A Comparison of Methods to Bias Translation Toward Portuguese Variants

Catarina Costa and Sebastian Padó
IMS, University of Stuttgart, Germany
catcosta98@gmail.com, pado@ims.uni-stuttgart.de

Abstract

Portuguese serves as the official language of multiple countries across four continents. It is classified into two primary variants (European Portuguese and Brazilian Portuguese), but there is limited research on and resources for European Portuguese compared to the Brazilian variant. In this paper, we consider the task of Machine Translation (MT) into Portuguese. Given the resource imbalance, standard MT systems produce translations that are typically closer to the Brazilian standard. We compare four methods available to bias the translation toward the minority European Portuguese variant that target different places in the MT lifecycle: (1) reranking n-best MT outputs according to a variant classifier; (2) biasing hypothesis generation at inference time toward the target variant; (3) fine-tuning for the target variants; (4) moving completely to an LLM-based approach. We find that all methods can bias translation outputs to an extent. The LLM-based approach yields numerically the highest results, but the impact of memorisation remains unclear.

1 Introduction

The field of Natural Language Processing was biased toward the English language for a long time, given the predominance of both competencies and language resources for that language (Qin et al., 2025). This situation was problematic not just at the level of developing generalisable models but also regarding fairness and participation. In the last ten years, initiatives to promote globalisation and enhance communication have gained traction, with Machine Translation (MT) playing a pivotal role in these efforts (Mercan et al., 2024).

Nevertheless, languages still have very different statuses. In MT, the amount of parallel data available between a language and English largely determines its status as a high- or low-resource language (Ranathunga et al., 2023). Languages with abundant parallel data with English receive greater research attention, while languages with little or no parallel data with English remain less studied (Ikae and Kurpicz-Briki, 2024). In many cases, this gives rise to subpar translation outputs, which may culminate in unequal representations and problems such as misgendering, among others (Savoldi et al., 2025). Furthermore, inequality does not only exist between languages but also between variants of a language, with better resources and models available for more prestigious, or simply for more widely spoken, variants (Joshi et al., 2025). This is a particular issue for languages that are spoken globally.

In this paper, we consider the case of Portuguese and its two main variants, European Portuguese and Brazilian Portuguese (EP and BP). Portuguese falls into the category of a high-resource language; however, since Brazil's population is approximately 20 times larger than that of Portugal, EP is more appropriately considered a low-resource variant, with fewer linguistic resources and less available research compared to its Brazilian counterpart (Medeiros et al., 2023). For example, standard models for MT into Portuguese have been trained predominantly on BP and consequently tend to create output that conforms much more closely to the Brazilian standard.

The goal of our study is to investigate strategies that enable us to bias translation in the direction of the minority variant (here, EP). We explore three different methods that target different aspects of the life cycle of pre-trained MT models and apply them to a state-of-the-art model (OPUS MT en-pt). In addition, we consider LLM prompting as an alternative. We observe significant improvements of the MT model in generating EP, in particular for fine-tuning. These are exceeded by the results we obtain with prompting the LLM; however, it is unclear to what extent the LLM has memorised properties of the evaluation data.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 78–92

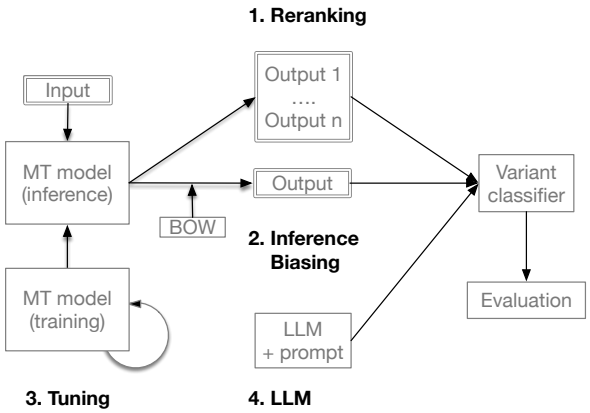

Figure 1: Our setup for evaluating different approaches to translating into specific target language varieties (1-3: by biasing MT models).

