Leveraging Portuguese Latent Spaces for Kriolu to English Translation with Monolingual Language Model Priors

Kevin Santiago
kevin711@mit.edu

Matthew Wilmot
mcwilmot@mit.edu

Peter Lin
plin503@mit.edu

Abstract

This project seeks to investigate low-resource translation for Cape Verdean Kriolu by leveraging its historical and structural proximity to Portuguese and improving English generation with a language model prior. We propose an encoder-decoder framework that combines Portuguese-based transformer encoders as the latent space with frozen English language models as decoder priors. We evaluate three Portuguese encoders (BERTimbau, Albertina PT-PT, and Albertina PT-BR) paired with three English models (BERT-base, RoBERTa-base, and Electra), yielding nine encoder-decoder configurations. Models are trained with AdamW using cross-entropy loss augmented by a KL-divergence regularizer that aligns translation model outputs with softened English LM distributions. Performance is assessed using BLEU and chrF on validation and test splits, alongside a baseline employing a randomly initialized latent-space encoder. We demonstrate improvements in validation loss and test BLEU and chrF scores across all pretrained Portuguese encoder pairings relative to the baseline, with BERT-base-Portuguese delivering the most consistent BLEU gains and Brazilian Portuguese latent spaces showing particularly strong chrF performance. These results suggest that related language latent representations, combined with an English LM prior, can meaningfully improve translation quality and semantic stability in low-resource settings.

A. Introduction

In the current world environment, neural machine translation systems (NMT) have become the default infrastructure for multilingual communication through natural language. Yet, the practical coverage of these systems is heavily skewed towards more commonly known high-resource languages with over-abundancies of written text. In fact, languages that are primarily oral and lack stable orthography are often invisible in commercial machine translation (MT) systems [8]. This is somewhat counter-intuitive, since accurate MT is arguably better served focusing on conversion between a known and lesser known language rather than between two languages that are popular and easy to learn.

Cape Verdean Creole, more commonly called Kriolu or Kabuverdianu, is a low-resource dialect with an estimated 90–95 percent of its vocabulary derived from Portuguese [1]. At the same time, its grammar differs sharply from standard European or Brazilian Portuguese in that verbs have minimal inflection, tense and aspect are expressed via preverbal particles, and features such as double or even triple negation are commonplace [1]. For MT, this means that systems are faced with not only the task of translating this language, but also being robust enough to handle these variations despite an extremely small amount of parallel data.

As it stands, conventional low-resource MT approaches tend to treat each language as an independent unit and often under-exploit known genealogical and lexical relationships [2, 8]. This is very problematic for Kriolu because it ignores the obvious fact that many tokens are derived from Portuguese and that high-capacity encoders trained on Portuguese already exist. As a result, we propose a different strategy. Rather than asking a translation model (TM) to discover structure from a small and noisy corpus, we propose injecting prior knowledge at both ends of the model. Because Kriolu lexicon and word order are so strongly anchored in Portuguese, we hypothesize that mapping Kriolu strings into some sort of Portuguese latent space will help the model generalize across orthographic variants and recognize typical Portuguese substrings even when they appear in creole spelling. Therefore in order to do this, we train a Kriolu→English MT under an LM-prior framework where

(i) the encoder is initialized from either a Portuguese BERT model (BERTimbau) or an Albertina encoder trained on European or Brazilian Portuguese [9, 11] , (ii) the decoder is regularized by a pre-trained English LM (BERT-base, RoBERTa-base, or ELECTRA-base) [3,4,6], and (iii) training takes place on the Kabuverdianu subset of the Kreyòl-MT corpus [8].

From a modeling perspective, we believe that this approach is particularly well-suited to capturing Kriolu's variation from Portuguese. Investigating standard subword embedding methods such as BPE would reveal that they are agnostic to genealogical relationships and learn merges purely from surface frequency. However, in a creole like Kriolu, where the same Portuguese-derived lexeme can appear across a spectrum of more phonetic or more Portuguese-like spellings, this often results in inconsistent segmentations that fail to tie the related forms together. Furthermore, character-level or byte-level models avoid fixed vocabularies but pay for this with longer sequences, higher data requirements, and no explicit way to actually exploit the fact that many tokens are noisy Portuguese rather than arbitrary strings.

