without domain adaptation [38], [39]. Additionally, Portuguese influence on formal registers, and a CVC's spoken-dominant moments can create code-switching forms of language use, which affect the corpus creation and require particular preprocessing.

Digital resources for CVC are still limited, inconsistent, and vary in quality. The majority of existing resources are the result of academic fieldwork, community projects, or lexicographic projects (e.g. [1]). Although there exists a small collection of lexical databases, grammatical descriptions, and collections of folklore, the chance of finding them in machine-readable formats, or licensed for open distribution is still highly unlikely, thus hindering the use of them in NLP workflows.

Initiatives in other creole languages provide models for resource development. Haitian Creole has been supported by joint campaigns for data annotation and inclusion into multilingual benchmarks [40], [41], and resources for Mauritian Creole [42], [43] show the advantages of community-based standardization of orthography, between many other related works [41], [44]. Engaging in comparable activities for CVC, especially participatory data collection and standardization of orthography, would likely lead to even faster development of viable NLP systems for the language.

The CrioleSet dataset [1], which contains translation pairs between CVC, English, Portuguese, and French, was created to enable MT experiments while explicitly addressing the larger issue of the linguistic challenges and lack of a standardized resources for CVC. To the best of our knowledge, there are no standardized POS-tagged corpora, no dependency treebanks, and no speech datasets available. Researchers are left with several limited options such as engaging in transfer learning from high-resource languages, manually collecting and generating data, and creating synthetic resources. CrioleSet fills this large gap by providing a publicly available, well-formed parallel corpus, which lays groundwork for MT for CVC via [1].

3.4. Datasets for Creole and Underrepresented Languages

The growth of linguistic resources for low-resource and underrepresented languages has come a long way, although the coverage is inconsistent. Large multilingual resources such as JW300 [45], OPUS [46], and the datasets developed by the Masakhane project [2], [47] cover a majority of African languages and some Creoles, but there is little or no coverage for CVC or very insignificant coverage. The original challenge of lacking standard orthography and multiple dialects is unique to dataset construction [34].

In particular for creole languages, targeted corpora have been of use to enable downstream NLP tasks. The datasets of Haitian Creole [40], [41] and resources for Mauritian Creole [42], [43] highlight the potential of targeted data curation and working with communities to develop sustainable resources. Likewise, both those projects, as well as CCAligned [48] and the FLORES benchmark [49] indicate that, even when limited bilingual data is available, a new multilingual context allows it to be used together with a training paradigm that can yield better performance. However, unlike with Haitian or Mauritian Creole, there are currently no large, open-access corpora for CVC, which reduces the applicability of any existing pretrained multilingual models without applying substantial adaptations.

3.5. Machine Translation in Low-Resource Scenarios

MT in low-resource scenarios faces challenges that go beyond data scarcity, often involving orthographic variation, code-switching, and morphosyntactic irregularities that limit the direct application of high-resource models. Recent developments exploit transfer learning [50], multilingual fine-tuning [12], and back-translation [51] techniques to leverage cross-lingual transfer. These methods allow models trained in resource-rich languages to bootstrap performance in resource-poor, yet linguistically related, languages.

For Creole languages such as Haitian Creole [40], [41], NMT methods, especially Transformer-based architectures, have significantly outperformed traditional phrase-based statistical systems. However, the performance gap between Creoles and dominant world languages remains considerable, highlighting the persistent challenges in low-resource MT.

Synthetic data generation [18] and multilingual joint training further improve model robustness. Adaptive pretraining using domain-specific or dialect-specific corpora can also align model representations more closely to the target linguistic features. In the case of CVC, dialectal fragmentation introduces additional complexity: a model trained on one creole variant may perform poorly on others, and without explicit variant labeling, standard MT systems struggle to generalize across the language.

Point-to-point learning treats the entire NLP system for MT, as a single mapping function f. Adopting the CVC to English example, if f is a CVC-to-English translation model then the transformation of input CVC sentence (F_{cv}) to their English equivalent occurs through (F_{en}) such that $F_{\text{en}} = f(F_{\text{cv}})$. For English-to-CVC it is $F_{\text{cv}} = f'(F_{\text{en}})$ instead. Figure 3.1 depicts this mapping a MT function f, where each mapping follows the formalism detailed in Russell and Norvig [52].

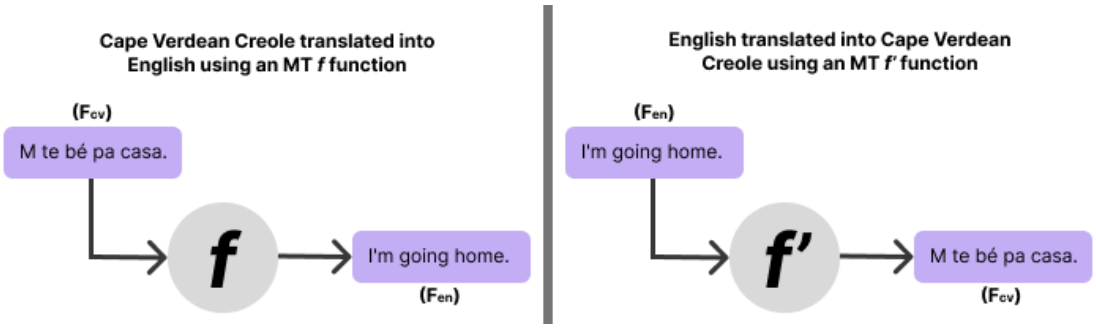

FIGURE 3.1. MT f function representation between CVC and English

3.6. Dialect and Variant Classification in Low-Resource Languages

Automatic classification of dialects and language variants is a fundamental task for LRLs, particularly when linguistic diversity and limited resources hinder standard NLP approaches. The primary goal of classification is to accurately assign input text to a predefined variant or dialect label, which is critical for downstream tasks such as machine translation, sentiment analysis, or information retrieval.

Existing approaches for variant classification leverage different neural architectures. Character-level convolutional networks capture subword and orthographic patterns that are especially relevant for morphologically rich or orthographically inconsistent languages. Hierarchical recurrent architectures model longer contextual dependencies, which can disambiguate variants in texts with subtle syntactic differences. Transformer-based classifiers offer state-of-the-art performance by attending to long-range dependencies and capturing nuanced contextual relationships, making them particularly effective for distinguishing closely related creole variants [53]–[55].

In low-resource scenarios, one major challenge is the scarcity of labeled examples. Multitask learning [56]–[58] addresses this by jointly training models on variant classification alongside other related tasks, such as translation. By sharing representations across tasks, the model can generalize better even with limited annotated data. Additionally, data augmentation techniques such as back-translation or synthetic text generation can expand the training set and improve classifier robustness.

From a formal perspective [52], classification can be modeled as a mapping function f (illustrated in Figure 3.1) from input sentences F_{cv} to discrete label space L_{var}, represented by the function $L_{\text{var}} = f(F_{\text{cv}})$, where L_{var} represents the predicted dialect or variant label, while f encapsulates the learned parameters of the classification model. Effective classification not only enables accurate identification of linguistic variants but also supports downstream tasks such as targeted MT, corpus curation, and language-specific resource creation.

3.7. Pre-trained Models for Machine Translation

Research on multilingual transformer-based models has expanded rapidly in the context of MT for LRLs. For CVC, this is especially relevant due to both the scarcity of annotated

resources and the high degree of dialectal variability. Several studies demonstrate how multilingual models can improve translation performance in such low-resource contexts.

Robinson *et al.* [59] introduced *Kreyòl-MT*, a large-scale machine translation initiative covering 41 Creole languages and 172 translation directions. By fine-tuning mBART-50 on 14.5 million Creole sentences, the authors demonstrated the ability of multilingual models to generalize across multiple Creole varieties, substantially improving translation quality despite limited language-specific data.

Parameter-efficient fine-tuning strategies have also been explored. Fekete *et al.* [60] investigated adapter-based tuning for Creole MT, showing that adapters allow large multilingual models to effectively transfer knowledge between high-resource and low-resource languages while significantly reducing training costs.

Other work highlights the advantages of multilingual encoder-decoder architectures. Zheng *et al.* [61] pretrained mBART on Indigenous American languages, achieving consistent improvements in BLEU and chrF scores relative to baselines without multilingual pretraining. Similarly, Tonja *et al.* [62] compared bilingual and multilingual systems for Indigenous languages, concluding that multilingual fine-tuning consistently yields better performance.

Earlier research by Lakew, Cettolo, and Federico [63] showed that multilingual NMT can mitigate overfitting and improve generalization in low-resource settings by sharing parameters across languages. Building on this, more recent work such as Sarkar *et al.* [64] demonstrated that NLLB can be successfully applied to spoken language translation tasks for Indic languages, where fine-tuned NLLB models produced strong results even in noisy, cascaded pipelines. While not directly targeting Creole languages, this reinforces the robustness of NLLB in low-resource and challenging conditions, making it a promising baseline for CVC.

Finally, research on in-context learning, such as the study by Akallouch and Fardousse [65] for Tarifit, suggests that LLMs can achieve usable translation quality with few-shot demonstrations, even without large parallel corpora. This line of work highlights the potential of prompt-based methods as a complementary approach for CVC translation.

Taken together, these studies confirm that multilingual transformer-based architectures, including mBART, MarianNMT, and NLLB are highly effective tools for advancing MT in low-resource and Creole contexts.

3.8. Pre-trained Models for Classification

While much of the focus in LRLs has been on translation, classification tasks such as sentiment analysis, NER, and dialect identification are fundamental components of natural language processing, especially in low-resource contexts. For CVC, where dialectal fragmentation between Barlavento and Sotavento adds complexity, variant classification provides an important complementary task to translation, supporting the design of more robust end-to-end NLP pipelines.

Recent advances in transformer-based architectures have demonstrated that multilingual pre-trained models can generalize effectively to LRLs, including in dialect identification scenarios. Fouadi *et al.* [66] applied BERT-based models to classify Arabic dialects from social media data and reported strong performance in distinguishing closely related varieties, illustrating that fine-tuning pre-trained encoders can capture subtle dialectal signals. Similarly, Yadav *et al.* [67] evaluated mBERT, DistilBERT, and RemBERT on Hindi and Nepali NER, showing that multilingual BERT-variants transfer knowledge well between related languages, even with limited annotated data. These findings are especially relevant to CVC, where dialect boundaries share similar challenges of fine-grained discrimination.

Beyond BERT, lightweight alternatives such as DistilBERT have also been shown to perform competitively in low-resource classification. For instance, Agbesi *et al.* [68] compared BERT, DistilBERT, RoBERTa, and DeBERTa for text classification in Ewe, a low-resource African language, reporting that DistilBERT offered strong trade-offs between efficiency and accuracy. These results suggest that reduced-size models can be leveraged in resource constrained scenarios, which is relevant for deployment-oriented CVC applications.

Salleh *et al.* [69] explored a hybrid model for LRLs text classification and compared XLM-RoBERTa with other tools, showing that it performs strongly in low-resource scenarios. Its demonstrated ability to handle multiple languages and domains underpins its selection for the variant classification experiments in this thesis.

While BART has primarily been employed in sequence-to-sequence tasks, it has also been adapted for classification when combined with appropriate fine-tuning strategies. In multilingual contexts, encoder-decoder architectures such as BART and mBART have shown potential for few-shot or zero-shot classification, further expanding the range of transformer models applicable to low-resource settings [12], [18].

This prior work shows that BERT, DistilBERT, BART, and XLM-RoBERTa can be effectively adapted for classification in low-resource contexts, supporting their use in CVC variant classification where subtle dialectal differences must be captured despite limited data.

CHAPTER 4

The Morabeza Corpus

This chapter documents how Morabeza Corpus, a new and improved version of CrioleSet [1], was collected, assembled, and prepared for NLP tasks.

4.1. Sources and Selection Criteria

To account for actual usage and increase language diversity, text was combined from (i) consensually collected private messaging (Facebook, WhatsApp), (ii) Cape Verdean literature and other authored texts (e.g., public prose/poetry, essays), (iii) web content harvested from directed web scraping (e.g. news/blog posts), and (iv) any other contributions from the community. Comparatively to large multilingual collections such as JW300, OPUS, or Masakhane datasets [45]–[47], the amount of CVC content is very minimal. The intended preliminary objective was to curate CVC content across respective registers and variants. The collection framework was in alignment with community first principles as demonstrated in participatory work for African languages [44], as the collection process emphasized: consent, transparency, and representativeness.

Texts kept based on the set of principles: (1) *showcase variant coverage* of Barlavento and Sotavento varieties (cf. the linguistic overview in Chapter 3; (2) *domain balance*, informal chats versus edited prose versus web texts; (3) *clean extractability* sentence segmentation feasible without excessive noise; (4) *legal clarity permission* or public availability with a compatible licence; (5) *parallelizability availability* or feasibility of aligning with English for MT; and (6) *ethics*, no sensitive personal information (Personally Identifiable Information (PII)) or content subject to special procedures (Section 4.5).

4.2. Data Collection, Preprocessing, and Normalization

Considering the orthographic instability and intra-language variation in CVC [38], [39], the model for collecting, processing, and normalizing the data was designed as a modular pipeline. The goal was to adopt a *lightweight*, *variant-aware* normalization approach rather than a prescriptive unification, thereby ensuring that linguistic diversity remained intact while improving usability for downstream NLP applications.

As shown in Figure 4.1, the pipeline is organized into three main phases: *Data Collection Sources*, *Data Processing / Extraction / Storing*, and *NLP Applications*.

In the first phase, *Data Collection Sources*, CVC texts were gathered from multiple channels (Facebook and WhatsApp messages, CVC literature, web scraping, and other contributions). For text and messages extracted from Facebook and WhatsApp, the data was consensually collected with explicit permission from the owners. Web-scraped data was sourced from forums, social media, online translations of song lyrics, and other web

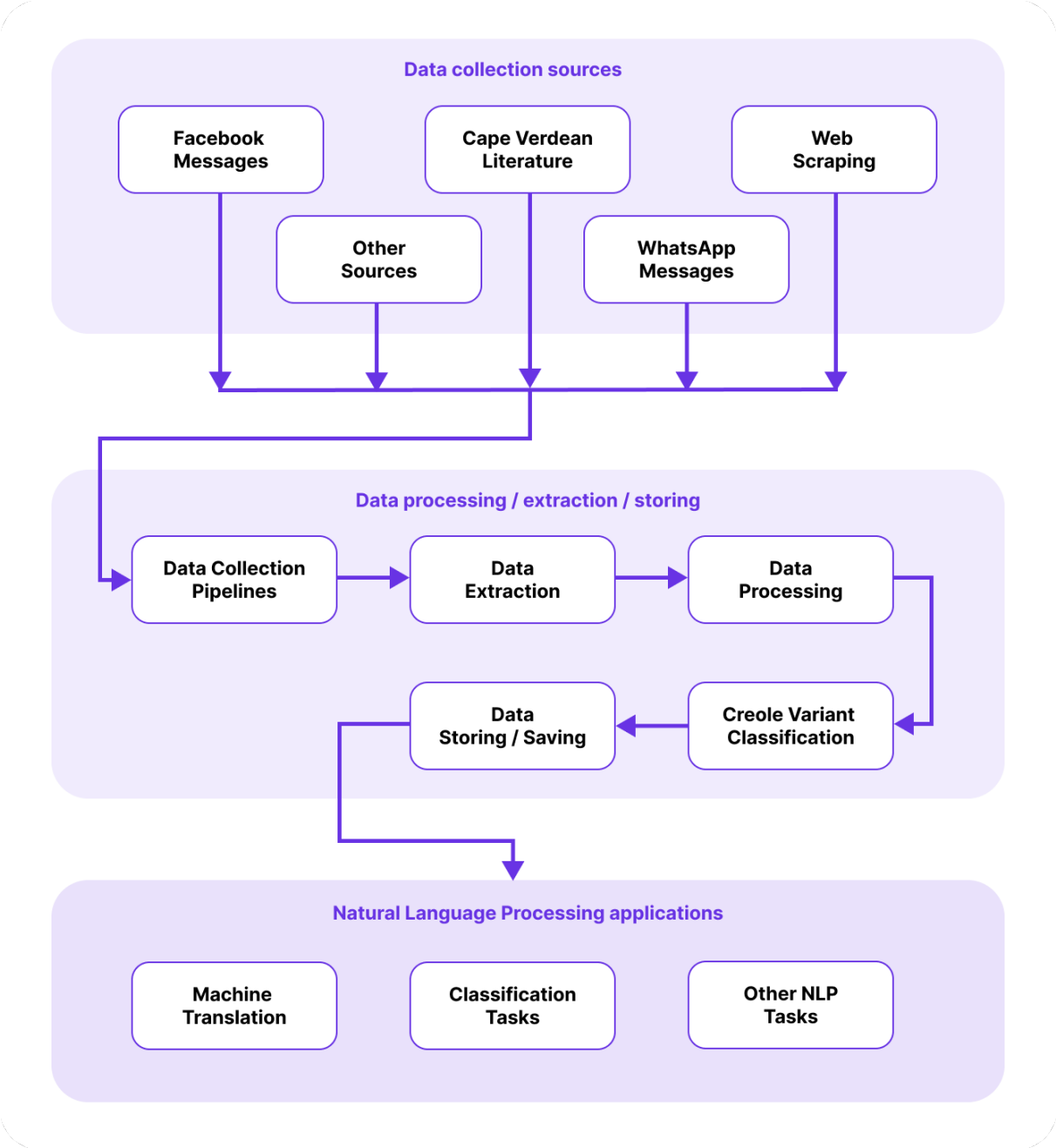

FIGURE 4.1. Data collection workflow for Morabeza Corpus: sources (top), processing / extraction / storing (middle), and NLP applications (bottom)

pages containing CVC text. In addition, further material was provided by volunteers who wished to contribute to the project.