2 Methods

In our study, we will use a set of approaches aimed at exploring how variant-specific control can be achieved with minimal resources through adaptations of an existing MT pipeline. Figure 1 visualises the different points at which an MT model can be modified to bias it toward a particular variant of the target language. We consider three different approaches, moving from the end of the pipeline to earlier phases. Approach (1) does not touch the MT system but targets its output, attempting to recover outputs for different variants from an n-best list. Approach (2) is applied during inference time and aims at biasing the highest-ranked hypotheses toward a target variant. Approach (3) targets the MT model itself and fine-tunes it for the target variant. Finally, Approach (4) moves away from a traditional MT model and prompts an LLM to generate a translation in a specific target variant.

2.1 Reranking

This approach is based on ranking different system outputs. It follows in particular the work of Troiano et al. (2020), where this method was used to rerank emotion-based translation candidates. After observing that emotions are partially lost during translation, the authors explored reranking of several outputs, informed by an emotion classifier, to reverse this tendency, obtaining a model for emotion style transfer. The core idea in Troiano et al. (2020), of having a classifier informing us about the presence or absence of particular characteristics in the MT outputs, can be adapted to prioritise translations that best align with linguistic characteristics indicative of Brazilian or European Portuguese. This is possible by leveraging the beam search decoding capabilities of the MT model to generate a diverse set of candidate translations for each input sentence, and then selecting the highest-scoring candidate for each variant using a variant classifier (see Section 3.3).

2.2 Inference Biasing

By treating each variant as a separate domain, it is possible to adapt a pre-trained NMT model to produce outputs that conform to the linguistic norms of the desired variant. This approach is particularly useful when retraining or fine-tuning the full model is impractical due to limited data or computational constraints.

Building on Saunders and Byrne (2020) and drawing inspiration from Dathathri et al. (2020), one effective strategy involves guiding model outputs at inference time using lightweight attribute models. In Dathathri et al. (2020), the authors used a pre-trained transformer language model as-is and augmented the decoding process with a small auxiliary model, in the form of a bag-of-words (BoW) or a shallow classifier, that steered the output in the desired direction. Kadiķis (2023) already explored this method with OPUS-MT models, using them without any weight changes, and adapting them in the decoding step through small domain discriminators that nudged the model's output closer to the desired domains (in their case, formality and gender). This happens by encoding the source text and then iteratively adjusting the decoder's hidden states during generation. For each token, the model first performs a forward pass to predict the next token distribution. That output is then evaluated by a domain attribute model (BoW), which scores how well it matches the target domain. Using this score, a backward pass computes gradients with respect to the decoder's hidden states (not the model parameters), and the hidden states are updated accordingly. These adjustment steps may be repeated several times. A final forward pass with the perturbed hidden states produces a modified probability distribution from which the next token is sampled. Additional components—warm-up steps and a negative bag-of-words—help adapt the method to MT and to contrasting domains.

2.3 Fine-Tuning

Fine-tuning offers another, fairly established, avenue for improving variant-specific translation (Ouyang et al., 2022; Zhou et al., 2023). In

MT, fine-tuning involves continuing the training of a general-purpose translation model on a smaller, domain-specific parallel corpus, allowing the model to better capture the linguistic characteristics of the target variant. This approach leverages the knowledge acquired during pre-training while specialising the model for the desired output. In the context of European and Brazilian Portuguese, fine-tuning will allow the MT model to learn variant-specific lexical choices, morphological patterns, and syntactic structures, helping it generate outputs consistent with the intended variant.

Arguably one of the most popular and effective parameter efficient methods for optimising the fine-tuning of models pre-trained on a large amount of data is LoRA (Hu et al., 2022), which makes fine-tuning computationally efficient by freezing pre-trained weights and injecting trainable rank decomposition matrices into the model, significantly reducing trainable parameters, aiming to balance the trade-off between the model's performance and the computational resources required for its adaptation. The new matrices are trained on new data while keeping the overall number of changes low. We adopt this method due to its practicality in our resource-constrained setting.