B. Related Work

B.1. Kriolu Syntax and Structural Background

Descriptive work on Cape Verdean Kriolu provides the linguistic foundation for our modeling choices. In her book, Baptista's detailed analysis of the Sotavento varieties highlights several core properties that distinguish Kriolu from its Portuguese lexifier. These include a relatively fixed subject-verb-object (SVO) word order, preverbal tense–aspect–mood (TAM) participles, and the use of multiple negation strategies [1]. It is these structural rules that make it clear that Kriolu cannot just be treated as "noisy Portuguese" at the grammatical level, and instead need to be treated as additional features of the language itself [1]. In the context of our research, this suggests that models must be able to leverage lexifier information without collapsing Kriolu onto standard Portuguese. Hence that is why our approach explicitly assumes this split. As said before, we rely on Portuguese-based encoders to provide robust lexical and orthographic priors, whilst leaving room in the learned representations to capture rest of Kriolu specific syntax.

Creole languages frequently have a large variety in writing convention from speaker to speaker, leading to the same word holding multiple different spellings [1]. This can be seen in Cape Verde, where variations across the Sotavento (southern) and Barlavento (northern) islands cause words such as "handsome" to be spelled as "bunitu", "bnit", and "benite".

B.2. Using Language Models as Priors for Low Resource MT

Our approach is directly inspired by the language model prior framework introduced by Baziotis et al. [2]. In their formulation, an external monolingual LM over the target language is treated as a prior, and the NMT model is regularized via a Kullback–Leibler (KL) divergence term that penalizes discrepancies between the TM decoder's output distribution and the LM's distribution. This approach differs from classical shallow or deep fusion approaches in two crucial ways. Firstly, the LM does not require joint decoding with the translation model at inference time, and secondly, the prior acts only on the loss function itself which biases learning toward regions of the output space that correspond to fluent target language text [2]. In fact, the results of their research show that in low-resource conditions, this prior can improve both BLEU and training stability across multiple language pairs. As a result, we had confidence in the approach's potential to help regulate how much non-targeted noise we add during finetuning (to avoid the pitfalls discussed in the previous section).

Beyond Baziotis et al., a growing body of research has demonstrated that pretrained LMs can serve as powerful inductive biases for low-resource MT. Cold fusion and its extensions [12] incorporate a pretrained LM during training to stabilize decoder behavior, while work on LM-guided decoding [13] shows that target-side distributional knowledge can substantially improve fluency and reduce hallucinations. Other approaches integrate LMs through pseudo-monolingual regularization [10], effectively using monolingual dataset to augment scarce parallel data. In low-resource settings, monolingual LMs provide a robust structural prior that reduces overfitting and increases output quality. Using this research, we further explore supports the idea to employ a frozen English LM as a prior, ensuring that improvements stem from genuine target-language knowledge rather than artifacts of the limited Kriolu–English corpus.

B.3. Portuguese Latent Spaces as Lexifier Encoders

Finally, our modification to this approach is to add another prior in the form of a latent space prior upstream. Recent transformer encoders for Portuguese supply precisely this kind of lexifier-specific representation that our setting requires. Souza et al. introduce BERTimbau, a family of BERT-base models trained on large Brazilian Portuguese corpora and show that they achieve strong performance on a range of downstream Portuguese NLP tasks [8]. BERTimbau inherits the architecture and training objectives of BERT but is exposed exclusively to Portuguese text, resulting in subword segmentations and contextual embeddings that capture orthographic regularities and lexical patterns specific to Portuguese [11]. Building on this, Rodrigues et al. propose the Albertina PT-* models which

include encoders trained on both European (PT-PT) and Brazilian (PT-BR) Portuguese [9]. They frame Albertina as an attempt to level up Portuguese foundational models in order to close the gap between English and Portuguese in terms of general purpose encoders [9].

C. Methodology

C.1. Dataset

We use the Kriolu to English portion of the Kreyòl-MT corpus as our source of parallel data. In our code, we load it directly from HuggingFace [8] with the language pair kea-eng and then flatten the translation field into typical source and target columns. We also keep the original train, validation, and test splits and remove only the malformed examples. This produces a small but realistic dataset of 1470 sentences for training, 84 sentences for validation and 164 sentences for metric evaluation.