In the second phase, *Data Processing / Extraction / Storing*, pipelines were created to extract, clean, and normalize the CVC text. After processing, the texts were classified by regional variant (Barlavento or Sotavento) and stored in the dataset. During this phase, the following steps were performed:

- **Language identification**: Depending on the source, this was performed either automatically or manually. Automatic identification was applied when the source

was known to be exclusively CVC, while manual filtering was necessary for mixed-language sources (e.g., Facebook, WhatsApp, and web scraping). Scripts were developed to detect CVC text and exclude other languages.

- **Cleaning and normalization**: After identifying CVC sentences, texts underwent normalization steps including whitespace standardization, removal of special Unicode characters and emojis (common in chat data), normalization of quotation marks and symbols, and lowercasing (excluding named entities).
- **Anonymization of writers**: To protect sensitive information in Facebook and WhatsApp data, user identifiers were anonymized. Personal names were replaced with generic placeholders, and any personal or sensitive information was removed.
- **Variant classification**: Before being stored, texts were classified by island group (Barlavento or Sotavento). Initially, this was performed manually, but later automated using classification models described in Chapter 5
- **Data storage**: Finally, the processed and classified texts were stored in the dataset, forming a structured and reusable resource.

In the third and final phase, *NLP Applications*, the dataset became ready for downstream tasks such as MT and classification (see Chapter 5).

This workflow ensured that text from all raw sources (Figure 4.1) was consistently transformed into a normalized, variant-aware dataset, enabling robust NLP experimentation and application.

4.3. Dataset Structure, Statistics and Linguistic Characteristics

Morabeza Corpus is organized as a parallel corpus, between 4 languages, CVC, English, Portuguese and French. This corpus are prepared to be used for MT, but also Morabeza Corpus, provides and extra set that can be used for classification tasks, between the creole sentence source (Barlavento or Sotavento). Building on the earlier public release [1], this new version of the dataset extends the resource with standardized splits and richer annotations.

4.3.1. Data Structure

For this study there are two stages of the dataset, one that is the development stage and another one that is the final stage, were it can be used to perform NLP tasks. Either for MT or classifications there is structure that the data is organized. The datasets build for all the stages and for the NLP tasks are all stored on Comma-Separated Values (CSV) files, each of them with their on structure.

During the development stage, and for statistical purposes, the data collected had the following headers:

- **timestamp** - When the sentence / word was collected
- **kea** - CVC version of the sentence
- **en** - English version of the sentence
- **pt** - Portuguese version of the sentence

- **fr** - French version of the sentence
- **island_group** - The island group of the sentence (Barlavento or Sotavento)
- **writer** - The message writer or the person that created the sentence (in order to maintain the privacy of the people how contributed with their WhatsApp and Facebook messages the writer name was associated with some alias, for example "writer_1" till "writer_n")
- **source_type** - Source from which the sentence was extracted. Possible values: null, literature, web-scraping, message-whatsapp, and message-facebook.
- **writer_age_group** - The age group of the person who created the sentence. Possible values: null, between-10-and-20, between-20-and-30, between-30-and-40, between-40-and-50, between-50-and-60, and above-60.
- **writer_education_level** - The education level of the person who created the sentence. Possible values: null, no-education, elementary, secondary, bachelor, master, and PhD.

Table 4.1 presents the final version of the dataset, following statistical analysis and the removal of contributor information. The contributors information was used solely for statistical purposes during dataset analysis. The final version is prepared to be used in NLP tasks such as classification and MT.

TABLE 4.1. Morabeza Corpus dataset

island_group	kea	en	pt	fr	source_type
barlavento	m oiá ukê ke bô fazê lá.	i see what you did there.	eu vi o que você fez lá.	je vois ce que tu as fait là.	message-facebook
barlavento	manham no te falá ntom	tomorrow we talked then	amanhã nós falámos então	demain nous avons parlé alors	message-facebook
barlavento	sorre demora p respondê, ê ke m tite falá má nhe irmã	sorry to take a long time to answer, i was talking to my sister	desculpa demorar para responder, é que eu estava a falar com a minha irmã	désolé de prendre beaucoup de temps pour répondre, je parlais à ma sœur	message-whatsapp
sotavento	mundu sta cheio di malvadéza	the world is full of evil	o mundo está cheio de maldade	le monde est plein de mal	literature
sotavento	cusas di coraçom ta mexe cu mim	heart stuff messing with me	coisas do coração mexem comigo	des trucs de cœur qui me dérangent	web-scraping
sotavento	brasâ-m pa n podi ser senpri es mudjer	embrace me so that i can always be this woman	abraça-me para que eu possa ser sempre essa mulher	embrasse-moi pour que je puisse toujours être cette femme	web-scraping

Based on the information reported in Table 4.1, another subset containing all the tokens from CVC was created. Table 4.2 shows example of this tokens, that are also been translated to English, Portuguese and French.

TABLE 4.2. Morabeza Corpus tokens dataset

island_group	kea	en	pt	fr
barlavento	felicidad	happiness	felicidade	bonheur
barlavento	trozód	late	atrazado	en retard
barlavento	cansód	tired	cansado	fatiguée
sotavento	subrivivi	survive	sobreviver	survivre
sotavento	nisisidadi	necessity	necessidade	nécessité
sotavento	combersu	conversation	conversa	conversation

4.3.2. Data analysis

In order to provide a richer perspective on the dataset, we examined its composition in a number of dimensions, including the distribution by island group, source type, writer characteristics, and level of education. This demonstrates the representativeness, and variety, of Morabeza Corpus, while also indicating the distributions that may cause distortion in downstream tasks.

The entirety of the dataset contains 581,793 unique sentence records. When the multiple sentence copies are also considered, the dataset then contains 1,364,498 total sentences. The dataset contains 148,880 unique tokens, which is a considerable amount of lexical coverage of CVC in multiple variants and contexts.

Most of the data, as shown in Figure 4.2, is from Barlavento group, 74.9% of the records (427,071 sentences), and 25.1% from Sotavento (143,055 sentences). While many factors could contribute to this disparity, a plausible explanation lies in the fact that the authors and message writers are primarily from Santo Antão and São Vicente, two of the Barlavento islands. Moreover, many of the writers with whom the author collaborated, through messages or other sentence sources, are also primarily from this island group.

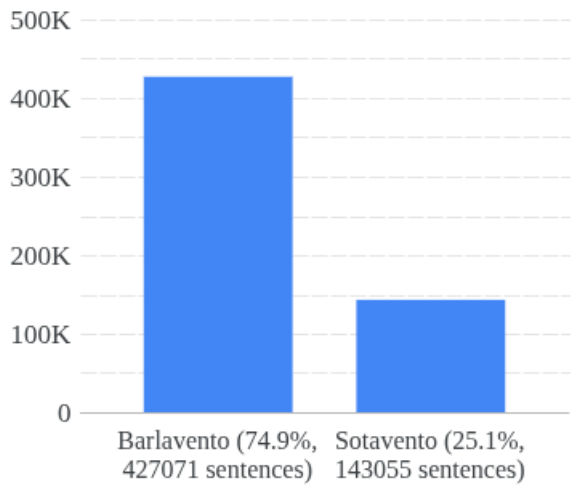

FIGURE 4.2. Morabeza Corpus sentences per island group

The dataset brings together collections from different sources, ensuring coverage of both formal and informal registers. The biggest share of data is taken from web scraping (56.7% 323,081 records), Facebook messages (36.6% 208,409 records), WhatsApp messages (6.2% 35,263 records) and Cape Verdean literature (0.6% 3,373 records) (Figure 4.3). The predominance of digital communications for data sources demonstrates the realization of language use by contemporary speakers, while literature collection provide a somewhat more standardized and historical antecedents to language use, and this contributes for a more diverse and rich dataset.

When analyzing only those relevant records with information on writer island, Sotavento writers cover 57.2% (1,294 writers) and Barlavento writers cover 42.8% (969 writers) (Figure 4.4). They are inverses of the overall dataset distribution, as while Barlavento

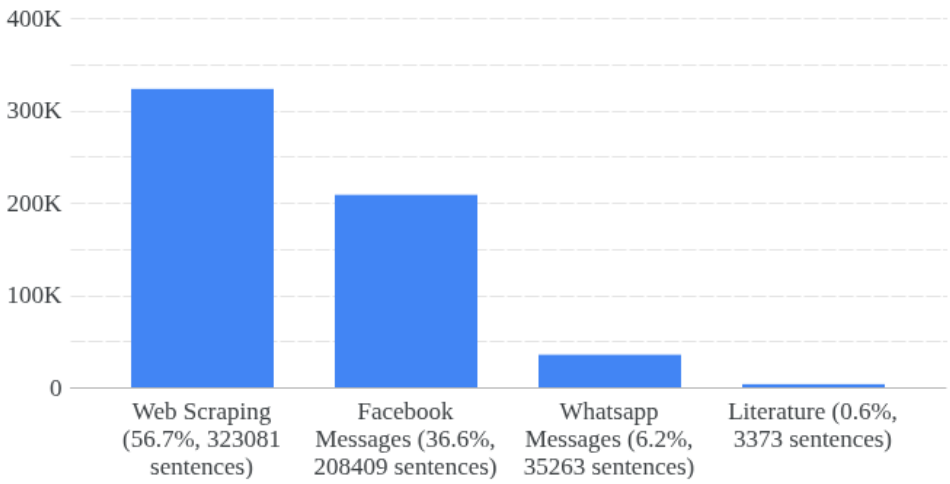

FIGURE 4.3. Morabeza Corpus sentence sources

can have more sentences (Figure 4.2), there were a relatively great number of identified writers from Sotavento, meaning that authorship and sentences do not necessarily have to correlate. After careful analysis of the dataset, it appears that this is predominantly due to the fact that the majority of the message writers from Sotavento, have fewer contributions of messages than from writers in Barlavento.

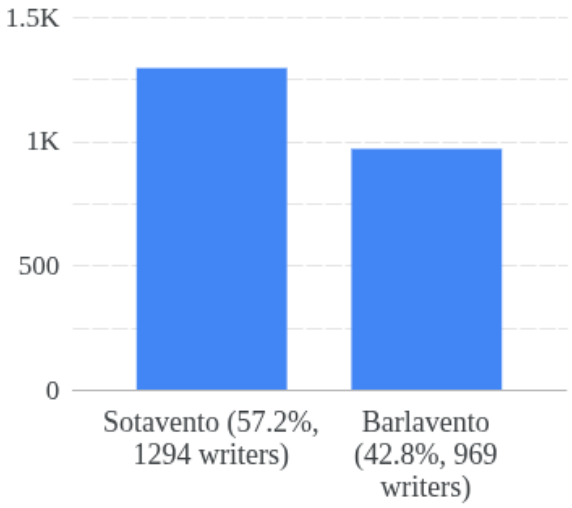

FIGURE 4.4. Morabeza Corpus writers island group origin

The dataset does also contain demographic differences by age (Figure 4.5). The majority of contributors are between the ages of 20 and 30 (36.3%, 821 writers), and subsequently between the ages of 10 and 20 (22.2%, 502 writers) as well as between the ages of 30 and 40 (19.3%, 437 writers). The number of contributions declined in the older groups, 9.2% were between the ages of 40 and 50 (209 writers), 3.0% between the ages of 50 and 60 (68 writers), and 4.6% beyond the age of 60 (105 writers). A small share of contributors (5.3%, 121 writers) could not be designated to a specific age group. This distribution reflects a predominance of younger speakers / writers in digital communication, which

is the principal source of the data, but it does also reflect some representation from the generations of older speakers / writers.

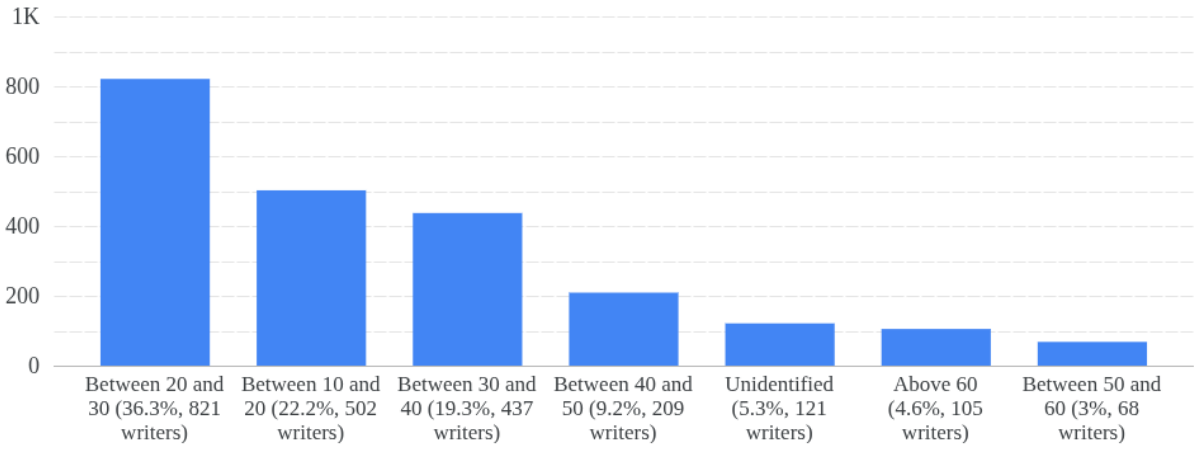

FIGURE 4.5. Morabeza Corpus writers age group

In addition to educational background, the dataset's characterization includes information about the contributors' educational background (Figure 4.6). The largest degree group in the dataset is secondary education with 888 writers (39.2%), followed by bachelor degree holders, 707 writers (31.2%). Contributors with elementary-level education account for 15.5% (350 writers), and smaller proportions are reflected in master's education (5.1%, 115 writers) and PhD education (2.1%, 48 writers). Additionally, there are 33 writers (1.5%) which reported no formal education. This spread indicates that Morabeza Corpus reflect a wide range of language use from informal and conversational registers to educated writing styles.

