ISCTE IUL
REPOSITÓRIO
INSTITUTO UNIVERSITÁRIO DE LISBOA

Repositório ISCTE-IUL

Deposited in *Repositório ISCTE-IUL*:
2026-09-28

Deposited version:
Accepted Version

Peer-review status of attached file:
Peer-reviewed

Citation for published item:
Medina, R., Batista, F. & Lima, E. (2025). CrioleSet: A new corpus for Cape Verdean creole, towards robust machine translation. In José Valente de Oliveira, João Leite, João Rodrigues, João Dias, Pedro Cardoso (Ed.), Progress in Artificial Intelligence: 24th EPIA Conference on Artificial Intelligence, EPIA 2025 (pp. 326?338). Springer Nature. https://doi.org/10.1007/978-3-032-05179-0

Further information on publisher's website:
10.1007/978-3-032-05179-0

Publisher's copyright statement:
This is the peer reviewed version of the following article: Medina, R., Batista, F. & Lima, E. (2025). CrioleSet: A new corpus for Cape Verdean creole, towards robust machine translation. In José Valente de Oliveira, João Leite, João Rodrigues, João Dias, Pedro Cardoso (Ed.), Progress in Artificial Intelligence: 24th EPIA Conference on Artificial Intelligence, EPIA 2025 (pp. 326?338). Springer Nature. https://doi.org/10.1007/978-3-032-05179-0, which has been published in final form at https://dx.doi.org/10.1007/978-3-032-05179-0. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

Use policy

Creative Commons CC BY 4.0
The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Serviços de Informação e Documentação, Instituto Universitário de Lisboa (ISCTE-IUL)
Av. das Forças Armadas, Edifício II, 1649-026 Lisboa Portugal
Phone: +(351) 217 903 024 | e-mail: administrador.repositorio@iscte-iul.pt
https://repositorio.iscte-iul.pt

CrioleSet: A New Corpus for Cape Verdean Creole, Towards Robust Machine Translation

Roberto Carlos Medina[1,2][0009−0005−7623−8781], Fernando Batista[1,2][0000−0002−1075−0177], and Estanislau Lima[3][0000−0002−8373−5357]

[1] ISCTE – Instituto Universitário de Lisboa, Portugal
https://www.iscte-iul.pt/
[2] INESC-ID Lisboa, Portugal
https://inesc-id.pt/
[3] UTA - Atlantic Technical University, Cabo Verde
https://uta.cv/

Abstract. This paper presents CrioleSet, a new parallel corpus designed to facilitate neural machine translation (NMT) for Cape Verdean Creole (CVC), a low-resource language spoken by the majority of Cape Verdeans. Comprising over 6,000 translation pairs in English, Portuguese, French, and CVC, the dataset addresses the scarcity of annotated resources for CVC, which is further challenged by dialectal variation across the archipelago. We trained and evaluated three neural network architectures LSTM, GRU with gated attention (GAtt), and Transformer—base to perform English-CVC translation tasks. Experimental results demonstrate that the Transformer-base models significantly outperforms the others, achieving the highest BLEU and METEOR scores and the lowest TER, reflecting better translation quality and robustness. This confirms that attention-based architectures can effectively handle low-resource translation, even with relatively modest datasets. The work underscores the potential of focused data curation and deep learning to advance NLP resources for underrepresented languages. Future work includes expanding the dataset to encompass more dialects and refining models for broader NLP tasks in CVC, thereby contributing to linguistic inclusivity and cross-cultural communication.

Keywords: Machine Translation · Natural Language Processing · Deep Learning · dataset

1 Introduction

Although Portuguese is still the nation's official language, the vast majority of Cape Verdeans speak CVC as their mother tongue. In terms of computational linguistic resources, CVC has received little attention despite being an essential part of Cape Verdean culture and everyday communication. Standardization and language processing tasks are especially difficult because each of the nine inhabited islands in the archipelago has developed its own variant of CVC.

Given the expanding tourism industry of the nation, more than 819,999 hotel guests were recorded in 2019 [1], most of them from English, French, and Portuguese speaking nations, language barriers between residents and tourists can make communication difficult. Systems that can translate between CVC and other commonly spoken languages are becoming more and more necessary because many Cape Verdeans do not speak English, or other languages, fluently. Additionally training professional translators is expensive and time consuming, automated solutions are a desirable substitute.