As mentioned before, Kriolu is a spoken dialect which often has a wide range of writing conventions from island to island. Because of this lack of an official writing scheme, and reliance on Portuguese for official written documents, there is a severe lack of resources for the language. NLLB-200, [5] is a MT model supporting over 200 languages. Included in this research is a larger Kriolu dataset. However, it relies on a translation of the Bible into Kriolu, one of the few pieces of text written in the language. This led to an inaccurate dataset, as many of the direct translations were replace for indirect, culturally equivalent sayings and analogies.

C.2. Architecture

Our architecture follows the LM-prior framework introduced by Baziotis et al., adapting it to the low-resource Portuguese→Kriolu translation setting. However, we incorporate a Portuguese-initialized encoder to strengthen cross-lingual transfer. The core component of the system is a Transformer-based Translation Model (TM) that encodes Cape Verdean Kriolu input sequences and autoregressively embeds output tokens in a Portuguese latent space. To improve representational quality in this low-data regime, the encoder is initialized with weights from a Portuguese BERT model. This Portuguese-initialized TM provides a linguistically grounded latent space aligned with the source language, enabling the model to more effectively capture morphological and syntactic structure as it translates Kriolu. The architecture remains fully swappable: different TM configurations or pretrained Portuguese models can be substituted without modifying the overall framework.

While this is going on, a monolingual English language model serves as an auxiliary prior during training. For each target English sequence associated with a Kriolu reference, the LM generates its own next-token distribution based solely on the gold history. This LM operates independently from the TM and does not observe the source Portuguese input. When combined, the TM and LM produce two complementary probability distributions: the TM distribution reflects source-conditioned translation dynamics, while the LM distribution reflects the fluency constraints derived from high-resource English data. It is important to note that the LM branch is used only during training, rather than during inference time.

The TM and LM interact through a KL-based regularization pathway. The hidden representations produced by the TM encoder flow through the decoder to yield token-level logits, while the LM simultaneously produces its own logits for the same timestep. The KL pathway between these two distributions encourages the TM to align with fluent target-side patterns without overriding its capacity to diverge from traditional English when necessary for accurate translation.

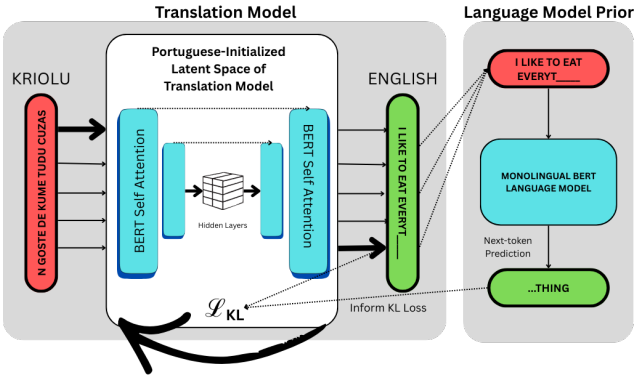

Figure 1. Conceptual diagram showing TM architecture with upstream and downstream priors.

C.3. Loss function

Following Baziotis et al. [2], at each decoding step t, given the Kriolu source sentence x and the gold English prefix $y_{<t}$, the TM produces a distribution

$$p_\theta(\cdot \mid x, y_{<t})$$

over the next token, while the LM produces its own distribution

$$p_{\text{LM}}(\cdot \mid y_{<t}).$$

Through this, it is able to capture distributional regularities of English without conditioning on the source sentence. The TM and LM are still in distinct representational spaces, but their outputs are linked through a KL-Divergence. For this, both distributions are softened via a temperature parameter τ, producing $\tilde{p}_\theta$ and $\tilde{p}_{\text{LM}}$, and resulting in the divergence

loss:

$$\mathcal{L}_{KL} = \tau^2 \, \mathbb{E}_t \Big[\mathrm{KL}\Big(\tilde{p}_\theta(\cdot \mid x, y_{<t}) \,\Big\|\, \tilde{p}_{\mathrm{LM}}(\cdot \mid y_{<t}) \Big) \Big] ,$$

We can then integrate the loss into the full training objective:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(p_\theta, y) \; + \; \lambda_{\mathrm{kl}} \, \mathcal{L}_{KL}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the standard cross-entropy loss on the gold target tokens. $\tilde{p}_\theta$ and $\tilde{p}_{\mathrm{LM}}$ are softened by a temperature τ via softmax over logits divided by τ, and λ_{kl} controls the strength of the prior.