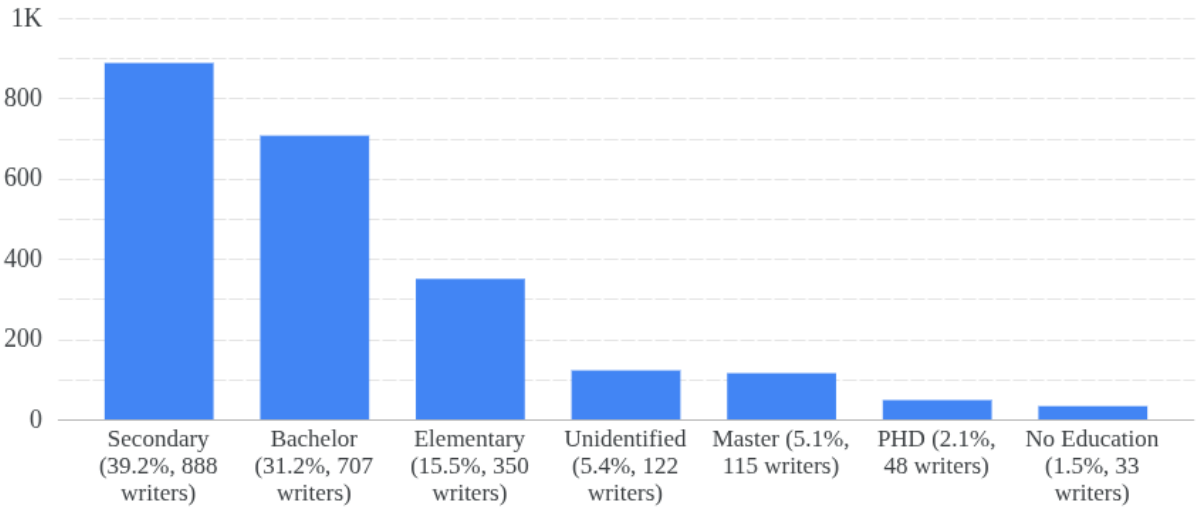

FIGURE 4.6. Morabeza Corpus writers education level

In summary, these statistics demonstrate both the strength and diversity of Morabeza Corpus, with stronger representation for both island groups (Barlavento and Sotavento),

ages, and education levels and with range across sources), which is important for downstream NLP applications. This variation was build to support tasks like MT, making classification of variants, and creating a wider linguistic analysis.

4.3.3. Token Embedding Visualization

Token frequency statistics and distributional token embeddings were used to examine the lexical profile of the Morabeza Corpus. Taken collectively, the findings emphasize the preponderance of function words, recurring content words, and interesting regional variation between the Barlavento and Sotavento varieties of CVC.

Table 4.3 and Table 4.4 show most frequent stop-words for Barlavento and Sotavento. The stop-word lists have very clear differences in what is ranked highest: Barlavento to stop-word tokens are "bo" (48,529 times) and "um" (43,978), while Sotavento most prevalent stop-word tokens, are "ta" (24,174) and "na" (18,090). These differences indicated that the two regional varieties featured different function words in high frequency, which may illustrate syntactic preferences, different uses of particles, or tokenization effects.

TABLE 4.3. Barlavento 40 most frequent stop words and their frequencies

Word	**Freq**	**Word**	**Freq**	**Word**	**Freq**	**Word**	**Freq**
bo	48529	ten	1048	li	4847	ums	42
um	43978	ki	975	já	4515	más	39
ma	25110	podê	781	nha	4453	algun	42
mas	16665	un	764	nada	2847	ora	73
na	15862	mal	713	gora	2154	ali	53
la	15741	ku	659	uma	1805	sin	74
ta	14888	enton	600	pa	6919	ken	139
bem	13483	di	543	ou	5950	sê	143
ja	12119	cedo	328	mi	9406	nôs	146
i	11234	sabi	208	tud	8441	tê	148

TABLE 4.4. Sotavento 40 most frequent stop words and their frequencies

Word	**Freq**	**Word**	**Freq**	**Word**	**Freq**	**Word**	**Freq**
ta	24174	pa	11218	bo	7316	mi	5190
na	18090	sta	10392	di	6912	se	4624
ka	12717	um	8005	i	6623	la	4047
bu	7685	ma	7701	ki	6105	ten	3827
um	8005	bem	2057	mas	5732	el	3705
li	3645	un	2342	nha	5514	ke	597
kel	3338	ou	2430	ku	5332	tud	574
es	2800	tem	2800	ja	1138	undi	525
kes	1949	já	1715	nada	1633	sin	708
bem	2057	uma	2025	pode	1199	mal	633

Table 4.3 and Table 4.4 illustrate both overlap and divergence in formal, non-stop and token lexical types. Many common content words (e.g., "pensa", "mesmo") exist

in both varieties, showing a common core vocabulary. Other frequent lexical types are variety-specific (e.g., Barlavento presents frequent tokens such as “xcola”, “meste”, “diaza” while Sotavento lists tokens such as “staba”, “xinti”, “djobi”). These discrepancies are likely sourced from topical and stylistic biases in the data (domains, genres, or collection method) as well as from real regional lexical preferences.

TABLE 4.5. Barlavento most frequent tokens and their frequencies

Word	**Freq**	**Word**	**Freq**	**Word**	**Freq**	**Word**	**Freq**	**Word**	**Freq**
essim	4922	prala	1337	maltas	2131	intende	1061	mensagem	975
einda	3401	durmi	1303	traboi	2102	gente	1050	xcola	972
pensa	2718	sobre	1293	tiver	1974	meste	1034	festa	990
manera	2533	primer	1270	escola	1908	diaza	1033	depos	1017
mesmo	2521	noite	1257	lembra	1888	dvera	1031	consegui	1017
gosta	2269	espia	1246	falta	1834	manham	1491	ideia	1456
manda	2144	conta	1221	semana	1824	algum	1425	tenta	1364
panha	1176	termina	1122	tembê	1061	volta	1553	manhã	1619
pergunta	1138	mandam	1097	feria	1090	entra	1620	manha	1667

TABLE 4.6. Sotavento most frequent tokens and their frequencies

Word	**Freq**	**Word**	**Freq**	**Word**	**Freq**	**Word**	**Freq**	**Word**	**Freq**
staba	1127	algum	429	minis	590	exame	365	mundo	343
studa	875	lembra	428	semana	580	manha	358	linda	343
mesmo	853	forma	424	mesti	565	munti	356	festa	340
pensa	814	djobi	423	xinti	557	amigo	355	normal	337
txiga	687	dipos	407	noite	545	família	354	sobre	335
feliz	684	lugar	402	midjor	515	coisa	353	pergunta	333
manda	669	filha	397	spera	513	alguém	345	força	328
gosta	663	conta	390	tenta	487	praia	487	volta	474
tempo	657	amiga	381	amigos	486	cabeça	484	panha	464

Figure 4.7 and Figure 4.8 show the three-dimensional projections of token embeddings computed from the corpus[1] , allowing inspection of their semantic and distributional structure. The x, y, and z axes correspond to the first three principal components (PC1, PC2, and PC3) obtained through Principal Component Analysis, which capture the major directions of variance in the embedding space. Although these axes are not directly interpretable in linguistic terms, they effectively expose patterns of similarity and divergence among tokens.

From these projections, it becomes clear that Barlavento and Sotavento exhibit partially separable clustering neighborhoods in multiple regions of the space. These clusters are conventionally and comparatively composed of tokens that are more frequent in only one of the two variants of CVC, pointing to distributional differences associated with regional usage. Ultimately, the visual patterns and groupings demonstrate that the token

[1]An interactive visualization can be accessed online at `https://robertocarlosmedina.github.io/morabeza-corpus-tokens-3d-visualizer/`

embeddings encode meaningful semantic and syntactic distinctions within the Morabeza Corpus and reflect regional variation between Barlavento and Sotavento in the learned embedding space.

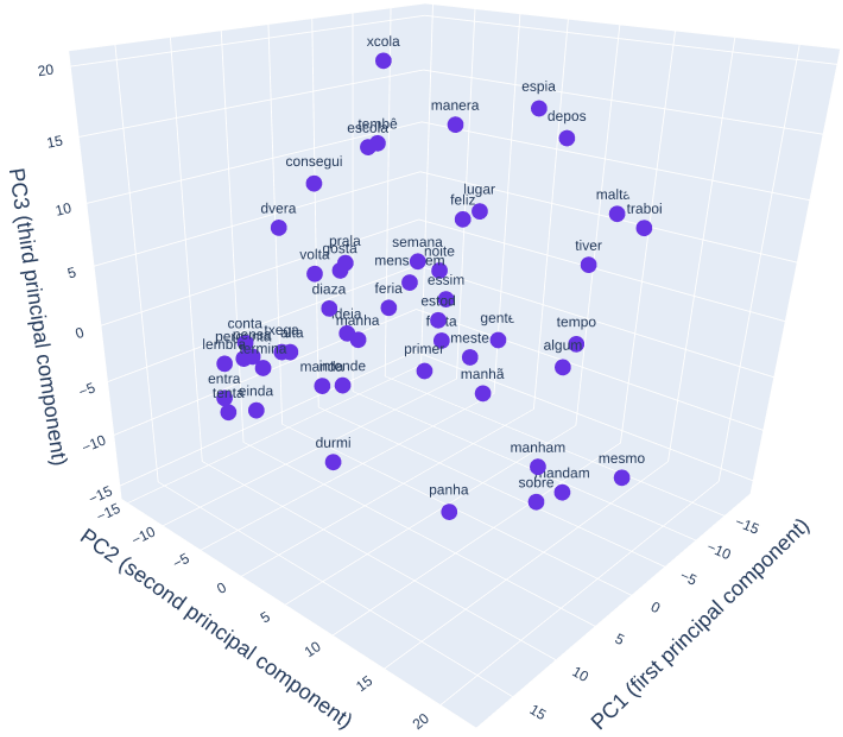

FIGURE 4.7. Morabeza Corpus 50 Barlavento most frequent token 3D embeddings, without stop words

Manual inspection of the visualizations and tables showed some striking patterns: (i) medium-frequency tokens clear semantic cohesion which could be used as seeds for the lexicon; (ii) overlap tokens (e.g., "pensa", "gosta") tend to occupy the middle ground among regional clusters reflecting their cross-regional value which explains their relatively informativeness of a dialect classification; and (iii) low-frequency tokens that are region-specific are positioned in more isolated regions embedding space reflecting they may be a good marker for dialects although with more variation.

The combined frequency and embedding analyses demonstrate that the Morabeza Corpus captures both a shared core lexicon and meaningful regional variation, highlighting clear patterns in token usage and distribution across Barlavento and Sotavento. These analyses reveal how semantic and syntactic relationships are encoded in the corpus and show that regional differences are systematically reflected in the data, offering valuable insights into the structure and variation of CVC.

4.3.4. Linguistic Characteristics of the Dataset

Beyond the statistical summary and token embedding analyzed, knowledge of the dataset can also be considered through a linguistic lens. In this case, four aspects were analyzed:

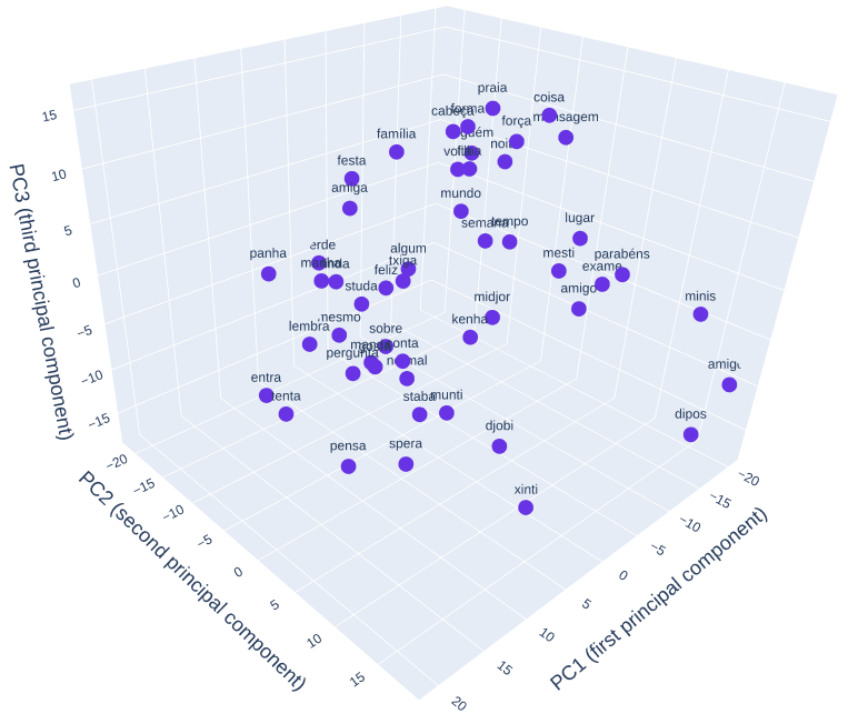

FIGURE 4.8. Morabeza Corpus 50 Sotavento most frequent token 3D embeddings, without stop words

the distribution of sentence length, lexical diversity, orthographic variation, and occurrences of code-switching with Portuguese. Aspects analyzed together provide more depth of understanding of the properties and challenges of Morabeza Corpus.

A sentence length analysis shows that most of the sentences are relatively short, over 65% have 15 or fewer tokens (see Figure 4.9). The informal and conversational origins of many of the sources drew from (social media, instant messaging, as shown on Figure 4.3) is reflected in the short sentences. Longer sentences (greater than 20) are drawn from text types that were literary and web materials. The combination of these two sources shows the range of informal and formal registers produced in the dataset. Differences between dialect groups are also observed in the average length of sentences since Sotavento sentences were generally longer than the length of Barlavento sentences.

Lexical richness was assessed using the Type–Token Ratio (TTR). In aggregate, the dataset has a TTR of 0.18, which is in line with other conversational corpora [70], [71]. If calculated by dialect, we find the TTR of Sotavento was 0.21 and the TTR of Barlavento was 0.17, indicating that, while the Barlavento portion is substantially bigger the Sotavento dialect shows greater variety of vocabulary per unit volume. The apparent greater diversity for Sotavento is likely due, in part, to the more heterogeneous nature of sources included in the Sotavento subset.

CVC has no standardized orthography, which is evident in this dataset. Many lexemes emerge with multiple forms. For instance, the adverb “assim” occurred in both the Portuguese-based version (“assim”, 2,830 times) and the more phonetic based Creole version (“essim”, 4,922 times) (this can be visualized in Figure 4.7 and Figure 4.8).

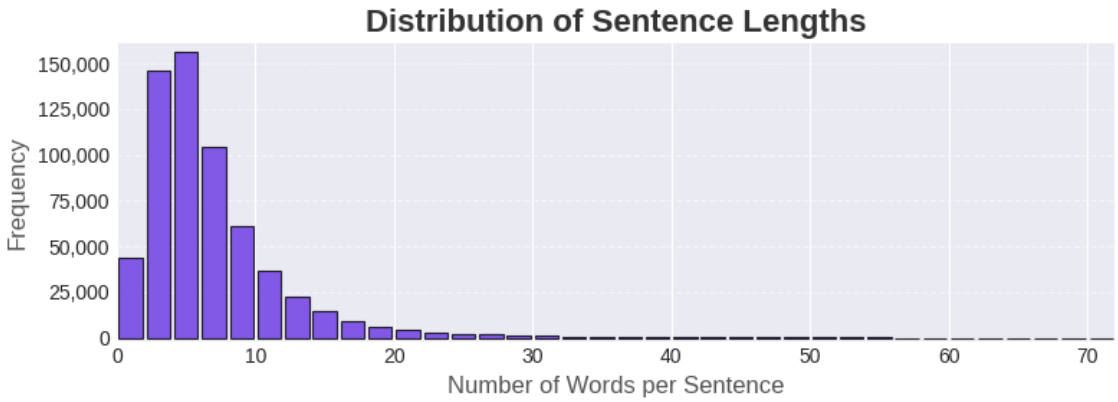

FIGURE 4.9. Morabeza Corpus sentence length distribution

Similarly, the verb "chega" (corresponding to the verb "arrive" in English) as it occurred both as "chega", "txega" and also as "txiga". The different spellings lead to more token sparsity, complicating embedding learning, and reinforcing the varied normalization practices needed for downstream NLP tasks. There are examples of common orthographic variations in Table 4.7.