2.4 LLM-based Translation with Prompting

The success of generative models comes with the prospect of doing away with bespoke strategies such as those presented above. In the context of generative LLMs, machine translation targeted at specific low-resource languages, variants, and dialects can be achieved through prompting to produce outputs that adhere to the desired linguistic norms. This approach allows for flexible adaptations without modifying model parameters or retraining. We explore whether prompting can steer the model to generate translations consistent with European or Brazilian Portuguese.

3 Data

To carry out experiments with the approaches sketched in Section 2, we need two types of datasets: (a) development data for the target variant (European Portuguese); and (b) unseen test data for both variants (EP and BP) for unbiased evaluation. We also require a classifier that distinguishes Brazilian from European Portuguese.

3.1 Tatoeba

The Tatoeba dataset is a special compilation of data released under the label of the Tatoeba Translation Challenge.[1] It is a collection of example sentences with translations geared toward language learners trained on OPUS data (Tiedemann et al., 2023).

Several versions of this dataset already exist, and for Portuguese, translations are mixed and provided in both Brazilian and European Portuguese. To obtain a usable development corpus, we identify a subset of Tatoeba that is not yet part of the 2021 release of the OPUS MT model we use in our experiments (cf. Section 4.1) and classify them as either EP or BP. Concretely, we start from the 2023 Tatoeba dataset for Portuguese (227,657 sentences) and remove all sentences that were present in Portuguese datasets earlier than 2021. This yields a total of 172,121 English-Portuguese sentence pairs that are unseen for our MT model. These sentences are then pre-classified into BP or EP with the PeroVaz_PT-BR classifier described in Section 3.3. We create a dataset with 50K sentences for each variant. For EP, each sentence is manually verified and adapted where necessary.[2]

3.2 FRMT

FRMT (Riley et al., 2023) is a dataset and evaluation benchmark for Few-shot Region-aware Machine Translation, consisting of professional translations from English into two regional variants each of Portuguese (BP and EP) and Mandarin Chinese.

The dataset comes divided into three buckets (lexical, entity, and random), consisting of sourced Wikipedia English sentences and their human translations. The lexical bucket is a subset where the English Wikipedia pages were selected according to a manual list of terms and phrases that tend to be translated differently for the targeted regions. In the entity bucket, the English Wikipedia pages are about entities that were deemed to have a strong connection to a particular targeted region. The random bucket is a subset of random English Wikipedia pages appearing in the "good articles" or "featured articles" collections of Wikipedia.

The datasets are split at the document-level into "exemplar", "dev", and "test" splits. Further examination of the dataset before the beginning of the experiments revealed that some sentences were

[1] https://github.com/Helsinki-NLP/Tatoeba-Challenge

[2] The datasets are available at https://github.com/catvcosta/epbp_datasets.

not properly split. In addition, for the goals of this research, we combine the exemplar and dev buckets in order to use them for fine-tuning. We end up with a total of 2847 sentences in the extended dev set (dev + exemplar), alone and combined with different splits of the Tatoeba dataset (explained above) and 2597 sentences in the test set. The test set sentences are translated across all our methods and used for calculating our evaluation scores.[3]

3.3 BP/EP Classifier

The differences between EP and BP span multiple linguistic dimensions, including phonetics, syntax, lexicon, and orthography[4] (Barreiro et al., 1996). For the purposes of this paper, we used an automatic classification method consisting of a pre-trained BERT model trained on a collection of 500,000 sentences evenly split between EP and BP to classify our MT outputs as Brazilian or European Portuguese: the PeroVaz_PT-BR classifier.[5] According to the model developers, this classifier achieves a binary classification accuracy of 94.6% on an evaluation set (and its softmax architecture enables it to provide confidence scores).