The LM prior is instantiated via `AutoModelForMaskedLM` using the same decoder model name as the TM, all LM parameters are frozen, and the LM is kept in evaluation mode throughout training so that it functions purely as a fixed prior and is never fine-tuned on the Kriolu corpus.

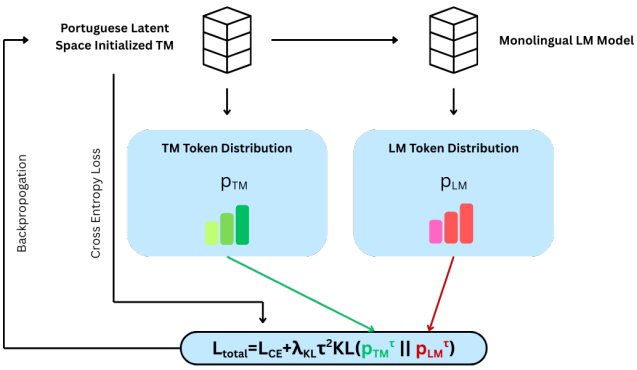

Figure 2. Conceptual diagram showing Loss calculation.

D. Experiments

D.1. Models

As mentioned previously, we use a combination of Portuguese-based transformer encoders and English-based language models for the architecture. Specifically, we rely on 3 Portuguese models and 3 English models. For the latent space, we use the BERTimbau model, the Albertina PT-PT model, and the Albertina PT-BR model. It is important to note that BERTimbau and Albertina PT-BR are both trained in Brazilian Portuguese, whereas Albertina PT-PT specializes in European Portuguese. There are key linguistic and formality differences between both, which may have an influence in which Kriolu can map more easily to.

For the English LMs and decoders, we utilize BERT (specifically BERT-base), RoBERTa (also base), and Electra. Both BERT and Electra use a WordPiece tokenizer,

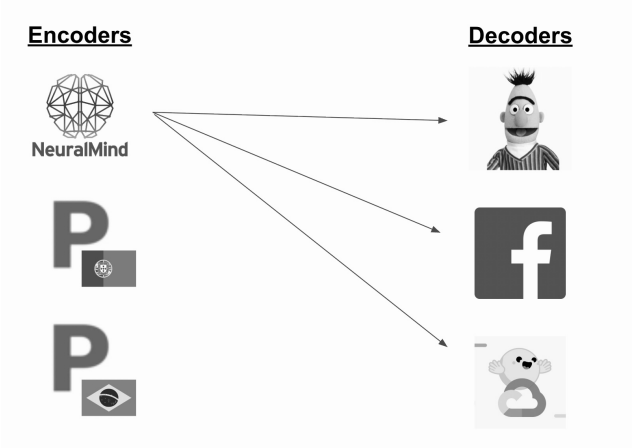

Figure 3. Model Pairings: Each encoder is paired with every decoder, resulting in 9 pairings. Albertina PT-PT and Albertina PT-BR are distinguished by the country flag

while RoBERTa uses Byte Pair Encoding. All the models average around 110-125 million parameters. The vocab size ranges from 50k to 30k, but didn't have an impact on the final results.

Each Portuguese model is paired with each of the English models to evaluate which pairings perform better or worse. This allows us to view how our model architecture and training procedure perform with a variety of different base models, especially compared to the general baseline. To add to this, we will also be able to observe if differences in the latent space brought about by the Brazilian and European Portuguese models play a role in the models ability to learn Kriolu

D.2. Model Configurations

We train the encoder–decoder pair using the AdamW optimizer. We initialize the training environment with a learning rate of 5×10^{-5} and mini-batches of size 32. Additionally, we employ early stopping based on validation loss with a patience of 12 epochs and a minimum improvement threshold of 0.01. Training is terminated if the model doesn't decrease in validation over the course of 12 epochs. This is to prevent the model from overfitting to the training set and performing worse. Furthermore, we implement a hard cap of 100 epochs for training. This is due to a lack of computational resources and time constraints.