TABLE 4.7. Examples of orthographic variation across Morabeza Corpus, showing alternative spellings and their frequency

Word	**Variant**	**Appearance count**	**English translation**
kenhem	Barlavento / Sotavento	923	who
kenha	Sotavento	810	
gosi	Sotavento	1302	now
gora	Barlavento	1203	
agora	Barlavento / Sotavento	701	
chega	Barlavento	234	arrive
txega	Barlavento	803	
txiga	Sotavento	921	

Code-switching is a prominent characteristic of CVC speech and writing. The dataset documents approximately 12.6% of sentences including at least one Portuguese token. This is worth noting because it demonstrates the code-switching that occurs in bilingual environments in which the language is spoken. These tokens include phrases, function words (e.g., "que", "porque", etc.), and full phrases; younger speakers, in particular agency, display this in their written messages. Code-switching details more considerations for tasks such as MT and classification because models need to disambiguate code-switching token (e..g. similar structures, different contexts) in both languages. As seen in Table 4.8, examples of code-switched sentences are presented, highlighting the language use and code-switching concept.

TABLE 4.8. Illustrative examples of code-switching between CVC and Portuguese within Morabeza Corpus

Type	**Sentence Example**
Mixed Creole/Portuguese	Mi sta content porque m consegui kel trabalho.
Intra-sentential switch	Nos ta papia kriolu, mas depois ele falou em português.
Borrowing	Bo pode manda mensagem, tudo bem?

Overall, these analyses highlight the linguistic richness and challenges that occur in Morabeza Corpus. The data set represents, not only, a large-scale resource of under-resourced languages, but highlights features such as orthographic variation and code-switching which capture critical aspects of how languages are used in the day to day live in Cape Verde.

4.4. Challenges Building the Dataset

Morabeza Corpus exemplified various obstacles, linguistically and technologically, that detail the challenges of working with under-resourced and non-standardized languages. One of the main problems was the lack of a standardized orthography in CVC because a many words appear in multiple forms (see sub Section 4.3.4) and there was considered two main variants: Barlavento and Sotavento. Given this, instead of opting for a single unified spelling, the data were lightly normalized. and in a variation-aware way to standardize the texts as much as possible, yet retaining the original forms as features.

Another set of challenges arose largely from the heterogenity of data sources. Social media and messaging data (e.g., WhatsApp, and Facebook) had features such as informal spellings and abbreviations, emojis, and inconsistent spelling. Web scraping sometimes presented instances of CVC mixed with Portuguese or English, so it was necessary to have reliable language identification filters. In contrast, literary texts to some extent conformed to conventions which contrasted starkly with the sources. This type of variety also meant that we needed flexible preprocessing pipelines that could respond to different registers of language use.

Ethical considerations were also important. There was needed to render sensitive information from personal messages anonymous, to uphold contributors confidentiality. Author names and personal identifiers were substituted with generic equivalents, while taking care to make sure not to impact the naturalities of the text. Simultaneously, another recurring issue arose in that there is often substantial redundancy in datasets: messaging platforms typically produced many near-duplicate or repeated sentences. To avoid contaminating frequency counts and training data, near-duplicate detection was used to manage excessive repetition while maintaining authentic conversational features.

Another difficulty was variant classification. Determining if a sentence was part of the Barlavento or Sotavento variant was done manually during the beginning of the project, but subsequently was automated (see Chapter 5). Regardless, many unclear findings were still present, especially short messages with little amount of contextual information. At

the same time, the constant code-switching between CVC and Portuguese created another hurdle, as many speakers naturally switch and interchange the two languages in what they might call a single sentence. Rather than eliminating these cases from our data set, we used care in filtering and tagging, so that code-switching could be studied in terms of natural language use.

Finally, a central issue was finding a balance between linguistic differentiation and computational usability: too much normalization could account for variation that brought valuable linguistic diversity, while not enough would result in data that was too noisy for downstream NLP models. The strategy employed was to find a happy medium where we applied only minimal cleaning and normalization, and managed all of the orthographic and lexical diversity in descriptive documentation. This was done so that a dataset would be as true to the real language as possible, but still useful for MT purposes.

Overall, the construction of Morabeza Corpus needed to take into account; the many aspects of orthographic insecurity, dialectal diversity, heterogeneous sources, anonymization, duplication, and code-switching; to name only a few issues. These issues are common in LRLs, and unlike Haitian or Mauritian Creole, where targeted corpora and community projects exist [42], [43], CVC lacks large, open corpora.

4.5. Ethical Considerations and Licensing

In terms of consent and privacy, all private messages (Facebook/WhatsApp) were collected based on informed consent. Any PII (names, phone numbers, emails, and locations) and sensitive material was removed or masked until released. Only de-identified, non-sensitive parts of messages were kept.

The Community participation and benefit, Followed participatory frameworks [44], contributors were informed about goals, scope, and possible uses of the resource. There were participatory feedback loops with speakers and educators to jointly make variant tags choices and light-touch normalization choices.

The Morabeza Corpus, therefore, is an ethically-sourced, reproducible parallel corpus with relevant variant annotations for (i) MT (across CVC, English, Portuguese, and French), and (ii) variant/genre classification two NLP tasks that can be used, in the context of LRLs for CVC.

CHAPTER 5

Methodologies

In this chapter, the methodological process taken for the experiments is provided which includes preparation of the dataset, experimental setups, methods of classifying, and MT model configurations. It also provides description of the evaluation protocols and metrics to measure performance so that the design decisions are made in the spirit of transparency and reproducibility. The focus of this chapter is to describe how the DL models used in this work, introduced in Section 2, were configured and integrated into the overall experimental pipeline.

5.1. Dataset Selection and Preparation

Due to the extensive nature of Morabeza Corpus, as detailed in Chapter 4, and the exceeded computational capability to create and train large-scale deep learning models with the dataset (see Section 2), a limited context of the data was used for the classification and MT tasks. The intention was to balance the representativeness of the data and its computational feasibility, as well as provide the linguistic diversity required in the subsets to support an experiment with a meaningful human-centric goal.

In order to conduct the classification experiments, a balanced subset of 12,000 sentences was extracted from the full dataset, 6,000 sentences were from Sotavento variant and 6,000 sentences from the Barlavento variant. The balanced subset was extracted in order to avoid any imbalance in the learning process of the model, represent both major variants of CVC equally. This subset of data was split into training, validation, and test sets (70%, 15%, and 15% of the data respectively) using standard procedures in supervised learning [72]. The subset was partitioned in order to have sufficient examples for training, while still maintaining proper evaluation and test data, to assess the model performance.

In terms of MT tasks a subset of 20,000 sentence pairs was taken from Morabeza Corpus. The subset is parallel data from CVC, English, Portuguese, and French demonstrating that CVC is situated in a multilingual setting. The data was divided, with 90% for training purposes and 10% for testing purposes following current practices in MT research [73], where the larger training sets were preferred due to the data-hungry nature of neural MT, and a smaller but representative set of data for testing. A separate validation set was not created for MT, as early stopping and tuning of hyperparameters were completed through cross-validation within the training set.

All sentences for both NLP tasks underwent preprocessing and tokenization prior to being fed into the models. The preprocessing stage consisted of normalization, deduplication, and limited orthographic harmonization, as outlined in Chapter 4. Tokenization was achieved via the pretrained tokenizers provided for each of the models. This included SentencePiece for XLM-RoBERTa [12], BERT WordPiece [6], and byte-pair encoding for MarianNMT [51]. Using the original pretrained tokenizers ensured that we were aligned with the models' training regime, and it also helped us capture a great deal of morphological and subword structure in CVC, which is essential for low-resource, high-orthographic variation contexts.

Consequently, this method of preparing the dataset was developed to optimize both experimental integrity and feasibility, providing balanced and linguistically diverse sources for both, the classification and MT experiments while still balancing best-practices in multilingual NLP.

5.2. Training Environment Setup

All experiments were done on Google Colab Pro, running Python 3 on an NVIDIA T4 GPU and high-memory runtime, in order to efficiently train the DL models while minimizing the need for capable hardware. The main libraries used included *scikit-learn* for evaluation, *transformers* for fine-tuning the models, and *nltk* for preprocessing tasks like tokenization and text normalization.

5.3. Classification Models

To classify between the different variants of CVC (Barlavento and Sotavento, presented on the Chapter 4), experiments were done using four multilingual transformer-based models, BART, BERT base multilingual, DistilBERT, and XLM-RoBERTa. The models were selected given their demonstrated applicability to multilingual classification tasks, and their ability to transfer knowledge to LRLs tasks, as described previously.

Each model was fine-tuned for binary sequence classification. Table 5.1 summarizes the main architectural properties of the models used.

TABLE 5.1. Summary of classification models and key architectural configurations

Model	**Layers**	**Hidden Size**	**Attention Heads**	**Vocab Size**
BART (base)	6 encoder / 6 decoder	768	12	50,265
BERT (multilingual)	12	768	12	119,547
DistilBERT (multilingual)	6	768	12	119,547
XLM-RoBERTa (base)	12	768	12	250,002

Fine-tuning the models was done using a cross-entropy loss function, and using an early stopping to prevent overfitting. Dropout rates varied from 0.1 to 0.2 according to the model's default parameters in order to ensure sufficient generalization ability.

Each pretrained model is associated with a corresponding tokenizer and the tokenizers were used to preprocess the dataset (see Section 2). For BERT and DistilBERT, the multilingual, WordPiece tokenizer was used; for XLM-RoBERTa, a SentencePiece tokenizer; and for BART, the Byte-Pair Encoding (BPE) tokenizer was used for preprocessing. With this all done, this ensures that the subword segmentation of the dataset is in alignment with the model vocabularies.

Models were fine-tuned using the AdamW optimizer with a learning rate of 2×10^{-5}, with batch sizes set as either 16 or 32 depending on GPU memory limits. Training epochs were set between 3 and 5, and the models had early stopping based on validation loss. Each model took up to half a day of training, including the changes to its parameters for the experimental tests. So, in total times related to the training, testing, and evaluation for the whole collection of classification models were around two days.

TABLE 5.2. Training hyperparameters for classification models

Parameter	**Value Range**	**Notes**
Optimizer	AdamW	Standard for transformers
Learning rate	2×10^{-5}	Tuned from preliminary experiments
Batch size	16–32	Depending on GPU memory
Epochs	3–5	Early stopping applied
Dropout	0.1–0.2	As per model defaults
Loss function	Cross-entropy	For binary classification

This training setup ensured a balance between computational efficiency and model robustness, enabling reliable evaluation of multilingual models for CVC variant classification.

5.4. Machine Translation Models

In the MT tasks, four pretrained encoder–decoder systems were fine-tuned to provide translation between CVC and English in both directions: mBART-50, Helsinki-NLP's MarianNMT (ROMANCE-English), Helsinki-NLP's MarianNMT (multilingual-English), and NLLB-200 Distilled. These models were chosen based on their demonstrated effectiveness with low-resource and multilingual translation tasks (see Chapter 3).

In accordance with the computational restrictions and resources that were available to run the experiments, the configurations of the models were changed from how they were originally implemented (see Section 2). These modifications were done to use the resources available and still try to get the best results possible. All the MT model configurations used, mBART, the Helsinki-NLP MarianNMT models (ROMANCE–English and Multilingual–English), and the NLLB Distilled, are presented in Table 5.4.

All the models were trained on the training subset, as previously described, to allow for bidirectional translation to CVC and to English. This framework offers an organized way to study how the models performed on a LRLs, and whether differing architectural issues and hyperparameters had any effect on how small amounts of data governed the translation quality, even in a rectilinear mode.

TABLE 5.3. Machine Translation models configurations

Parameter	mBART-50	Helsinki-NLP MarianNMT	NLLB-200 Distilled
Encoder / Decoder Layers	12 / 12	6 / 6	12 / 12
Attention Heads	16	8	16
Hidden Size (d_{model})	1024	512	1024
FFN Size	4096	2048	4096
Activation	ReLU	Swish	ReLU
Dropout	0.1	0.1	0.1
Layer Norm	Yes	No	Yes
Max Position Embeddings	1024	512	1024
Vocab Size	250,054	64,172	256,206
Shared Embeddings	–	Yes	–
Tokenizer	MBart50Tokenizer	MarianTokenizer	NllbTokenizer

Training epochs were set between 2 and 5, and the models had early stopping based on the loss. Each model took up to two days of training, including the changes to its parameters for the experimental tests. So, in total times related to the training, testing, and evaluation for the whole collection of the MT models were around two to three weeks.

5.5. Evaluation Metrics

Evaluation was done as is standard in both text classification and MT. In the classification experiments, evaluation metrics were accuracy, macro-averaged F1, and confusion matrices to evaluate robustness across CVC variants to demonstrate balanced performance even in the presence of class imbalance [53], [56].

In assessing MT, different automatic metrics were computed in order to capture multiple facets of translation quality.

- **BLEU** [74]: a precision-based metric which focuses on n-gram overlap between system/golden translations and has been widely accepted as the default metric for MT evaluation.
- **METEOR** [75]: has been created to best correlate with human judgements with the use of stemming, synonymy (via WordNet), and recall to alleviate some of the disadvantages of BLEU in accounting for lexical variation.
- **ROUGE-L** [76]: was firstly designed for text summarization but otherwise has been broadly used in MT, it counts and measures the longest common sequence between the candidate and reference translations in order to reflect fluency and adequacy.
- **BERTScore** [77]: uses contextual embeddings from pretrained transformers as a metric for semantic similarity, evaluating translations beyond surface-level overlapping. This is useful in low-resource settings where literal overlap typically is minimal.
- **Translation Edit Rate (TER)** [78]: counts the number of edits (insertions, deletions, substitutions, shifts) needed to transform the system output into the

reference translation. The similarity with post-editing effort underneath is intuitive.

- **chrF** [79]: is a character n-gram F-score metric that includes precision and recall at the character-level, which is especially useful for morphologically rich and low resource languages where overlap at the word level is less trustworthy.

These metrics together provide a comprehensive perspective [80], BLEU and ROUGE provide surface overlap, METEOR mediates between recall and synonymy, TER reflects edit distance, and BERTScore adds in semantic adequacy.

CHAPTER 6

Experiments and Results

This chapter presents the results obtained from the dataset described in Chapter 4, focusing on two core NLP tasks: text classification (distinguishing between Barlavento and Sotavento variants) and MT (between CVC and English in both directions). Results are reported according to the evaluation metrics described in Chapter 5, with particular attention to robustness, adequacy, and fidelity of the models.

6.1. Classification Results

As described before, the classification experiments were performed on a subset of 12k records from the Morabeza Corpus dataset and were assessed within four transformer architectures, BART (base), BERT (multilingual), XLM-RoBERTa (base), and DistilBERT (multilingual). Figure 6.1 shows the training and validation loss curves, as well as validation accuracy, are all displayed for all models. All models converge smoothly and exhibit only minor overfitting, signifying that they generalize well to previously unseen data. DistilBERT shows some signs of convergence slightly faster than all the other models tested, likely due to its distilled architecture, and accordingly fewer parameters.