Recent developments in machine translation and natural language processing (NLP), especially with regard to Deep Learning (DL) techniques, have greatly improved the quality of translation systems [2][3]. However, these models usually need extensive annotated datasets, which low-resource languages like CVC frequently lack. For NLP to be more inclusive and to benefit underrepresented languages, this gap must be closed.

A study on the creation of DL-based models for automatic translation between English and CVC is presented in this paper. As part of this work, a custom dataset called CrioleSet was created, which contains over 6,000 translation pairs in English, Portuguese, French, and CVC. We assess and contrast three neural architectures for translation: Transformer [3], GAtt (GRU-Gated Attention) [4][5], and LSTM [6].

The development of NLP resources for CVC is hindered by several key challenges. These include limited annotated datasets, the lack of a standardized orthography, and the presence of multiple dialects that complicate linguistic analysis [7]. Additionally, as a low-resource language, CVC faces difficulties in leveraging existing NLP technologies, which are predominantly designed for high-resource languages like English, Spanish, and Portuguese [8][9].

Low-resource languages often struggle with data scarcity, making it difficult to train robust NLP models [10]. Transfer learning and multilingual models, such as mBERT [11][12], offer potential solutions, but their effectiveness in languages with significant linguistic variation remains an open question. Furthermore, the absence of computational tools tailored for CVC limits applications in essential NLP tasks such as machine translation [13], named entity recognition [14], and sentiment analysis [15].

The main contributions of this paper are as follows: (1) We present CrioleSet, a carefully selected parallel dataset that consists of more than 6,000 translation pairs in English, Portuguese, French, and CVC; (2) We implement and compare three DL-based MT models: LSTM, GAtt, and Transformer; (3) We evaluate the models using common metrics (BLEU, METEOR, TER) and show that the Transformer model achieves the highest performance; and (4) We show that, by creating targeted datasets and adapting DL, it is possible to develop machine translation systems for underrepresented languages.

2 Related Work

This section outlines key developments in NLP and MT for low-resource languages and reviews deep learning-based translation models relevant to this study.

2.1 NLP and MT for Low-Resource Language

Due in large part to the availability of deep learning architectures and large-scale datasets, NLP has advanced significantly over the last ten years. However, high-resource languages like English, Chinese, and French have disproportionately benefited from this advancement, with large language datasets [16][17]. The dearth of standardized orthographies, linguistic tools, and annotated corpora for low-resource languages like CVC limits NLP research in these languages.

Three methods are usually used in efforts to develop NLP resources for under-represented languages: data augmentation, transfer learning, and manual corpus creation [18]. Notably, transfer learning has demonstrated encouraging outcomes in tasks like machine translation and part-of-speech tagging for morphologically complex or non-standardized languages [19].

2.2 Deep Learning in Machine Translation

Rule-based systems gave way to statistical machine translation (SMT) and, more recently, NMT in the evolution of machine translation [20]. Translation performance underwent a sea change with the introduction of encoder-decoder architectures utilizing Recurrent Neural Networks (RNNs), such as LSTM [6] and GRU [4][21]. This was especially true when combined with attention mechanisms [3][5].

RNN-based systems have limitations when it comes to modeling long-range dependencies, despite their benefits. The Transformer architecture [3], which is based solely on self-attention mechanisms and has since emerged as the state-of-the-art in NMT tasks, was introduced to address this shortcoming.

According to recent comparative studies, the Transformer performs better than RNN-based models in terms of training efficiency and translation quality, especially when applied to larger datasets [16]. Its performance on small, low-resource datasets, like those for Creole languages, is still understudied, though, and this paper attempts to address that.

3 The CrioleSet Dataset

The creation of CrioleSet, a parallel corpus of sentences in English and CVC as well as other languages, is a key element of this work. It was created to facilitate NMT tasks. The dataset was created from scratch using a combination of manual translation, semi-automatic alignment, and carefully selected content from both spoken and written sources because there was no publicly accessible corpus before this work.