For each epoch, we compute:

1. the standard cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ from the TM (as returned by the HuggingFace model),
2. the LM prior KL loss $\mathcal{L}_{\mathrm{KL}}$ using the softened TM and LM distributions with temperature $\tau = 2.0$,
3. the combined loss $\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{kl}} \mathcal{L}_{\mathrm{KL}}$ where $\lambda_{\mathrm{kl}} = 0.5$.

D.3. Training Procedure

For each patch, the encoder processes the source sequence to produce contextualized hidden states, which the decoder then consumes autoregressively to predict the next English token. Since the LM prior is independent of the translation model, the gold English prefix is passed separately through the frozen LM to obtain the reference distribution for each position. Furthermore, the TM and LM logits are both softened using the same temperature, and the KL divergence term is computed only over non-padding positions to ensure that the prior influences learning only where the model has real supervision. At validation time however, we evaluate the TM without applying the LM prior. This comes from the fact that the KL term is a training only regularizer [2]. The model predicts English tokens using beam search and computes validation loss strictly using the TM's own cross-entropy.

D.4. Evaluation

After training, we evaluate the final checkpoint on the held-out test split using token-level loss and corpus-level BLEU [6]. Moreover, we also evaluate the pairings on chrF [7] for further measure. Because BLEU compares values using n-grams of words, the model can often be harsh on morphologically rich languages like Kriolu and Portuguese. Additionally, the wide range and unstandardized spelling convention of Kriolu can lead to lower performance despite having an understandable translation. However, BLEU tends to be more stable with shorter data and still captures translation fluency. In contrast, by using character n-grams, chrF can be more robust to morphologically rich languages, but can also overestimate the quality of a model. Both BLEU and chrF are computed from model-generated translations produced with a maximum length of 64 tokens and a beam size of 4.

In practice, we generally observe that BLEU scores are more conservative, often penalizing translations that are semantically correct but use alternative word forms or word orders, which is common in Kriolu and Portuguese. chrF, on the other hand, tends to give higher scores in these cases, since it captures partial matches at the character level and is less sensitive to exact word boundaries. Therefore, it is useful to consider both metrics together: BLEU provides a stricter measure of n-gram precision and fluency, while chrF can highlight a model's ability to produce translations that preserve meaning despite surface differences.

D.5. Baseline

To assess the impact of initializing the Translation Model with a Portuguese latent space, we construct a baseline system that removes all source-language pretraining. In this baseline, the encoder–decoder architecture is randomly initialized across all layers, providing a Sequence-to-Sequence model with no prior exposure to Portuguese or related Romance-language structure. The remainder of the framework is kept identical to the proposed model to ensure a controlled comparison. In particular, the baseline TM is paired with the same frozen BERT-base English LM prior, instantiated via `AutoModelForMaskedLM`, in order to attribute differences in performance solely to the presence or absence of Portuguese pretraining rather than to changes in the LM prior or optimization setup.

The baseline system is trained under the same low-resource conditions and with the same training procedure as the Portuguese-initialized model. We allow the model to train for 30 epochs, using the same KL-regularized objective and the same teacher-forcing decoding strategy during training. After training, we evaluate the baseline using both BLEU and chrF, enabling a direct performance comparison across metrics that capture complementary aspects of translation quality. This baseline provides a reference point for quantifying the contribution of source-language pretraining to translation performance in the Portuguese→Kriolu setting.

E. Results

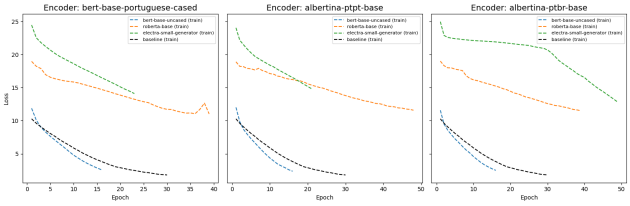

Figure 4. Training Results of Model Pairings

As mentioned above, the models had variable training epochs based on their validation performance. Although the models with the RoBERTa-base English LM tended to train for longer, with the Albertina PT-PT and RoBERTa-base training for a joint longest 48 epochs, they often started with a higher training loss and validation loss. In contrast, models with a BERT-base English LM tended to reach their minimum validation quickly and terminate with a training loss just behind the Baseline (2.46 vs 1.78). Overall, based on the similarity in training in Figure (num), most of the training trajectories were influenced by the English LM rather than the latent space. This could be representative of the influence the KL Divergence loss, and LM predictions, had on the training process.