An overview of the accuracy, precision, recall, and F1-score for each model is provided in Table 6.1. In general, all models demonstrate high classification performance, achieving accuracies exceeding 97.8%, reflecting both the quality of the Morabeza Corpus dataset and the appropriateness of transformer-based architectures for classification of CVC language variants (Sotavento and Barlavento). Specifically, DistilBERT achieved the highest performance overall, with accuracy of 98.44%, precision, recall, and F1-score of 0.984, and slightly better than larger models, with lower computation costs.

The confusion matrices for every model are presented in Figures 6.2. Misclassifications are very low among all models, with many appearing between related CVC variants. DistilBERT appears to have the least amount of off-diagonal errors, as expected given the good performance metrics. BERT and XLM-RoBERTa are both competently discriminatory, but do have a couple of misclassifications in the minority classes, indicating that they may be sensitive to class imbalance or sample size in specific variants.

6.1.1. Model Comparison and Performance

When evaluating the models, it is clear that DistilBERT obtains the best accuracy (98.44%) while being the most lightweight in terms of computation, allowing it to emerge as advantageous for deployment use-cases with limited resources. XLM-RoBERTa gained nearly indistinguishable accuracy from DistilBERT (98.16%), as it concurrently shows

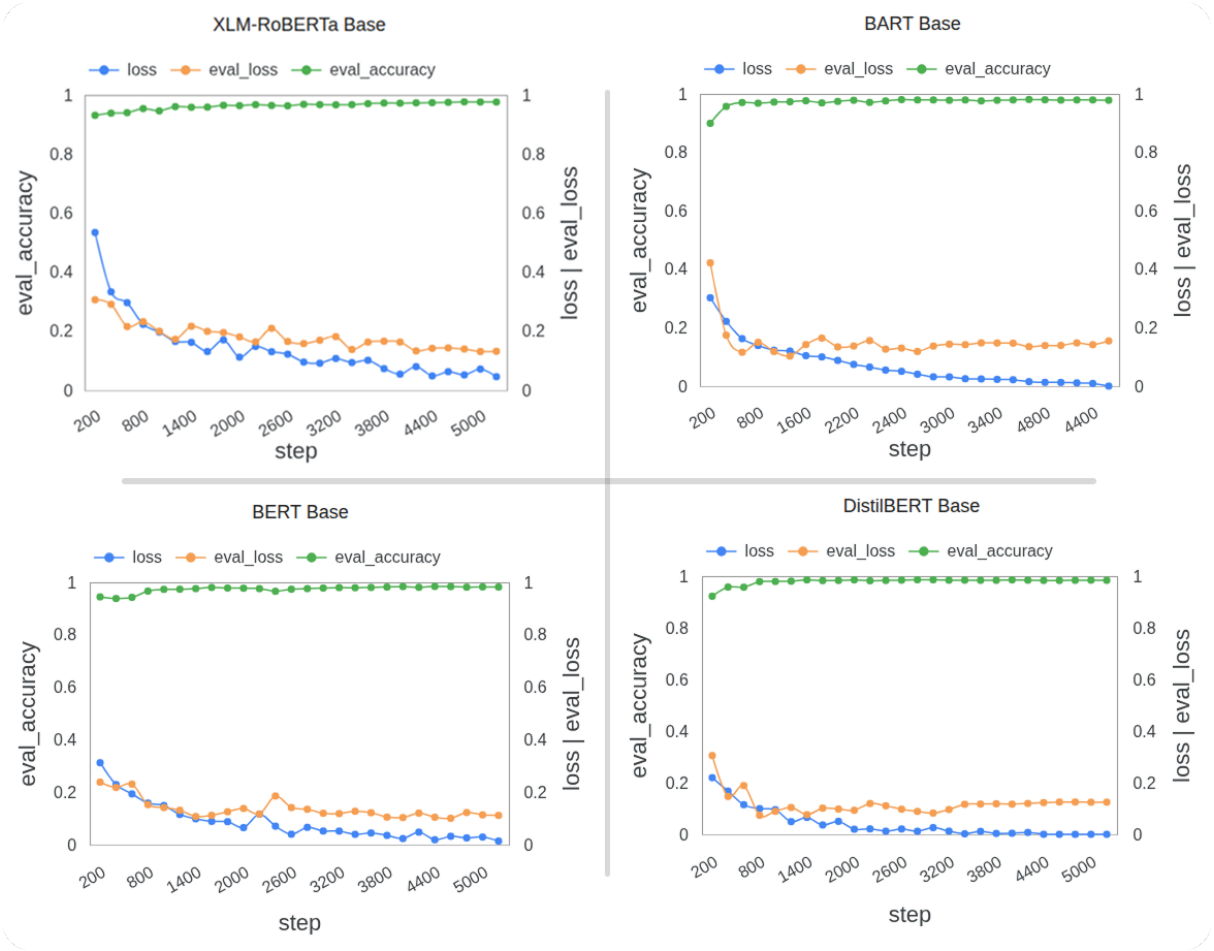

FIGURE 6.1. Classification models training loss, validation loss and validation accuracy

impressive robustness across all multilingual variants, reflecting strong cross-lingual pre-training. BERT was slightly lower than XLM-RoBERTa (98.05%). It has relatively good precision and recall across classes, indicating sufficient consistency across the classes. BART also performed reasonable well (97.88%) but slightly behind the other models; however, it could be an effect of the generative pre-training objectives of BART not being fully aligned, or beneficial, in a task class that was discriminative in nature (see Table 6.1).

TABLE 6.1. Classification performance on the 12k Morabeza Corpus subset.

Model	**Acc. (%)**	**Precision**	**Recall**	**F1-Score**
BART (base)	97.88	0.978	0.978	0.978
BERT (multilingual)	98.05	0.980	0.980	0.980
XLM-RoBERTa (base)	98.16	0.981	0.981	0.981
DistilBERT (multilingual)	**98.44**	**0.984**	**0.984**	**0.984**

By manually testing the models on them to assess their performance, it is possible to see that DistilBERT has the best results, although, overall, all models did well on the manual tests, which can be seen in Table 6.2 when testing them with some sentences, notably all the models made a fair share of correct predictions.

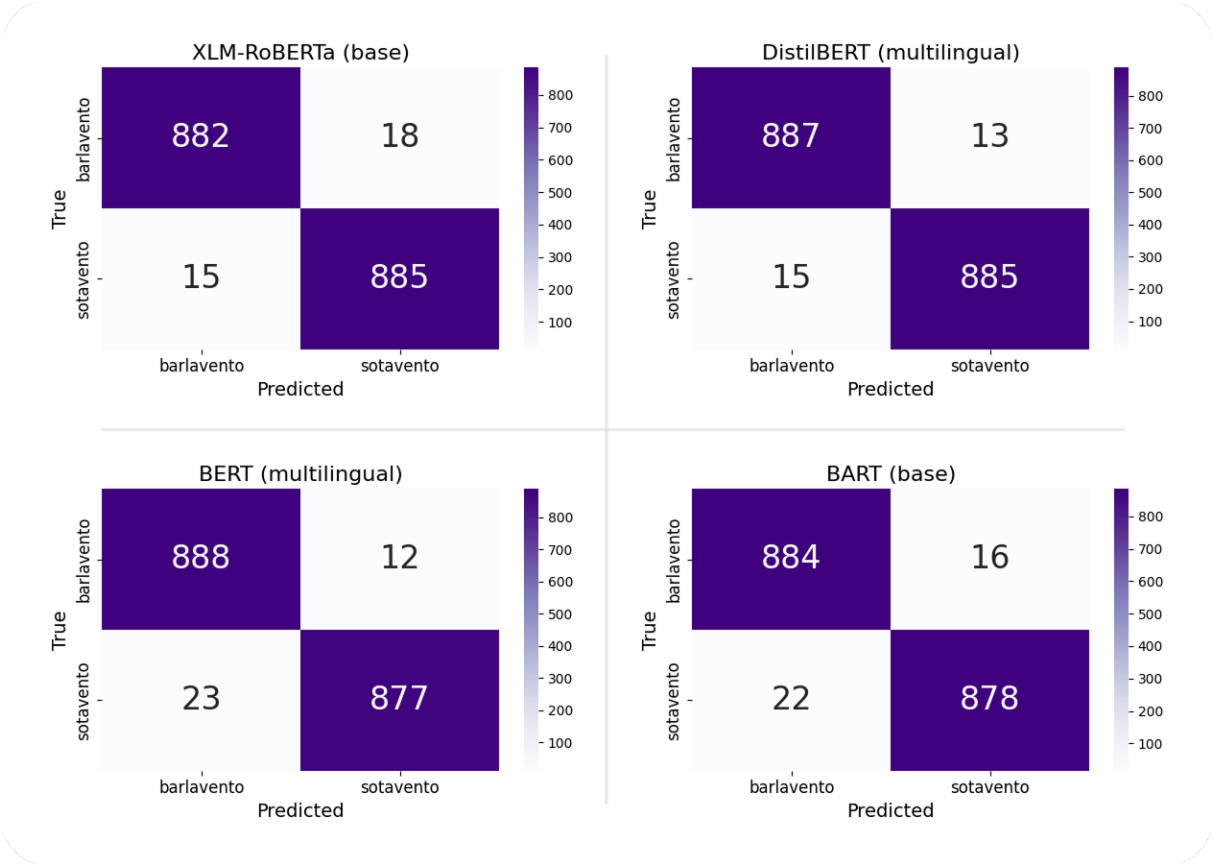

FIGURE 6.2. Classification models confusion metrics

TABLE 6.2. Example classification predictions on Morabeza Corpus by four models. Correct predictions are highlighted in green, incorrect in red.

Sentence	**True Value**	**BART**	**BERT**	**XLM-RoBERTa**	**DistilBERT**
Ami m pensa ma tudo staba dretu	sotavento	True	True	True	True
Embora m te otxá ke cosa ene bem essim, mas pronto	barlavento	True	True	True	True
Keli go, m ka speraba di bu	sotavento	True	True	True	True
Bai deli, que krê sabe de bô mas ne nhe vida	barlavento	True	True	True	True
Sempri m gosta de planeja cusas di nha manera	sotavento	True	True	True	True
El ta vivê ma se pai y se mãi ne Sonsent, desdi pikinoti.	sotavento	True	False	True	True
Dentu mi tem 10 ilha	sotavento	True	False	True	True
Foi lá k m machucá nha tornozel	barlavento	False	True	False	True
Um hom ta sentóde frente de um ecrã de cumptador ta segura um teclado e ta oia pa câmera.	barlavento	False	False	False	False

6.2. Machine Translation Results

Figures 6.3 and 6.4 show the loss curves of the training for the machine translation models from CVC to English and from English to CVC respectively. The convergence patterns are marginally suggestive of stable training across all models. The mBART-50 exhibited a faster and more consistent loss reduction when compared to the other models, which

potentially indicates that the multi-lingual encoder-decoder pre-training of the mBART-50 provided it with an adaptable application to the CVC translation tasks.

Table 6.3 outlines the evaluation results across numerous automatic metrics, which include BLEU, chrF, METEOR, ROUGE-L, BERTScore, and TER. In the comparison, a clear ranking of performance is present among the models tested. Additionally, Tables 6.4 and 6.5 provide instances of some manual translation evaluations applied to the models and their outputs.

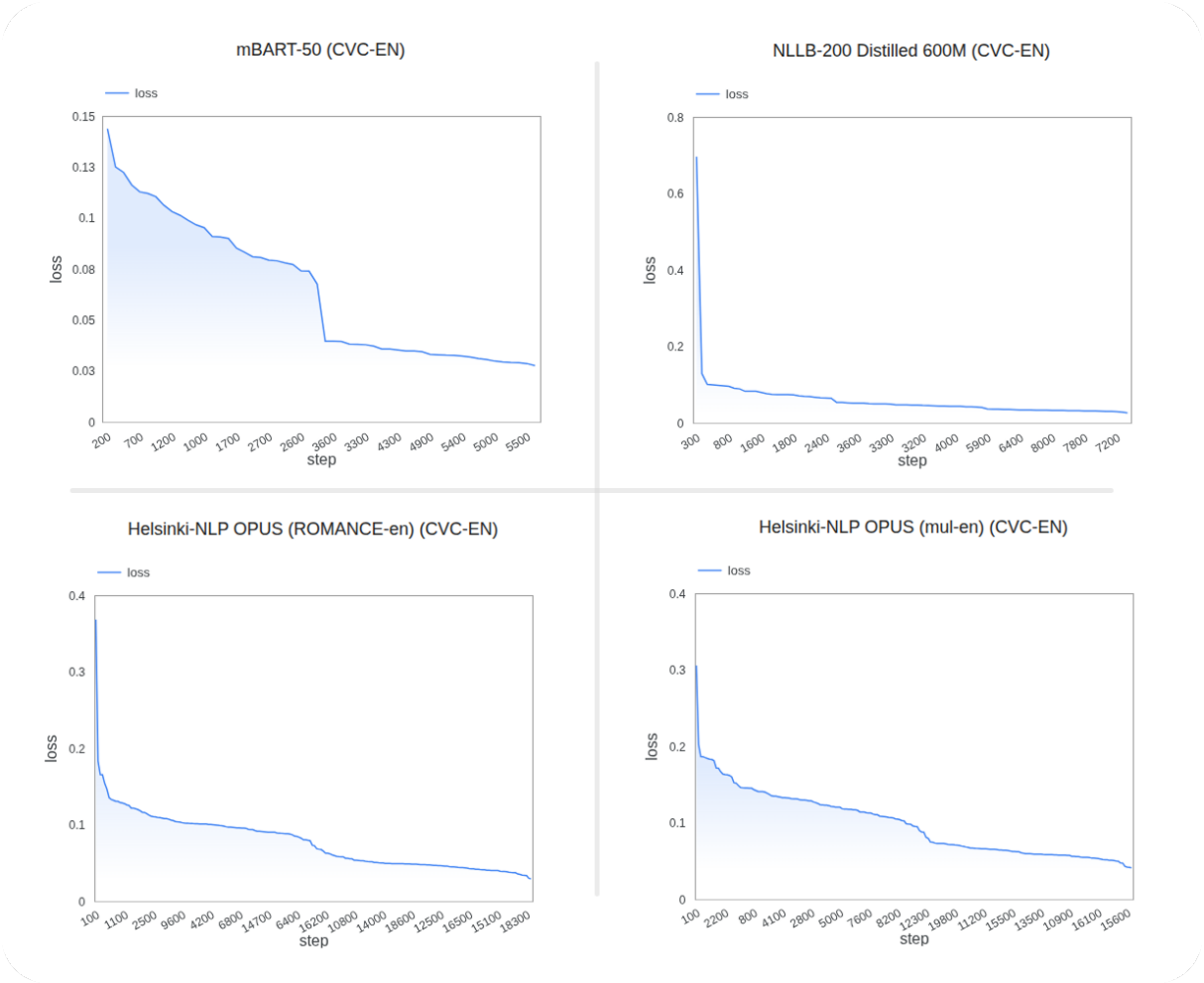

FIGURE 6.3. MT models loss during training to translate from CVC (kea) to English (en)

TABLE 6.3. MT results for model trained to translate from CVC (kea) to English (en)