However, there is no guarantee that the evaluation results obtained by the authors generalise directly to the data that we work with. Therefore, we carry out a small-scale study to assess the classifier's performance on our experimental data. To do so, we randomly select 100 sentences to evaluate the classifier predictions (EP vs. BP) against the intuition of one of the authors, a native speaker of European Portuguese.[6] Out of the 100 sampled sentences, manual and automatic classification disagree on 16. However, these cases correspond mostly to short or structurally simple sentences that can plausibly be interpreted as either EP or BP. This puts the classifier's accuracy on our data at least 84%. The remaining labels (most frequently BP) are not necessarily incorrect; however, this result suggests a tendency for the model to favour Brazilian Portuguese when the sentence does not contain clear variant-specific markers.

[3]The edited FRMT datasets are available at https://github.com/catvcosta/epbp_datasets.

[4]See Appendix A for details.

[5]https://huggingface.co/bastao/PeroVaz_PT-BR_Classifier

[6]The sentences and evaluation can be consulted at https://github.com/catvcosta/epbp_datasets.

4 Experimental Setup

We now describe the baseline MT model we use as well as the concrete implementation of each of the methods described in Section 2.

4.1 Baseline Machine Translation Model

The model in use throughout the experiments is the pre-trained OPUS-MT[7] en-pt model[8] (Tiedemann and Thottingal, 2020), available through the Hugging Face Transformers library. This model is based on state-of-the-art transformer-based neural machine translation. It uses the Marian-NMT framework[9] (Junczys-Dowmunt et al., 2018) with an architecture consisting of six self-attentive layers in the encoder and decoder network and eight attention heads in each layer, and with hyperparameters following the general recommendations of Vaswani et al. (2017). Tokenisation of sentences before translation is performed via the SentencePiece model (Kudo and Richardson, 2018).

The adaptation of a multilingual translation model like OPUS-MT to generate outputs specific to European or Brazilian Portuguese requires introducing mechanisms to distinguish between the two variants. A popular technique when dealing with multiple variants in the target language involves introducing a variant-specific token at the beginning of each source sentence during training. In Tiedemann (2020), the author released new baseline models where said variant-specific tokens were used (Johnson et al., 2017); one of these models was trained to separate between EP (token *por*) and BP (token *pob*). We will adapt this model throughout this project and use it as a baseline to compare with the different adaptations. We used the model as-is, without further alterations.

Our baseline consists of translating the English FRMT test sentences with the pre-trained OPUS-MT en-pt model, without any modifications or adaptations, and using the language tokens for EP (*por*) and BP (*pob*). Since the baseline uses the non-altered original model, it enables us to measure to what extent variant-specific features are already implicitly captured, and it give us an overview of how much the BP data is affecting the EP output, even in the presence of the language token.

[7]https://github.com/Helsinki-NLP/Opus-MT

[8]https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-en-pt

[9]https://marian-nmt.github.io

4.2 Reranking

The reranking experiments extend the baseline by outputting several candidate translations. For each input sentence, the model's 50-best list of translations was generated. These lists were then reranked based on the likelihood that the candidates adhered to features characteristic of BP or EP, using the confidence scores provided by the PeroVaz_PT-BR classifier. A key challenge with this approach was that not all sentences yielded sufficient candidate diversity in the output list, meaning that, for some sentences, the ranking step could not effectively modify the outcome.

4.3 Inference Biasing

We adopt a strategy as defined in Section 2.2 by creating bag-of-words (BoW) attribute models for each Portuguese variant. The full BoWs can be consulted in Appendix B. During decoding, the system biases generation toward words in the positive bag for the target variant and away from the corresponding negative bag (containing words typical of the other variant). We hypothesise that during decoding, the BoW classifiers will increase the likelihood that the generated translations adhere to the target variant.