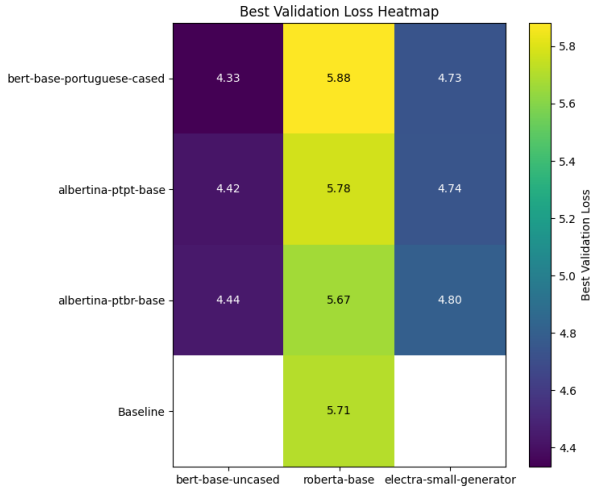

Figure 5. Best Validation Losses during Training

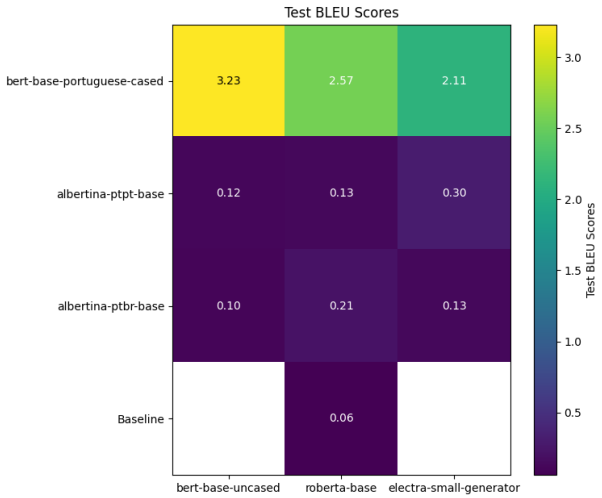

Figure 7. Test BLEU scores

Although the models trained worse to the baseline, 7 out of 9 models ended with a better validation loss over the course of training. As seen in training, models with the BERT-base English LM performed better independent of the latent space. Additionally, BERT-base-portuguese, one of the brazilian portuguese latent spaces tended to perform better compared to the other encoders (with the exception of RoBERTa-base).

Both the validation and test BLEU scores saw an immense outperformance by the pairing with the BERT-base-portuguese model. Each pairing with the model achieved a minimum 2 point increase in the BLEU score. In contrast, the pairing of Albertina PT-BR and BERT-base performed worse on both BLEU evaluations (0.13 and 0.10). However, all models were able to pass the baseline BLEU score, which received an extremely low score despite the low training loss.