3.1 Language Overview

Spoken throughout the nine inhabited islands of Cape Verde, CVC is a Creole language with a Portuguese foundation. Every island has its own variant with lexical, phonological, and syntactic differences, such as São Vicente's Sanpadjudu or Santiago's Badiu. Despite their differences, native speakers can still understand each other. However, natural language processing is hampered by the coexistence of several dialects and the absence of a standard orthography.

Text samples from the São Vicente and Santo Antão variants were used to create the dataset for this study in an effort to maintain common linguistic traits and guarantee cross-dialect comprehensibility.

3.2 Data Collection and Preprocessing

The CrioleSet dataset, presented in the scope of this article, includes translation pairs between CVC and three languages: Portuguese, English, and French. In addition to translating a few English phrases from the Multi30K [17] dataset, Creole sentences were also gathered from actual message exchanges between the author and other people (with permission from about 20 people).

The languages were chosen strategically: French because of a similarly high rate of tourists from France ($\approx$ 10.4% in 2019 [1]); English because of its international status and the high influx of UK tourists to Cape Verde ($\approx$ 24% in 2019 [1]); and Portuguese because it is the linguistic foundation of Creole.

Translations were done sequentially from Creole to Portuguese, English, and French after sentence validation and any necessary corrections. Examples of translations for all four languages are shown in Table 1, and token distributions are displayed in Table 2. More than 6,000 sentence pairs make up the final dataset, which is divided into **6,000 for training**, **222 for validation**, and **183 for testing**.

Table 1. Paired examples of translation of dataset CrioleSet

Sentence	Cape Verdean Creole	Portuguese	English	French
S1	Algo xtronhê tite contsê ei.	Algo estranho está acontecendo aqui.	Something weird's going on here.	Quelque chose de bizarre se passe ici.
S2	M te otxá ke bô ê mut exigente.	Eu acho que você é muito exigente.	I think you're too picky.	Je pense que tu es trop pointilleux.
S3	M percebê ke m tava te precisa de ajuda.	Eu percebi que precisava de ajuda.	I realized I needed help.	J'ai réalisé que j'avais besoin d'aide.

Table 2. Number of tokens per language in the CreoleSet dataset

Language	Number of Tokens
Cape Verdean Creole	5154
Portuguese	5251
English	4135
French	4810

4 Methodology

4.1 Model Architectures

Three distinct neural models, namely LSTM, GAtt, and Transformer-based, that are frequently employed in sequence-to-sequence (seq2seq) translation tasks were put into practice and assessed. All of the models are encoder-decoder based and were modified to work with token embeddings and variable-length sequences. An example of seq2seq models architecture can be seen in Figure 1 [22].

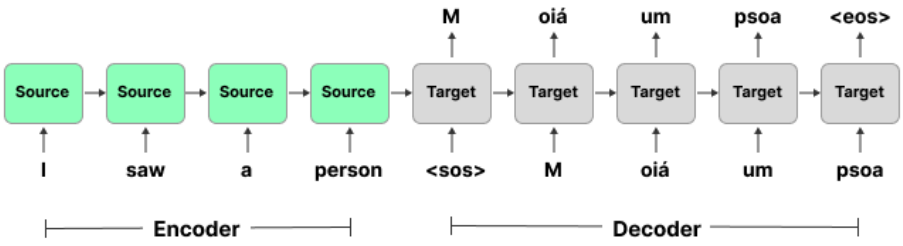


Fig. 1. Basic Sequence-by-Sequence Model

LSTM Long Short-Term Memory (LSTM) model, [6] is a type of RNN capable of capturing long-range dependencies in sequential data. The encoder reads the input CVC sentence and generates a fixed-length context vector, which the decoder uses to produce the corresponding English sentence token by token, using the sequence-to-sequence architecture on Figure 1.

GAtt (GRU-Gated Attention) The GAtt model integrates Gated Recurrent Units (GRU) [4] with an attention mechanism [5]. GRUs offer simpler computation than LSTMs, while attention helps focus on relevant parts of the input sequence at each decoding step, improving translation quality.

Transformer The Transformer model [3] eliminates recurrence entirely and uses multi-head self-attention and feedforward networks. Its parallelizable architecture accelerates training and yields state-of-the-art performance in translation

tasks. We use a 4-layer Transformer with positional encoding, and we also fine-tuned a pre-trained Portuguese model, from spacy package, to perform the MT task, as Portuguese is the closest language to CVC.