Model	BLEU(%)	chrF(%)	METEOR	ROUGE-L	BERTScore F1	TER(%)
NLLB-200-Distilled-600M$_{(kea-en)}$	37.11	73.32	0.75	0.68	0.914	75.86
NLLB-200-Distilled-600M$_{(en-kea)}$	28.83	61.87	0.59	0.61	0.887	63.74
Helsinki-NLP-OPUS-ROMANCE-en$_{(kea-en)}$	73.38	81.27	0.82	0.85	0.973	21.73
Helsinki-NLP-OPUS-ROMANCE-en$_{(en-kea)}$	73.36	81.13	0.82	0.85	0.971	21.64
Helsinki-NLP-OPUS-mul-en$_{(kea-en)}$	68.26	76.58	0.78	0.82	0.966	26.05
Helsinki-NLP-OPUS-mul-en$_{(en-kea)}$	69.37	77.58	0.79	0.82	0.967	25.73
mBART-50$_{(kea-en)}$	**81.26**	**86.31**	**0.87**	**0.90**	**0.981**	**15.80**
mBART-50$_{(en-kea)}$	**83.49**	**88.98**	**0.86**	**0.89**	**0.977**	**12.21**

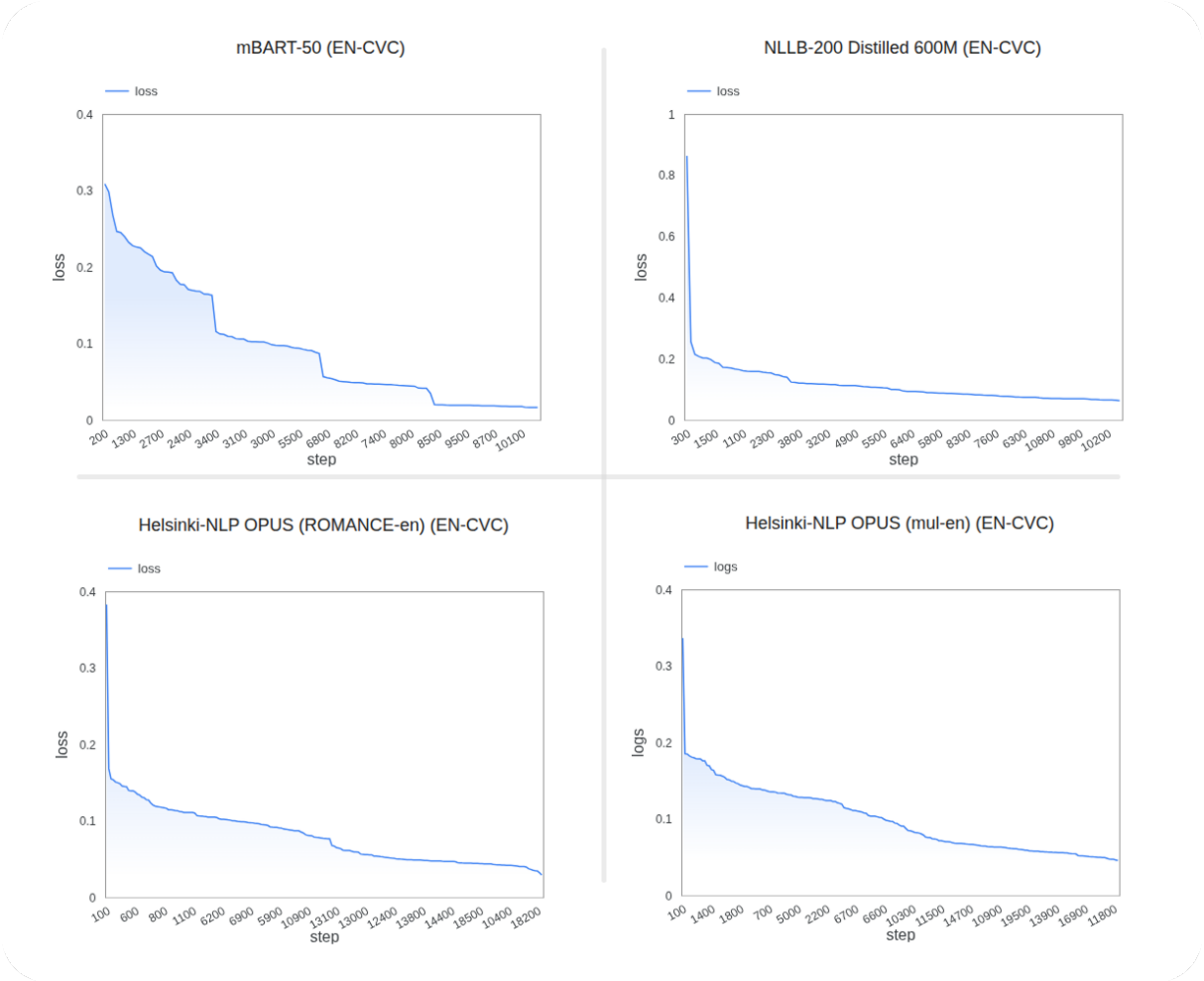

FIGURE 6.4. MT models loss during training to translate from English (en) to CVC (kea)

6.2.1. Discussion Evaluation Metrics Results

The results show that mBART-50 significantly outperformed all other models on every evaluation metric, with BLEU scores of 81.26 (mBART-$50_{(\text{kea-en})}$) and 83.49 (mBART-$50_{(\text{en-kea})}$), a large margin over the second-best models (Helsinki-NLP OPUS-ROMANCE-en variants), which achieved BLEU scores of only about 73. These results underscore the potential of multilingual sequence-to-sequence pretraining to account for the structural and orthographic variation associated with CVC (see Table 6.3).

mBART-50 also achieves higher chrF and METEOR scores, which means there is not only a higher n-gram lexical overlap at the surface level, but also a higher level of character-based similarity and semantic adequacy. Likewise, high BERTScore values (greater than 0.977) suggest a similar meaning with respect to the reference sentences. Similarly recognizing limited editing effort, lower TER scores (15.80 and 12.21) mean the machine outputs could be more usable for similar downstream applications.

On the other hand, the distilled NLLB-200 models performs significantly worse, with under 40 BLEU scores for NLLB-200-Distilled-600M$_{(\text{kea-en})}$ and under 30 for NLLB-200-Distilled-600M$_{(\text{en-kea})}$. Though NLLB has achieved strong performance in many low-resource scenarios, the relatively poor performance here indicates that its generalization may not work well when there is no direct training data for Creole, and when there is limited leverage of lexical similarities to the training languages.

The Helsinki-NLP Marian-based OPUS models are reasonably robust, especially those trained on ROMANCE-to-English pairs yielding BLEU scores over 70. They perform well in this regard, owing to the linguistic closeness between Portuguese and CVC. However, these models still fall short compared to the mBART-50 ones, which benefits from more advanced multilingual pretraining across diverse languages.

These findings not only showcase that mBART-50 is the best model (among the tested models), but they also demonstrate the overall quality and appropriateness of the Morabeza Corpus dataset. The strong results produced, compared to another low-resource translation project, specifically Tarifit [65], highlight that carefully curated data in tandem with a good model choice can provide quality translations exceeding previous work for similar underrepresented languages [1]. The dataset's balanced design, inclusion of multiple language pairs, and preprocessing ensure that the models can effectively learn mappings for both directions of translation.

In summary, according to the models metrics evaluation, the translation experiments prove that large parallel multilingual encoder-decoder architectures are capable of leading to state-of-the-art performance when fine-tuned on high-quality parallel data, even in extremely LRLs contexts like CVC. These findings also provide a clear foundation for future work on CVC translation beyond what has been discussed here, through hybrid methods that may blend the fine-tuning process described here with either in-context or prompt-based methods in training, which provides an additional verification of the methodological decisions made throughout this thesis.

6.2.2. Discussion Manual Testing Results

Tables 6.4 and 6.5 show translations across the four models, with background colors indicating quality: green (good translation), yellow (close / similar translation), orange (incomplete translation) and red (poor translation). Also the references are provided for comparison. This provide supplementary insights into the model's performance that automatic metrics are unable to show. These examples illustrate systematic tendencies of each system as well as qualitative strengths and weaknesses in their treatment of CVC.

When translating from CVC to English, mBART-50 provided translations that were consistently the closest to the reference translation in terms of both meaning and form. The mBART-50 outputs demonstrated fluency in English while maintaining a faithful translation of the meaning of the CVC source. The only difference was that in some

TABLE 6.4. Manual evaluation from CVC (kea) to English (en) translations

mBART-50	**NLLB-200-Distilled-600M**	**Helsinki-NLP-OPUS-ROMANCE-en**	**Helsinki-NLP-OPUS-mul-en**
Input (EN): ta parsi man ta fika duenti mas faxi ki otus algên **Reference**: i seem to get sick a little easier than other people			
i seem to get sick a little easier than other people	seems to get sick a little easier than other people	i look like i'm sick a lot easier than other people	it seems that the guy is sick but he does not do anything about other people
Input (EN): si bu tiver kes sintomas li, djobi bu asukar na sangi pa bu odja si ê pamo keli. **Reference**: if you have these symptoms, check your blood sugar level to see if that is the cause.			
if you have the following symptoms, check your blood sugar level to see if this is the cause.	if you have these symptoms, check your blood sugar to see if it's right.	if you have the following symptoms, check your blood sugar for your eyes if it is because of this.	if you have these symptoms, check your blood sugar to see if that is why.
Input (EN): ta manti en vigor ti kin revoga. **Reference**: remains in effect until i revoke it.			
remains in effect until revoke.	remains in effect until revoked.	remains in effect until i revoke.	it's still in use as wind.
Input (EN): bô ene dzem manera ke no tava te bem distribui es bolsa **Reference**: you didn't tell me how we would distribute these bags.			
you didn't tell me how we would distribute these bags.	you didn't tell me how we were going to distribute these bags	you didn't tell me how we would deliver this bag	you didn't tell me how we were going to distribute this bag

TABLE 6.5. Manual evaluation from English (en) to CVC (kea) translations

mBART-50	**NLLB-200-Distilled-600M**	**Helsinki-NLP-OPUS-ROMANCE-en**	**Helsinki-NLP-OPUS-mul-en**
Input (EN): i seem to get sick a little easier than other people **Reference**: ta parsi man ta fika duenti mas faxi ki otus algên			
ta parsi man ta duenti mas faxi ki otus algên	ta parsi ma ta duenti mas faxi ki otus algên	ta parse man fika doenti mas fasil ki otus genti	parse ma mim ta fika mal mas rapidi ki otus pessoa
Input (EN): if you have these symptoms, check your blood sugar level to see if that is the cause. **Reference**: si bu tiver kes sintomas li, djobi bu asukar na sangi pa bu odja si ê pamo keli.			
si bu tiver kes sintomas li, djobi bu asukar na sangi pa bu odja si ê pamo keli.	si bu tiver kês sintomas li, djobi bu nível di asukar na sangi pa bu odja si ê kauza li.	si bu tene kes sintomas, bu mesti djobi nivel di asucar pa odja razom	si bo ta sinti es sintomas li, kontrolla bu nivel di asucar pa konfirmal
Input (EN): remains in effect until i revoke it. **Reference**: ta manti en vigor ti kin revoga.			
ta manti en vigor ti kin revoga.	ta manti en vigor ti kin revoga.	ta fika en vigor ate mi revoga	ta kontina validu ti ku mi kansela
Input (EN): you didn't tell me how we would distribute these bags. **Reference**: bô ene dzem manera ke no tava te bem distribui es bolsa.			
bô ene dzem manera ke no tava te bem distribui es bolsa	bô ene dzêm manera ke no tava te distribuí es bolso.	bô ene dzêm manera no tava te bei distribui es bolsa.	bô ene dzêm manera no tava te bem faze pe distribui es bolsa.

translations, there were minor simplifications, even with verbs remaining the same. NLLB-200-Distilled-600M produced translations that were generally comprehensible and acceptable in English, however NLLB-200-Distilled-600M often omitted things or changed the structure of sentences enough to reduce adequacy. Helsinki-NLP OPUS-ROMANCE and Helsinki-NLP OPUS-multilingual differed in that they tended to show a significant lack of understanding: The OPUS systems produced translations that contained frequently incomplete or synthesized meanings which clearly reflected a significant deviation from the intended meaning. This translation output indicates an instructive reliance to transfer meaning from another Romance language which in turn became an interference versus a proper alignment from CVC to English.

The differences between models for translations to CVC were even more pronounced. As with English to CVC translations, mBART-50 had the highest quality and often produced outputs that could not be distinguished from the reference. In addition, mBART-50 was able to render grammatical structures and lexical items that are typical to CVC, which other models often did not. NLLB-200-Distilled-600M produced "adequate" translations, but was less consistent, with orthographic variation and always produced more rare lexical items as opposed to precision. On the other hand, Helsinki-NLP OPUS-ROMANCE and Helsinki-NLP OPUS-multilingual generated translation influenced by Portuguese or Spanish, introducing calques and unnatural constructions. Indeed some of models translations were semantically inaccurate or incomplete suggesting that the models had difficulty generalizing to CVC.

Overall, these findings indicate that larger multilingual systems are better equipped to handle a low-resource context, even if there is no adequate training data specific to CVC. In addition, however, these findings suggest that it is also necessary to build resources specifically for this context. While mBART-50 and NLLB-200-Distilled-600M get by with some reasonable results, they also create certain errors and inconsistencies that would not be suitable for a high-stakes task when translation accuracy is paramount. The qualitative differences noted through the manual evaluation intensify the argument that CVC continues to receive inadequate attention from the state of NMT technology, which is in line with previous analysis [1], and building resources specific to the language will be crucial to enable good translation quality.

6.2.3. Summary of Evaluation Results

The evaluation results provided in this study provide robust evidence that the combination of the Morabeza Corpus corpus and use of state-of-the art multilingual transformer-based architectures has dramatically improved the quality of MT for CVC. Compared to the authors previous research on CrioleSet [1], which utilized a smaller dataset of around 6,000 parallel pairs and classical sequence-to-sequence models such as Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) with gated attention, and Transformer-base, the current study resulted in substantial performance improvements, but was based on a larger dataset than the earlier work. In the previous study, while the Transformer-base model outperformed recurrent neural networks and indicated the benefits of attention mechanisms to low-resource translation, as noted, overall performance remained constrained due to data-scarcity and dialectal variability.

On the other hand, Morabeza Corpus both increases the size of the dataset and has wider dialectal coverage, but also employs cutting edge multilingual pre-trained models that offers significant gains in automatic metrics and human rating evaluations. mBART-50 is consistently the clear winner in this group of models achieving BLEU scores better than 80, near 90 for chrF and lower TER scores than in previous studies, while outperforming competitors in METEOR, ROUGE-L, and BERTScore. Manual assessments

verify that outputs align closely with the human references in fluency, idiomaticity, lexical accuracy and semantic accuracy. In contrast, Helsinki-NLP OPUS-ROMANCE and Helsinki-NLP OPUS-multilingual performed well on some measures, but outputs were often incomplete or awkward. NLLB-200-Distilled-600M patterns were more stable structurally, but not as close to the rated translations as mBART-50 on finer points of the translations. These results indicate that having more rich, diverse data along with high-performing multilingual architectures can greatly improve translation, consistent with prior evidence that pretraining on multilingual data improved outcomes in low-resource settings [18], [19], whereas training on narrow datasets struggles to generalize [51].

In light of earlier studies, in many cases in LRLs, reliance was placed either on small datasets for experimentation, or on strcutured evaluation protocols that did not closely consider semantic or idiomatic fidelity in assessment [1], [73]. In contrast, the present study integrates a well-defined, balanced dataset that allows for solid auto metrics evaluation, as well as rigorous well guided manual inspection. Drawing an important insight to highlight in this contribution, simply put, the performance differentials seen are an outcome of the quality or coverage of info included in the aggregation of training and evaluation material, not simply a matter of architectural choices for the models being assessed. So by offering a well-curated benchmark study, the present work fills a gap in the literature, suggesting the emphasis on working expeditiously towards developing datasets for addressing underrepresented languages is built on importance of establishing dataset for adequate testing, and from expertly generating constructed benchmarking evaluation.