4.4 Fine Tuning

We explore fine-tuning with LoRA across datasets of varying sizes to assess whether small amounts of data are sufficient to improve translation performance in low-resource settings. The training datasets consisted of parallel source-target sentences, which were split into training and validation sets with a 5% evaluation ratio. Each source sentence was prepended with the language-specific token (*pob* or *por*, depending on the variant) to guide the model during translation. We additionally relied on quantised training (using 8-bit precision) and gradient-checkpointing to reduce GPU memory consumption without substantially degrading performance. Sentences were tokenised with the model's built-in SentencePiece tokeniser.

LoRA was applied to the model's attention and feed-forward layers, and the hyperparameters were set as follows: LoRA rank set to 16, LoRA alpha set to 32, and dropout set to 0.1, while keeping the base model mostly frozen. Training was carried out for 5 epochs, with a batch size of 4 and gradient accumulation steps of 4, effectively simulating a batch size of 16. The optimiser used was Adam with a learning rate of 2e-4. Models were evaluated at the end of each epoch using the BLEU score on the validation set, and the checkpoint achieving the highest BLEU was automatically saved. Decoding during evaluation relied on the model's default beam search strategy.

The OPUS-MT model was fine-tuned on three subsets with 2.8K (data from FRMT), 10K, and 50K sentence pairs (data from FRMT and Tatoeba). The 2.8K subset is composed of the exemplars + dev sentence pairs from the FRMT dataset. While simulating the most extreme data scarcity scenario for the variants, this dataset is also the one with sentence pairs closer to the domain of the test dataset, so similar results to the higher dataset sizes were expected. The 10K subset (composed of the 2.8K dataset plus 7.2K sentences from the Tatoeba dataset) provides a mid-sized training set, and the 50K subset (composed of the 2.8K dataset plus 47.2K sentences from the Tatoeba dataset) approaches a high-resource scenario within the constraints of this project. This staged setup allows us to explore how sensitive the model is to the amount of available variant-specific data and whether improvements scale linearly with dataset size.

4.5 LLM-based Translation with Prompting

To explore translation with prompting, we used the prompt "Translate the following sentences into Brazilian/European Portuguese. Return the sentences in the same layout." We used Gemini 2.5 Pro with vanilla settings.

4.6 Evaluation

We use the FRMT test subset consisting of 2,597 sentences distributed across the three buckets for evaluation. We employ multiple metrics for evaluation to obtain a comprehensive assessment.

BLEU. First, BLEU (Papineni et al., 2002), which measures the similarity between a reference translation R and a candidate translation C as:

$$\text{BLEU}(R, C) = \min(1, \exp(1 - \frac{|R|}{|C|})) \prod_{i=1}^{4} \text{prec}_i{}^{\frac{1}{4}}$$

where prec_i is the precision of C with respect to R regarding i-grams. BLEU computes a score between 0 and 1 (typically reported as percentages) which reflects how closely the candidate matches the reference in terms of lexical choices.

chrF. Second, chrF (Popović, 2015) calculates the similarity between C and R using character n-grams rather than word n-grams:

$$\text{chrF}(R, C) = \frac{2 \cdot \text{chrP} \cdot \text{chrR}}{\text{chrP} + \text{chrR}}$$

where chrP stands for the percentage of character n-grams in the hypothesis, which have a counterpart in the reference, and $chrR$ is defined analogously. Like standard F-Score and BLEU, the values of chrF also range between 0 and 1.

BLEURT. Finally, we use BLEURT (Sellam et al., 2020), an evaluation metric for Natural Language Generation tasks that produces a score for a (C, R) pair reflecting both fluency and semantic similarity. The score is computed on the basis of BERT representations for C and R, mapped onto scalar with a function learned from gold-annotated MT datasets.

Statistical analysis. To evaluate whether system performance differences are statistically significant, we conducted two-tailed paired t-tests between each system and the baseline, while accounting for our multiple comparisons problem. Due to the referred problems of untranslated sentences in our reranking experiment, we computed the statistical analysis only for our fine-tuning systems (2.8K, 10K, and 50K), for our inference biasing system, and for our LLM-based approach.