Three standard metrics are used to evaluate the translation quality: BLEU, which measures n-gram precision with a brevity penalty [23]; METEOR, which accounts for stemming, synonymy, and word order [24]; and TER, which calculates the minimum number of edits needed to match the reference translation [25].

4.2 Training Configuration

All models were trained using the same train-validation-test split of the CrioleSet dataset. The models were trained using gradient clipping set to 1, batch size of 10, and learning rate of $3e^{-4}$. 256-dimensional embedding layers were used in all models. The Transformer used 256 hidden units, compared to 1024 in the LSTM and GAtt models. Eight attention heads, four encoder and decoder layers, and 512-dimensional encoder embeddings were added to the Transformer's configuration. During training, dropout was used, with a value of 0.5 for GAtt and LSTM and 0.3 for all models. Training was monitored by early stopping. The Transformer had 260 training epochs, GAtt had 230, and LSTM had 160. In Table 3, we can see more information about models training.

Moving forward, finding, for example, $\text{GAtt}_{(\text{cv-en})}$ means that a GAtt model was trained to translate from CVC to English, and $\text{GAtt}_{(\text{en-cv})}$ the other way around.

Table 3. Model training data

Model	**Number of trainable parameters**	**Number of training epochs**	**Training time (minutes)**
$\text{GAtt}_{(\text{cv-en})}$	16,472,222	250	80
$\text{GAtt}_{(\text{en-cv})}$		220	65
$\text{LSTM}_{(\text{cv-en})}$	29,133,470	156	45
$\text{LSTM}_{(\text{en-cv})}$		156	45
$\text{Transformer}_{(\text{cv-en})}$	6,722,772	150	70
$\text{Transformer}_{(\text{en-cv})}$		250	118

5 Results and Discussion

As already mentioned, in Table 3, for each model (GAtt, LSTM and Transformer) two different models are trained to perform machine translation tasks from CVC to English (cv-en), and from English to CVC (en-cv). Therefore, for each of these trained models there are results regarding their training parameters, in Table 4 you can see the results of all these models where it is possible to see that the Transformer models, $\text{Transformer}_{(\text{cv-en})}$ and $\text{Transformer}_{(\text{en-cv})}$, are

the models that present the best results over the others, and the models that obtained the worst results were the LSTM models, $\text{LSTM}_{(\text{cv-en})}$ and $\text{LSTM}_{(\text{en-cv})}$. Although the Transformer models ($\text{Transformer}_{(\text{cv-en})}$ and $\text{Transformer}_{(\text{en-cv})}$) do not attain the highest validation accuracy during training, we can see in Figure 2 that they perform better than the others in the final testing phase, yielding the best results throughout.

Table 4. Accuracy and Loss Model training and validation measures

	Accuracy		**Loss**	
Model	**Training**	**Validation**	**Training**	**Validation**
$\text{LSTM}_{(\text{cv-en})}$	72.96%	72.68%	16.11%	9.33%
$\text{LSTM}_{(\text{en-cv})}$	71.15%	72.02%	20.97%	10.91%
$\text{GAtt}_{(\text{cv-en})}$	74.58%	74.74%	15.85%	8.00%
$\text{GAtt}_{(\text{en-cv})}$	72.92%	74.02%	19.96%	9.83%
$\text{Transformer}_{(\text{cv-en})}$	**44.51%**	**84.21%**	**11.12%**	**3.50%**
$\text{Transformer}_{(\text{en-cv})}$	**44.81%**	**83.57%**	**12.17%**	**4.20%**

5.1 Quantitative Results

All models were evaluated using BLEU, METEOR, and TER on the held-out test set. As shown in Table 5, the Transformer model consistently outperformed both LSTM and GAtt across all metrics in both translation directions (cv-en and en-cv).