In conclusion, the comparative evidence suggests that that large-scale multilingual pretraining, as demonstrated by mBART-50, greatly benefits LRLs, such as CVC. Additionally, the high agreement between metric-based evaluation and manual evaluation not only confirms the effectiveness of mBART-50, but it also validates the Morabeza Corpus dataset. In this sense, this research indicates that mBART-50 is currently the state-of-the-art CVC to English translation model, and additionally serves as a framework that can be replicated in future research in LRLs spaces to improve MT.

6.3. Models Deployment

For dissemination and demonstration, the best-performing MT model (mBART-50 fine-tuned on the Morabeza Corpus subset for MT) and the dialect classification models were deployed through Hugging Face Spaces, each accompanied by a dedicated Gradio application. This setup enables users to interact with the models directly through a web interface.

The MT model was exported with the following files: *config.json*, *generation_ config.json*, *model.safetensors*, *sentencepiece.bpe.model*, *special_ tokens_ map.json*, *tokenizer_ config.json*, and *tokenizer.json*.

The CVC variant classification models were exported with: *config.json*, *merges.txt*, *model.safetensors*, *special_ tokens_ map.json*, *tokenizer_ config.json*, *tokenizer.json*, and *vocab.json*.

To showcase the MT best model and the variant classifications models, two Gradio apps were developed and made available on Hugging Face Spaces: the CVC–English translation system[1] (Figure 6.6), and the dialectal variant classifier[2] (Figure 6.5). Both offer real-time outputs in an easy-to-use interface which enables the wider community to experiment with the model capabilities without requiring any technical knowledge.

This implementation approach ensures that the thesis findings are both reproducible and usable, showcasing the concrete impact of the developed NLP resources for CVC.

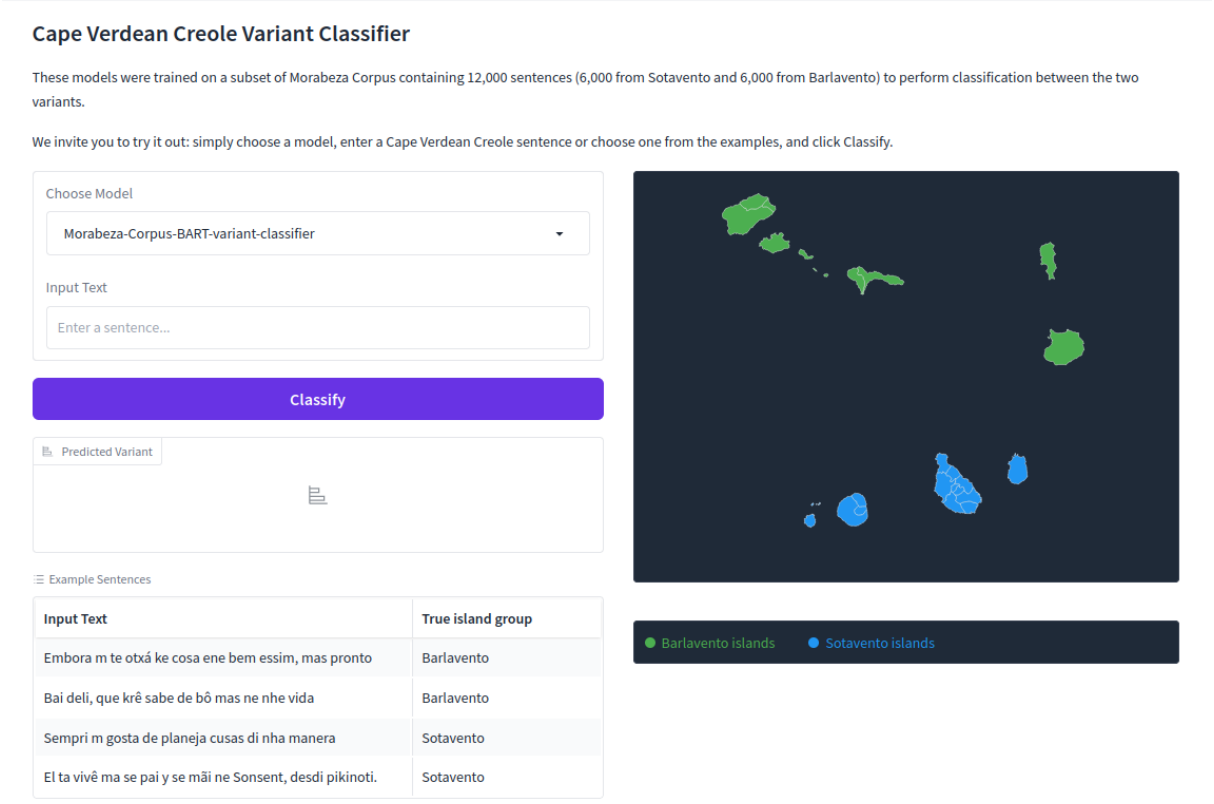

FIGURE 6.5. Gradio application demonstrating the CVC variant classifier. Users can input text in CVC and the app predicts the corresponding dialectal variant in real time.

[1]The variant translator application for demonstration can be found at `https://huggingface.co/spaces/morabeza/cape-verdean-creole-translator`

[2]The variant classifier application for demonstration can be found at `https://huggingface.co/spaces/morabeza/morabeza-corpus-variant-classifier`

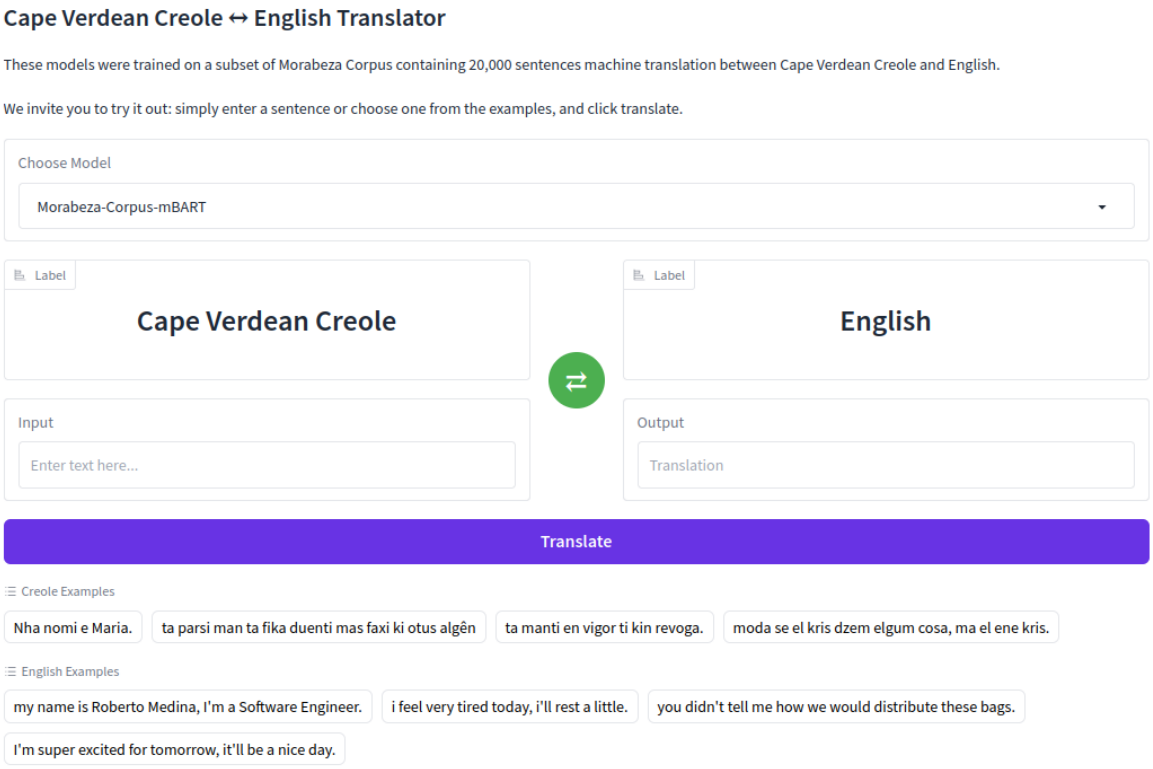

FIGURE 6.6. Gradio application demonstrating the CVC–English translation model. Users can input text in either Cape Verdean Creole or English and obtain the translation in real time.

CHAPTER 7

Conclusion

To the best of our knowledge, this thesis has presented one of the first broad efforts to build and assess NLP resources for CVC with a focus on translation tasks between CVC and English. The work has a direct contribution to making accessible a reliable, representative, and organized parallel corpus addressing the dearth of high-quality CVC resources, by presenting an extended and improved version of CrioleSet [1] called Morabeza Corpus. The dataset includes dialectal representation and has been cleaned and aligned to meet modern NLP pipeline definitions. In doing so, it answers **RQ1** directly by demonstrating that a systematic, replicable methodology can produce a resource that mediates between linguistic authenticity and computational use.

To address **RQ2**, the thesis thoroughly assessed a variety of MT methods, which included baseline models, transfer learning, and fine-tuning of multilingual pre-trained architectures. In all cases, the models, like mBART-50, when fine-tuned on Morabeza Corpus, provided state-of-the-art-quality outpourings, with notable fluency, idiomaticity, and semantic fidelity, in both automated and manual evaluations. This finding adds to the research on multi-lingual pre-training in low-resource environments, and is a significant contribution towards future research of CVC and English bilingual translation.

At the same time, this study investigated the automatic classification of CVC dialectal variants (Sotavento and Barlavento), thus addressing **RQ3**. The results indicate that classification models can distinguish between variants with promising accuracy, providing both a linguistic understanding and a practical tool for preprocessing and corpus enrichment. This is an important step toward making NLP systems treat CVC as internally diverse rather than monolithic.

In conclusion, while it provides a significant analysis of the issues that arise from the challenges experienced along the away, including lack of data, orthographic differences, and domain mismatch to pre-trained models, the study provides a roadmap for researchers who want to advance NLP development with deeply underrepresented and / or LRLs. By documenting and reflecting on these challenges, as well as offering solutions on how to work through these challenges, it answers **RQ4**.

To sum up, this dissertation has shown that although LRLs pose numerous challenges, it is still possible to build a reliable dataset, create high-quality MT, and investigate dialectal classification in CVC. This study has made contributions by directly answering the four research questions, illustrating the state of the art in CVC, NLP and within the field as a basis for sustainable future work. Striving to accentuate that every language,

no matter its size or power on the world stage, has a rightful place in the technological landscape, which is both a scientific urgency as well as a matter of personal importance.

7.1. Limitations

Though the study has merits, it has also a number of limitations. The data used, while bigger and more representative than past cases in the field, is nevertheless small by comparison to resource rich languages, which places an upper limit on the performance of the models we could achieve. Orthographic variation by dialects still makes data preprocessing, and human evaluation, difficult since annotators may have differing understandings of equivalence depending on the norms of their own dialects. Further, while fine-tuned multilingual models such as mBART-50 yielded impressive results, the fact that they needed pretrained models in unrelated languages to perform well exemplifies the ways in which low-resourced NLP relies on infrastructures outside of their own context or need. The dialect classification piece, while promising, faces similar challenges in the nuances of overlap, or code-switching, which are part of the socio-linguistic reality for CVC.

Another constraint was a limitation on computing resources used to train the MT models using larger parts of the dataset. Though Morabeza Corpus had more than 20.000 translation pairs, limitations on resources means only 20.000 pairs of those examples could be guaranteed during training. With greater computational resources, not only would be possible to use most or all of the original dataset, but it could have improved translations results even further. This also reflects on the classification models, were more data could also be used to train and validate the models.

7.2. Future Directions

Following the above contributions, several prospects for future work arise. First, the critical next step is to enhance the scale and domain of the data set in order to propel new translation performance, and allow the model to be applied to further domains of NLP, such as speech recognition and dialogue systems. Additionally, multimodal resources, such as embedding audio-text alignment may enhance the representation of CVC's oral tradition and the ease of dialect classification.

Secondly, in addition to this, future work should explore more complex adaptation methods, such as parameter-efficient fine-tuning, or the incorporation of CVC specific linguistic features, which might be able to lessen model dependency on external pretraining corpora and improve interpretability to aspects of linguistic change. Moreover, the creation of open-source tools geared towards CVC preprocessing and normalization would assist grant work and community projects.

Furthermore, while this project deals with MT between CVC and English, Morabeza Corpus also contains translation pairs to Portuguese and French. An interesting avenue for future research would be to build CVC–Portuguese and CVC–French models and compare these results to CVC–English. In a similar vein, the dataset could be labelled to reflect

the Creole variant for each island as a useful extension. At this time, the dataset has only been separated in terms of Barlavento and Sotavento, but classification schemes based on islands would allow for more nuanced evaluation and analysis of dialect classification tasks.

Ultimately, this work demonstrates the need to engage the Cape Verdean community in co-constructing NLP resources. As the author and a Cape Verdean researcher, this project can be viewed, not just as an academic effort, but as a cultural and social project that helps ensure our language is not left out of the digital age. Therefore, lessons for the future include a focus on participatory approaches and involving educators, linguists, and community members to create and validate resources, and in so doing we can ensure NLP systems for CVC reflect both linguistic diversity and cultural identity.

References

[1] R. C. Medina, F. Batista, and E. Lima, "Crioleset: A new corpus for cape verdean creole, towards robust machine translation," in *Progress in Artificial Intelligence*, J. Valente de Oliveira, J. Leite, J. Rodrigues, J. Dias, and P. Cardoso, Eds., Cham: Springer Nature Switzerland, 2026, pp. 326–338, ISBN: 978-3-032-05179-0.

[2] I. Orife *et al.*, "Masakhane: NLP for african languages by africans," *arXiv preprint arXiv:2003.11529*, 2020.

[3] N. Raychawdhary, N. Hughes, S. Bhattacharya, G. Dozier, and C. D. Seals, "A transformer-based language model for sentiment classification and cross-linguistic generalization: Empowering low-resource african languages," in *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, 2023, pp. 1–5. DOI: 10.1109/AIBThings58340.2023.10292494.

[4] N. Raychawdhary, A. Das, S. Bhattacharya, G. Dozier, and C. D. Seals, "Optimizing multilingual sentiment analysis in low-resource languages with adaptive pretraining and strategic language selection," in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, 2024, pp. 1–5. DOI: 10.1109/ICMI60790.2024.10585876.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017. [Online]. Available: https://doi.org/10.48550/ARXIV.1706.03762.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423/.

[7] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, T. Linzen, G. Chrupała, and A. Alishahi, Eds., Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446. [Online]. Available: https://aclanthology.org/W18-5446/.

[8] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds., Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. [Online]. Available: https://aclanthology.org/D16-1264/.

[9] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. [Online]. Available: https://aclanthology.org/P19-1493/.

[10] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[11] G. Hinton, O. Vinyals, and J. Dean, *Distilling the knowledge in a neural network*, 2015. arXiv: 1503.02531 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1503.02531.

[12] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. [Online]. Available: https://aclanthology.org/2020.acl-main.747/.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *Roberta: A robustly optimized bert pretraining approach*, 2019. arXiv: 1907.11692 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1907.11692.

[14] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov, "XNLI: Evaluating cross-lingual sentence representations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 2475–2485. DOI: 10.18653/v1/D18-1269. [Online]. Available: https://aclanthology.org/D18-1269/.

[15] P. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk, "MLQA: Evaluating cross-lingual extractive question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 7315–7330. DOI: 10.18653/v1/2020.acl-main.653. [Online]. Available: https://aclanthology.org/2020.acl-main.653/.