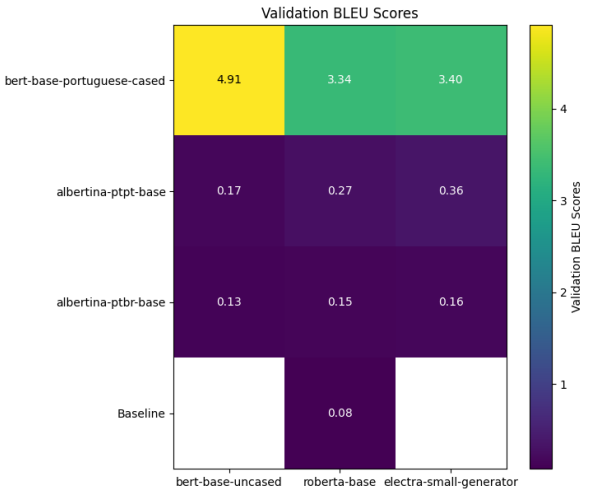

Figure 6. Validation BLEU scores

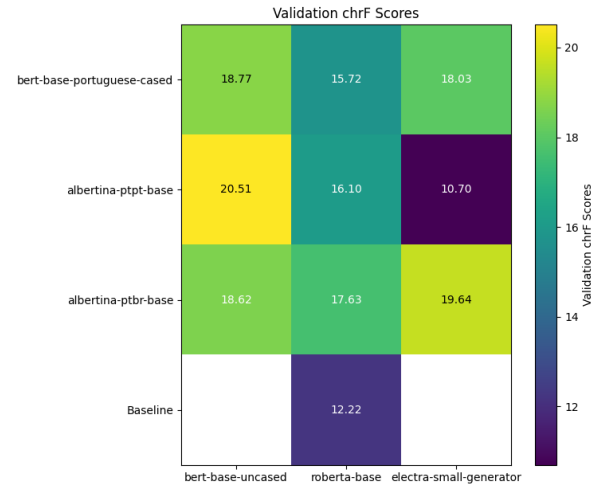

Figure 8. Validation chrF scores

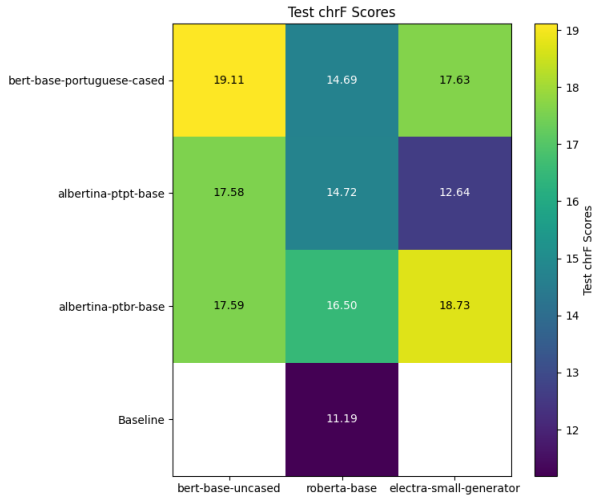

Figure 9. Test chrF scores

Similar to the BLEU scores, models with a BERT-base-portuguese latent space performed well overall. However, models with Albertina PT-BR ended up outperforming the other pairings. Altogether, the Brazilian Portuguese latent space tended to help with translation from the perspective of character analysis. All models were able to pass the Baseline chrF score for both validation and test

We believe the BERT-base-Portuguese encoder performed consistently well compared to the other pairings across both BLEU and chrF because it provides substantially richer and more accurate representations of Portuguese input text. Unlike the ALBERT models, whose parameter sharing compresses their capacity and weakens their ability to encode morphology, BERT-base-Portuguese is trained on over a billion tokens of high-quality Portuguese data, giving it stronger coverage of lexical, syntactic, and subword patterns. Since Cape Verdean Kriolu is closely related to Portuguese, high-quality Portuguese representations transfer directly to better Kriolu generation, especially in a low-resource setting where encoder quality dominates overall performance. Overall, all models were able to exceed the baseline metrics, implying that training with both a Portuguese latent space and English LM prior assisted the translation model in learning Kriolu.

F. Conclusion

This project set out to test a simple but meaningful idea. If Kriolu is historically and structurally close to Portuguese, then a strong Portuguese encoder should provide a better latent space for representing Kriolu text than a randomly initialized encoder. Further, if English fluency is difficult to learn from limited parallel data, then regularizing the decoder with an English LM prior should help stabilize generation and improve translation quality. Our results largely support both intuitions.

Across nine encoder-decoder pairings, we observe that most models with Brazilian Portuguese latent spaces achieve better validation loss and outperform the baseline on BLEU and chrF. In particular, the BERT-base-Portuguese encoder consistently delivers the highest BLEU improvements across its pairings and competitive chrF scores, suggesting that its larger capacity and richer Portuguese pretraining provide representations that transfer more directly to Kriolu.

This work highlights a path forward for machine translation in creole languages, which are routinely marginalized in commercial and academic NLP systems despite their significant speaker populations and cultural importance. By explicitly leveraging language proximity here, using the deep lexical and structural inheritance from Portuguese, we demonstrate that creole MT does not need to begin from a blank slate, nor rely on unrealistic amounts of supervised data. Instead, by anchoring the encoder in a related high-resource language and stabilizing the decoder with a strong English LM prior, our approach shows that it is possible to make meaningful progress in translation quality even when parallel corpora are sparse, noisy, or heterogeneous. Although large data is still required to make high quality translation models, this opens the door to potential methods for low resource languages. This is especially valuable for creoles, which often have a variety of orthographies and dialectal variation. A model grounded in the linguistic substrate of the creole can better generalize across this variation than systems trained purely from surface statistics.