Table 5. Metrics

Model	**BLUE**	**METEOR**	**TER**
$\text{LSTM}_{(\text{cv-en})}$	34.95%	66.43%	51.95%
$\text{LSTM}_{(\text{en-cv})}$	31.49%	68.83%	50.95%
$\text{GAtt}_{(\text{cv-en})}$	37.27%	73.69%	30.02%
$\text{GAtt}_{(\text{en-cv})}$	32.44%	71.38%	32.70%
$\text{Transformer}_{(\text{cv-en})}$	**41.63%**	**75.53%**	**27.01%**
$\text{Transformer}_{(\text{en-cv})}$	**36.72%**	**73.80%**	**31.01%**

The Transformer achieved the highest BLEU scores ($\text{Transformer}_{(\text{cv-en})}$: 41.2%; $\text{Transformer}_{(\text{en-cv})}$: 39.8%), indicating stronger n-gram overlap with the reference translations. It also scored highest in METEOR and lowest in TER, confirming its superior performance in terms of semantic accuracy and edit efficiency.

These results support the robustness of the Transformer architecture, even in low-resource settings when regularization and early stopping are applied effectively.

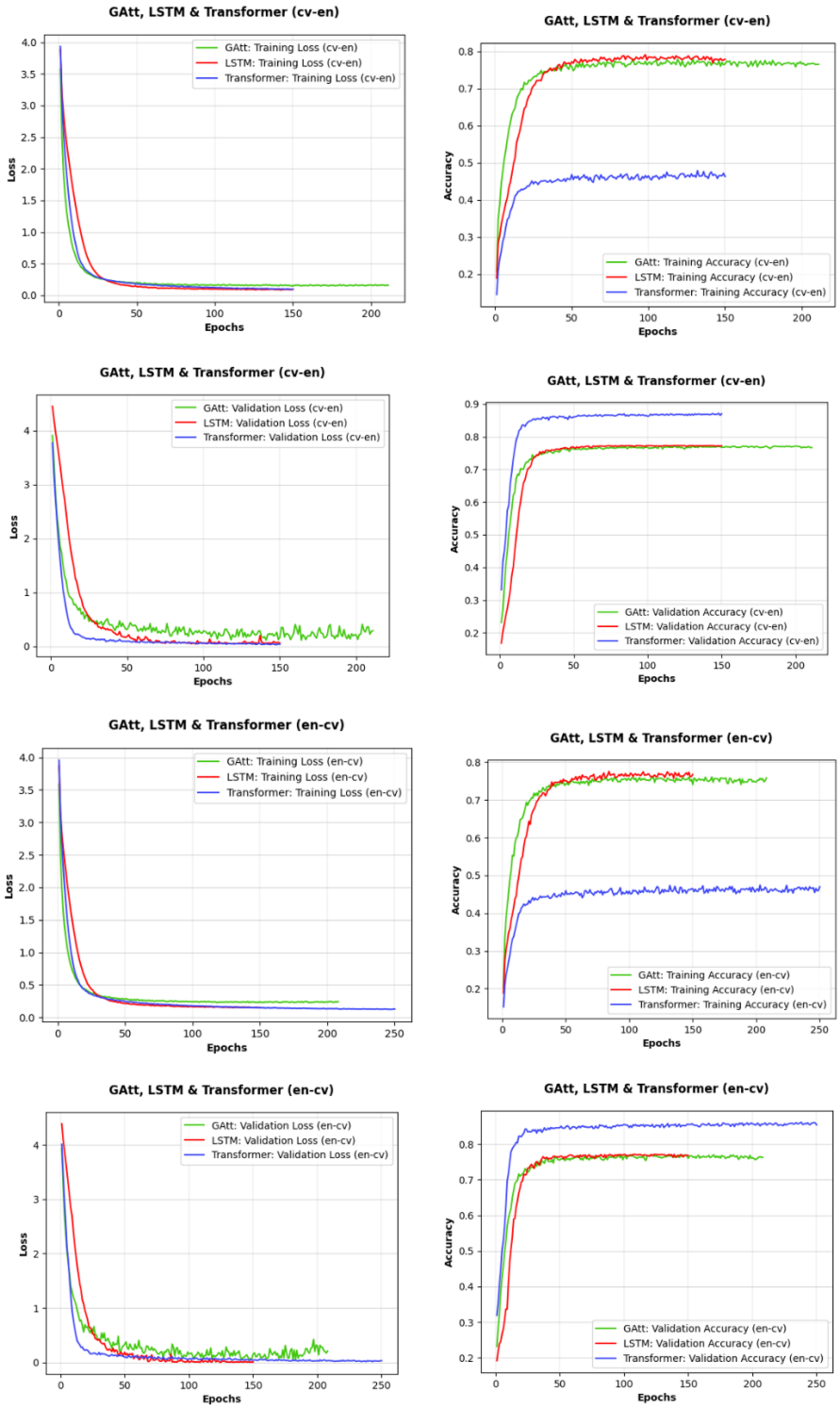


Fig. 2. Models accuracy and loos during training and evaluation in both directions (cv-en and en-cv)

5.2 Comparative Analysis

Although LSTM and GAtt models showed decent performance, they were notably less stable during training. The LSTM model exhibited higher variance in both training and validation losses, especially in the $\text{LSTM}_{(\text{en-cv})}$ direction. GAtt provided marginal improvements over plain GRUs by incorporating attention but was still limited in generalization capability.

Using the self-attention mechanism allows the Transformer to better capture global dependencies, which is particularly useful for languages like CVC that exhibit flexible word order and variation in morphology. Its faster convergence and lower overfitting risk made it more suitable for the CrioleSet's modest scale.

5.3 Qualitative Examples

Selected translation outputs demonstrate the Transformer's superiority not only in metrics but also in fluency and context preservation. Below are representative samples translated by each model.

Table 6. Model Translations examples

Example (CVC)	**System**	**Translation (EN)**
M te bem fca casa, pe pode relaxá.	Reference	I'm going to stay at home, so I can relax.
	Transformer	I'll stay home, to relax.
	GAtt	I stay at the house.
	LSTM	I'm in a house.
Talvez porkê nhe irmá ê advogada, agoh um kris ser també.	Reference	Maybe because my sister is a lawyer, now I want to be too.
	Transformer	Maybe because my sister is a lawyer, now I wanted to be too.
	GAtt	Maybe why lawyer, I want too
	LSTM	The and it is to. The I be was the of, and want.