[16] J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki, "TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages," *Transactions of the Association for Computational Linguistics*, vol. 8, M. Johnson, B. Roark, and A. Nenkova, Eds., pp. 454–470, 2020. DOI: 10.1162/tacl_a_00317. [Online]. Available: https://aclanthology.org/2020.tacl-1.30/.

[17] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "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*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. [Online]. Available: https://aclanthology.org/2020.acl-main.703/.

[18] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, M. Johnson, B. Roark, and A. Nenkova, Eds., pp. 726–742, 2020. DOI: 10.1162/tacl_a_00343. [Online]. Available: https://aclanthology.org/2020.tacl-1.47/.

[19] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," *arXiv preprint*, 2020, ML50 benchmark; multilingual finetuning over many directions. eprint: 2008.00401.

[20] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proceedings of ACL 2018, System Demonstrations*, F. Liu and T. Solorio, Eds., Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 116–121. DOI: 10.18653/v1/P18-4020. [Online]. Available: https://aclanthology.org/P18-4020/.

[21] J. Tiedemann and S. Thottingal, "OPUS-MT – building open translation services for the world," in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, and M. L. Forcada, Eds., Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, pp. 479–480. [Online]. Available: https://aclanthology.org/2020.eamt-1.61/.

[22] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, N. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, P. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. Sadagopan, H. Schwenk, B. Shao, H. Shi, J. Staiano, S. B. Sun, C. Tran,

B. Zoph, F. Guzmán, A. Fan, and S. Bhosale, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[23] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, 1950.

[24] A. Newell, J. C. Shaw, and H. A. Simon, "Elements of a theory of human problem solving," *Psychological Review*, vol. 65, no. 3, pp. 151–166, 1958.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[26] L. Deng, "Deep learning: Methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[27] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the EMNLP 2014*, 2014, pp. 1724–1734. [Online]. Available: http://arxiv.org/abs/1406.1078.

[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014. [Online]. Available: https://doi.org/10.48550/arXiv.1409.0473.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

[30] Z. Zhao, S. Li, R. Yang, and W. Sun, "Code-switching from english-chinese to chinese-english: An exploratory study," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 8182–8188. DOI: 10.18653/v1/2020.emnlp-main.663.

[31] M. Rios, W. Aziz, and K. Sima'an, "Deep generative model for joint alignment and word representation," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds., New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1011–1023. DOI: 10.18653/v1/N18-1092. [Online]. Available: https://aclanthology.org/N18-1092/.

[32] I. Adebara and M. Abdul-Mageed, "Towards afrocentric NLP for African languages: Where we are and where we can go," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3814–3841. DOI: 10.18653/v1/2022.acl-long.265. [Online]. Available: https://aclanthology.org/2022.acl-long.265/.

[33] B. Haddow, R. Bawden, A. V. Miceli Barone, J. Helcl, and A. Birch, "Survey of low-resource machine translation," *Computational Linguistics*, vol. 48, no. 3, pp. 673–732, Sep. 2022. DOI: 10.1162/coli_a_00446. [Online]. Available: https://aclanthology.org/2022.cl-3.6/.

[34] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 6282–6293. DOI: 10.18653/v1/2020.acl-main.560. [Online]. Available: https://aclanthology.org/2020.acl-main.560/.

[35] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. [Online]. Available: https://aclanthology.org/2020.acl-main.747/.

[36] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Transactions of the Association for Computational Linguistics*, vol. 7, L. Lee, M. Johnson, B. Roark, and A. Nenkova, Eds., pp. 597–610, 2019. DOI: 10.1162/tacl_a_00288. [Online]. Available: https://aclanthology.org/Q19-1038/.

[37] J. Lang, *Cape Verdean Creole - A Comparative Study of Sotavento and Barlavento Varieties* (Languages of the World/Materials). Munich: Lincom Europa, 2009.

[38] M. Baptista, *The Syntax of Cape Verdean Creole: The Sotavento Varieties* (Creole Language Library). Amsterdam/Philadelphia: John Benjamins Publishing Company, 2002.

[39] N. Alexandre, "Cape verdean creole," in *The Survey of Pidgin and Creole Languages, Volume III: Pidgins and Creoles beyond Africa and the Atlantic*, S. M. Michaelis, P. Maurer, M. Haspelmath, and M. Huber, Eds., Oxford University Press, 2015, pp. 43–52.

[40] L. Mompelat, "Recommendations for overcoming linguistic barriers in healthcare: Challenges and innovations in NLP for Haitian Creole," in *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, Š. A. Holdt, N. Ilinykh, B. Scalvini, M. Bruton, I. N. Debess, and C. M. Tudor, Eds., Tallinn, Estonia: University of Tartu Library, Estonia, Mar. 2025, pp. 20–31. [Online]. Available: https://aclanthology.org/2025.resourceful-1.6/.

[41] H. Lent, K. Tatariya, R. Dabre, Y. Chen, M. Fekete, E. Ploeger, L. Zhou, R.-A. Armstrong, A. Eijansantos, C. Malau, H. E. Heje, E. Lavrinovics, D. Kanojia, P.

Belony, M. Bollmann, L. Grobol, M. d. Lhoneux, D. Hershcovich, M. DeGraff, A. Søgaard, and J. Bjerva, "CreoleVal: Multilingual multitask benchmarks for creoles," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 950–978, 2024. DOI: 10.1162/tacl_a_00682. [Online]. Available: https://aclanthology.org/2024.tacl-1.53/.

[42] C. Zirn, A. Lüdeling, J. Nicolas, and B. Gehrke, "Creole talk: Investigating the use of mauritian creole in a social network," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association, 2016, pp. 4231–4235. [Online]. Available: https://aclanthology.org/L16-1683.

[43] R. Dabre and A. Sukhoo, "KreolMorisienMT: A dataset for mauritian creole machine translation," in *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, Eds., Online only: Association for Computational Linguistics, Nov. 2022, pp. 22–29. DOI: 10.18653/v1/2022.findings-aacl.3. [Online]. Available: https://aclanthology.org/2022.findings-aacl.3/.

[44] W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Fagbohungbe, S. O. Akinola, S. Muhammad, S. Kabongo Kabenamualu, S. Osei, F. Sackey, R. A. Niyongabo, R. Macharm, P. Ogayo, O. Ahia, M. M. Berhe, M. Adeyemi, M. Mokgesi-Selinga, L. Okegbemi, L. Martinus, K. Tajudeen, K. Degila, K. Ogueji, K. Siminyu, J. Kreutzer, J. Webster, J. T. Ali, J. Abbott, I. Orife, I. Ezeani, I. A. Dangana, H. Kamper, H. Elsahar, G. Duru, G. Kioko, M. Espoir, E. van Biljon, D. Whitenack, C. Onyefuluchi, C. C. Emezue, B. F. P. Dossou, B. Sibanda, B. Bassey, A. Olabiyi, A. Ramkilowan, A. Öktem, A. Akinfaderin, and A. Bashir, "Participatory research for low-resourced machine translation: A case study in African languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 2144–2160. DOI: 10.18653/v1/2020.findings-emnlp.195. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.195/.

[45] Ž. Agić and I. Vulić, "JW300: A wide-coverage parallel corpus for low-resource languages," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3204–3210. DOI: 10.18653/v1/P19-1310. [Online]. Available: https://aclanthology.org/P19-1310/.

[46] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A.

Moreno, J. Odijk, and S. Piperidis, Eds., Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2214–2218. [Online]. Available: https://aclanthology.org/L12-1246/.

[47] D. I. Adelani, D. Ruiter, J. O. Alabi, D. Adebonojo, A. Ayeni, M. Adeyemi, A. E. Awokoya, and C. España-Bonet, "The effect of domain and diacritics in Yoruba–English neural machine translation," in *Proceedings of Machine Translation Summit XVIII: Research Track*, K. Duh and F. Guzmán, Eds., Virtual: Association for Machine Translation in the Americas, Aug. 2021, pp. 61–75. [Online]. Available: https://aclanthology.org/2021.mtsummit-research.6/.

[48] A. El-Kishky, V. Chaudhary, F. Guzmán, and P. Koehn, "CCAligned: A massive collection of cross-lingual web-document pairs," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 5960–5969. DOI: 10.18653/v1/2020.emnlp-main.480. [Online]. Available: https://aclanthology.org/2020.emnlp-main.480/.

[49] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The Flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Transactions of the Association for Computational Linguistics*, vol. 10, B. Roark and A. Nenkova, Eds., pp. 522–538, 2022. DOI: 10.1162/tacl_a_00474. [Online]. Available: https://aclanthology.org/2022.tacl-1.30/.

[50] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds., Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1568–1575. DOI: 10.18653/v1/D16-1163. [Online]. Available: https://aclanthology.org/D16-1163/.

[51] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds., Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 86–96. DOI: 10.18653/v1/P16-1009. [Online]. Available: https://aclanthology.org/P16-1009/.

[52] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2020.

[53] T. Jauhiainen, M. Lui, M. Zampieri, T. Baldwin, and K. Lindén, "Automatic language identification in texts: A survey," *Journal of Artificial Intelligence Research*, vol. 65, pp. 675–782, 2019.

[54] F. Faisal *et al.*, "Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages," *Proceedings of the 62nd Annual Meeting of the Association for*

Computational Linguistics, pp. 1–12, 2024. DOI: 10.18653/v1/2024.acl-long.777. [Online]. Available: https://aclanthology.org/2024.acl-long.777/.

[55] A. Joshi *et al.*, "Natural language processing for dialects of a language: A survey," *arXiv preprint arXiv:2401.05632*, 2024. [Online]. Available: https://arxiv.org/abs/2401.05632.

[56] S. Ruder, "An overview of multi-task learning in deep neural networks," in *arXiv preprint arXiv:1706.05098*, 2017.

[57] W. Adouane, "When is multi-task learning beneficial for low-resource noisy texts?" In *Proceedings of the CALCS 2020 Workshop*, 2020, pp. 1–10. [Online]. Available: https://aclanthology.org/2020.calcs-1.3/.

[58] A. Hande *et al.*, "Findings of the shared task on multi-task learning in dravidian languages," in *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, 2022, pp. 1–10. [Online]. Available: https://aclanthology.org/2022.dravidianlangtech-1.43/.

[59] N. Robinson, R. Dabre, A. Shurtz, R. Dent, O. Onesi, C. Monroc, L. Grobol, H. Muhammad, A. Garg, N. Etori, V. M. Tiyyala, O. Samuel, M. Stutzman, B. Odoom, S. Khudanpur, S. Richardson, and K. Murray, "Kreyòl-MT: Building MT for Latin American, Caribbean and colonial African creole languages," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 3083–3110. DOI: 10.18653/v1/2024.naacl-long.170. [Online]. Available: https://aclanthology.org/2024.naacl-long.170/.

[60] M. R. Fekete, E. Lavrinovics, N. R. Robinson, H. Lent, R. Dabre, and J. Bjerva, "Leveraging adapters for improved cross-lingual transfer for low-resource creole MT," in *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, J. Sälevä and A. Owodunni, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 212–215. DOI: 10.18653/v1/2024.mrl-1.17. [Online]. Available: https://aclanthology.org/2024.mrl-1.17/.

[61] F. Zheng, M. Reid, E. Marrese-Taylor, and Y. Matsuo, "Low-resource machine translation using cross-lingual language model pretraining," in *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, M. Mager, A. Oncevay, A. Rios, I. V. M. Ruiz, A. Palmer, G. Neubig, and K. Kann, Eds., Online: Association for Computational Linguistics, Jun. 2021, pp. 234–240. DOI: 10.18653/v1/2021.americasnlp-1.26. [Online]. Available: https://aclanthology.org/2021.americasnlp-1.26/.

[62] A. L. Tonja, H. H. Nigatu, O. Kolesnikova, G. Sidorov, A. Gelbukh, and J. Kalita, "Enhancing translation for indigenous languages: Experiments with multilingual models," in *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, M. Mager, A. Ebrahimi, A.

Oncevay, E. Rice, S. Rijhwani, A. Palmer, and K. Kann, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 200–205. DOI: 10.18653/v1/2023.americasnlp-1.22. [Online]. Available: https://aclanthology.org/2023.americasnlp-1.22/.

[63] S. M. Lakew, M. Cettolo, and M. Federico, "A comparison of transformer and recurrent neural networks on multilingual neural machine translation," in *Proceedings of the 27th International Conference on Computational Linguistics*, E. M. Bender, L. Derczynski, and P. Isabelle, Eds., Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 641–652. [Online]. Available: https://aclanthology.org/C18-1054/.

[64] S. Sarkar, S. Kashyap, A. Joglekar, and S. Umesh, "Effectively combining phi-4 and NLLB for spoken language translation: SPRING lab IITM's submission to low resource multilingual Indic track," in *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, E. Salesky, M. Federico, and A. Anastasopoulos, Eds., Vienna, Austria (in-person and online): Association for Computational Linguistics, Jul. 2025, pp. 399–404, ISBN: 979-8-89176-272-5. DOI: 10.18653/v1/2025.iwslt-1.42. [Online]. Available: https://aclanthology.org/2025.iwslt-1.42/.

[65] O. Akallouch and K. Fardousse, "In-context learning for low-resource machine translation: A study on tarifit with large language models," *Algorithms*, vol. 18, no. 8, p. 489, 2025. DOI: 10.3390/a18080489.

[66] M. Fouadi *et al.*, "Bert-based models for classifying multi-dialect arabic texts," *Indonesian Journal of Artificial Intelligence*, 2024. [Online]. Available: https://ijai.iaescore.com/index.php/IJAI/article/view/23878/0.

[67] D. Yadav, S. Suravee, T. Strauß, and K. Yordanova, "Cross-lingual named entity recognition for low-resource languages: A Hindi-Nepali case study using multilingual BERT models," in *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, J. Sälevä and A. Owodunni, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 167–174. DOI: 10.18653/v1/2024.mrl-1.12. [Online]. Available: https://aclanthology.org/2024.mrl-1.12/.

[68] V. K. Agbesi, W. Chen, S. B. Yussif, M. A. Hossin, C. C. Ukwuoma, N. A. Kuadey, C. C. Agbesi, N. Abdel Samee, M. M. Jamjoom, and M. A. Al-antari, "Pre-trained transformer-based models for text classification using low-resourced ewe language," *Systems*, vol. 12, no. 1, 2024, ISSN: 2079-8954. DOI: 10.3390/systems12010001. [Online]. Available: https://www.mdpi.com/2079-8954/12/1/1.

[69] A. Salleh, M. H. Osman, S. Hassan, M. Y. Said, K. Y. Sharif, and K. T. Wei, "A hybrid model for low-resource language text classification and comparative analysis," *Knowledge-Based Systems*, vol. 326, p. 114 068, 2025, ISSN: 0950-7051. DOI: https:

//doi.org/10.1016/j.knosys.2025.114068. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095070512501113X.

[70] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, "Neural versus phrase-based machine translation quality: A case study," *CoRR*, vol. abs/1608.04631, 2016. arXiv: 1608.04631. [Online]. Available: http://arxiv.org/abs/1608.04631.

[71] E. Vanmassenhove, D. S. Shterionov, and A. Way, "Lost in translation: Loss and decay of linguistic richness in machine translation," *CoRR*, vol. abs/1906.12068, 2019. arXiv: 1906.12068. [Online]. Available: http://arxiv.org/abs/1906.12068.

[72] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org.

[73] P. Koehn, *Statistical Machine Translation*. Cambridge University Press, 2009.

[74] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[75] S. Banerjee and A. Lavie, "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*, 2005, pp. 65–72.

[76] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004, pp. 74–81.

[77] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations (ICLR)*, 2020.

[78] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of Association for Machine Translation in the Americas (AMTA)*, 2006, pp. 223–231.

[79] M. Popović, "Chrf: Character n-gram f-score for automatic mt evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 2015, pp. 392–395.

[80] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021, pp. 549–562.