More broadly, this work suggests a paradigm for developing NLP tools for creole and other low-resource languages that acknowledges their historical and linguistic relationships. Many creoles share similar properties, such as lexical inheritance from a dominant lexifier language, which make them both uniquely challenging and uniquely suited for transfer-based approaches like the one proposed here. By showing that a Portuguese-initialized latent space and an English LM prior can jointly improve robustness and generalization, we provide a template that can be adapted to other creole contexts, such as Haitian Creole, Papiamento, or Patoa. Ultimately, strengthening MT for creole languages contributes to broader linguistic equity: it enables access to digital information, supports language preservation efforts, and ensures that speakers of historically marginalized languages are not excluded from the global technological landscape.

G. Future Work

Future work should include controlled ablations (Portuguese encoder without LM prior, LM prior with English-only encoder, and no prior at all), along with expansion to

additional Kriolu varieties or related creoles to test generality. Since the main obstacle in our results was still due to lack of data, we could also explore data augmentation to artificially increase the size of the low-translation dataset, such as back-translation or synthetic Kriolu generation guided by Portuguese-Kriolu lexical overlap. Another promising direction to look at is knowledge distillation, where a teacher model (i.e. a multilingual model) supervises a smaller student model trained on Kriolu to improve generalization and fluency without requiring large-labeled datasets.

A promising direction for future work involves collecting more granular, locally sourced data across the different islands where Kriolu is spoken, in order to capture regional variation in vocabulary, grammar, and orthographic conventions. Such a dataset could be curated through field recordings, transcriptions, and community contributions, carefully annotated to reflect the speaker's island of origin. By mapping these variations, we can account for the variety that highlights both shared structures and unique local features, which are often underrepresented in existing datasets. This would allow us to evaluate whether the current model can generalize across dialectal differences or if it is biased toward the dominant or most standardized forms present in the training data. Additionally, analyzing model performance on this geographically annotated data could reveal specific patterns of errors or mistranslations linked to certain dialects, guiding future improvements in model architecture or training strategies, such as incorporating dialect-aware embeddings or multi-dialect pretraining. Finally, we can also explore various tokenization strategies other than BPE and WordPiece that's more tailored towards Kriolu orthographic variability.

References

[1] Marlyse Baptista. *The Syntax of Cape Verdean Creole: The Sotavento Varieties*, volume 54 of *Linguistik Aktuell / Linguistics Today*. John Benjamins, Amsterdam and Philadelphia, 2002. 1, 2

[2] Christos Baziotis, Barry Haddow, and Alexandra Birch. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Online, 2020. Association for Computational Linguistics. 1, 2, 3, 5

[3] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020. 2

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2

[5] NLLB Team et al. No language left behind: Scaling human-centered machine translation, 2022. 3

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 2, 5

[7] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. 5

[8] Nathaniel R. Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Bizon Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome A. Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Dean Stutzman, Bismarck Bamfo Odoom, Sanjeev Khudanpur, Stephen D. Richardson, and Kenton 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)*, pages 3083–3110, Mexico City, Mexico, 2024. Association for Computational Linguistics. 1, 2, 3

[9] João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. Advancing neural encoding of portuguese with transformer Albertina PT-*. In *Progress in Artificial Intelligence: 21st EPIA Conference on Artificial Intelligence*, pages 441–453. Springer, 2023. 2, 3

[10] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. 2

[11] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: Pretrained BERT models for brazilian portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417. Springer, 2020. 2

[12] Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. Cold fusion: Training seq2seq models together with language models, 2017. 2

[13] Özgür Uğur, Musa Yılmaz, Esra Şavirdi, Özay Ezerceli, Mahmut El Huseyni, Selva Taş, and Reyhan Bayraktar. Guided decoding and its critical role in retrieval-augmented generation. In *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, page 1–4. IEEE, June 2025. 2