Kel casal antig sentá lód a lód.	Reference	The old couple sat side by side.
	Transformer	The old couple sat side by side.
	GAtt	`The old sat side.`
	LSTM	diferent .

As seen above, the Transformer model produces contextually fluent translations, even when vocabulary or syntax diverge from English norms. Meanwhile, the LSTM model often generates overly literal or ambiguous phrases.

In the reverse direction (en-cv), similar trends were observed. The Transformer captured idiomatic expressions more naturally, while LSTM and GAtt models tended toward phrase-level literalism.

These results demonstrate that while all three architectures are viable for low-resource translation, Transformer models offer a clear advantage in both reliability and translation quality when applied with modest-sized but well-curated datasets like CrioleSet.

6 Limitations

Although this study demonstrates the feasibility of developing NMT systems for CVC, several limitations must be acknowledged. These are primarily related to the scope of the dataset, dialectal diversity, model robustness, and resource constraints.

Just over 6,000 sentence pairs make up CrioleSet, which is sufficient for a proof of concept but insufficient for training neural translation models that can recognize intricate syntactic and semantic patterns. The rich morphosyntactic diversity of CVC dialects is not fully reflected in the dataset, despite its careful linguistic curation. Performance may suffer on underrepresented or unseen variants since CVC differs greatly among islands in vocabulary, grammar, and phonology, and the models were not specifically trained to identify or adjust to these dialectal differences.

The trained models, especially Transformers and GAtt, demonstrate strong learning within the scope of the data set, but it is not clear how well they generalize to other contexts or genres. Informal or poetic language, domain-specific vocabulary (e.g., legal, medical), and idiomatic expressions are likely to pose challenges. Moreover, the absence of a standardized orthography introduces additional ambiguity in the tokenization and embedding processes.

7 Conclusion and Future Work

This paper presents a new corpus for Cape Verdean Creole and presents a deep learning-based approach to the bidirectional machine translation between CVC and English. Addresses the critical lack of computational resources and tools for this underrepresented language. The CrioleSet parallel corpus, comprises more than 6,000 sentence pairs, and was created to support model training and evaluation.

Three neural models, LSTM, GAtt, and Transformer-based, were implemented and compared using standard evaluation metrics (BLEU, METEOR, TER). The Transformer model significantly outperformed the others, producing more fluent and semantically accurate translations. These results confirm that even with limited training data, deep learning techniques, especially attention-based architectures, can effectively handle the translation needs of low-resource languages.

Beyond technical achievements, this work demonstrates that meaningful progress in NLP for Creole languages is possible with focused data curation, modest computational resources, and adaptation of state-of-the-art models.

The following steps include adding more dialects from different Cape Verdean islands and diverse sources to the CrioleSet dataset, training with more sophisticated state-of-the-art models, and integrating a wider variety of NLP tasks like named entity recognition and classification. As a first step toward improving NLP for CVC, the objective is to offer more comprehensive linguistic resources and tools.

The development of CrioleSet and the implementation of deep learning–based translation models mark a significant step toward addressing the systemic lack of resources for underrepresented languages in NLP. Cape Verdean Creole, like many African and Creole languages [15][26], has historically been excluded from mainstream computational linguistic research due to its low resource status, dialectal variation, and lack of standardized orthography.

The methodologies developed here can be extended to other low-resource settings, offering a replicable path toward language preservation, cross-cultural communication, and democratization of NLP technology.

Acknowledgments. This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020 (DOI:10.54499/UIDB/50021/2020).

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

8 References

[1] INE. Cape verde statistical yearbook 2019 – tourism. Instituto Nacional de Estatística, 2019. last accessed 2025/05/24.

[2] Li Deng. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, 2017.

[4] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the EMNLP 2014*, pages 1724–1734, 2014.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[7] Rahaf Abdul-Nabi, Ra'ed Obeidat, and Abdallah Bsoul. A survey on machine translation of low-resource arabic dialects. In *2024 15th International Conference on Information and Communication Systems (ICICS)*, pages 1–6, 2024.

[8] Pratik Joshi, Shruti Santy, Anjali Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL 2020*. Association for Computational Linguistics, jul 2020.

[9] Damian Blasi, Antonis Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of ACL 2022 (Volume 1: Long Papers)*, pages 5486–5505. Association for Computational Linguistics, may 2022.

[10] David Ifeoluwa Adelani, Dana Ruiter, Jesujoba O. Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. The effect of domain and diacritics in Yoruba–English neural machine

translation. In Kevin Duh and Francisco Guzmán, editors, *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual, August 2021. Association for Machine Translation in the Americas.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of NAACL 2019: HLT, Volume 1*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] Alexis Conneau, Kartik Khandelwal, Naman Goyal, Vishal Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*, pages 8440–8451. Association for Computational Linguistics, jul 2020.

[13] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Katrin Erk and Noah A. Smith, editors, *Proceedings of ACL 2016*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[14] Tao Xei, Qiang Li, Yansong Zhang, Zihan Liu, and Hongyuan Wang. Self-improving for zero-shot named entity recognition with large language models. In *Proceedings of NAACL 2024: HLT (Volume 2: Short Papers)*, pages 583–593. Association for Computational Linguistics, jun 2024.

[15] Nikita Raychawdhary, Arpan Das, Soumyajit Bhattacharya, Glenn Dozier, and Craig D. Seals. Enhancing sentiment analysis in amharic: Leveraging transformer-based language model for low-resource african languages. In *SoutheastCon 2024*, pages 50–55, 2024.

[16] Loïc Barrault, Marta Biesialska, Ondrej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Sheri Pal, and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, 2020.

[17] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, aug 2016.

[18] Antonis Anastasopoulos and Graham Neubig. Pushing the limits of low-resource morphological inflection. In *Proceedings of ACL 2019*, pages 2942–2952, 2019.

[19] Oliver Adams, Graham Neubig, Trevor Cohn, and Steven Bird. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 55th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 1100–1111, 2017.

[20] Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Jan Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, 2014.

[21] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.

[22] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 2020.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2022*, pages 311–318, 2002.

[24] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings ACL 2005*, pages 65–72, 2005.

[25] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, 2006.

[26] Nikita Raychawdhary, Arpan Das, Soumyajit Bhattacharya, Glenn Dozier, and Craig 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)*, pages 1–5, 2024.