When evaluating several systems against the same baseline, the likelihood of observing Type I errors increases if a standard significance threshold is applied independently to each test (Her and Kruschwitz, 2024). To mitigate this risk, we apply the Bonferroni correction, a simple yet conservative approach that adjusts the significance threshold by dividing α by the number of comparisons m. In our case, with $m = 5$ system–baseline comparisons, the corrected threshold becomes $\alpha = 0.05/5 = 0.01$. We therefore report results of two-tailed paired t-tests comparing each system against the baseline, and mark those differences in our tables as significant when $p < 0.05$ (*) and highly significant when $p < 0.01$ (**).

5 Results

We provide evaluation results for the baseline MT model on the different "buckets" of the FRMT test set, and then compare these results against the outcomes for the different biasing strategies.

Test Bucket	BLEU	chrF	BLEURT
Entity (EP)	46.05	69.43	76.74
Entity (BP)	57.29	76.54	79.85
Lexical (EP)	34.26	61.06	70.14
Lexical (BP)	52.36	73.30	75.99
Random (EP)	40.68	66.37	72.75
Random (BP)	55.13	74.86	75.98

Table 1: Baseline results for the three test buckets.

5.1 Baseline results

The baseline scores are provided in Table 1. Unsurprisingly, they show that all BP scores are substantially better than their EP counterpart. This confirms that the baseline OPUS-MT model is indeed biased toward Brazilian Portuguese, arguably due to a higher representation of BP in the training data. This validates the need for research toward improving the results for European Portuguese as a target variant.

For the sake of space, we report only chrF scores for the remaining methods, as this was consistently the most statistically significant metric. Full results, completed with statistical analysis checks, can be consulted in Appendix C.

5.2 Reranking

Recall that the reranking approach relies on the classifier (PeroVaz_PT-BR) to select the most likely translation variant. For some sentences, the classifier did not identify any candidate for the desired variant in the 50-best beam of the MT model. This was the case for roughly 35% of the EP inputs and 10% of the BP inputs. As a consequence, part of the test set remains untranslated, and the resulting scores cannot be directly compared between EP and BP, nor to the baseline scores from Table 1, which are computed on the full test set. For this reason, Table 2 presents results for each variant and test bucket that are computed on the *intersection* of sentences that received translations for both variants. We also recompute the baseline results on the intersection for direct comparability.

Overall, most reranking outputs are worse than the baseline, and for BP, there are no exceptions. For EP, we observe mild improvements in the lexical bucket, indicating that some EP lexical choices might indeed be contained in the pre-trained model but do not make it into the top-ranked translation. However, the gain is marginal and comes at a cost

Test Bucket	# Intersec.	chrF (baseline)	chrF (rerank)
Entity (EP)	390/973	68.56	66.34
Entity (BP)	(40%)	74.70	71.68
Lexical (EP)	491/873	60.66	**61.76**
Lexical (BP)	(56%)	72.87	70.68
Random (EP)	378/751	68.27	66.79
Random (BP)	(50%)	76.02	72.78

Table 2: Results for reranking (computed on the intersection between outputs of EP and BP for each bucket, cf. text). Boldface shows improvement over baseline.

Test Bucket	chrF (baseline)	chrF (inf. biasing)
Entity (EP)	69.43	**69.52**
Entity (BP)	76.54	76.20
Lexical (EP)	61.06	**61.57**
Lexical (BP)	73.30	72.84
Random (EP)	66.37	66.14
Random (BP)	74.86	74.16

Table 3: Results for inference biasing. Boldface shows improvement for inference biasing over baseline.

Test Bucket	chrF (BL)	chrF (2.8K)	chrF (10K)	chrF (50K)
Entity (EP)	69.43	***72.89***	**72.79**	**72.07**
Entity (BP)	76.54	**78.21**	***78.27***	**78.00**
Lexical (EP)	61.06	***66.96***	**66.88**	**66.36**
Lexical (BP)	73.30	***75.13***	**75.00**	**74.55**
Random (EP)	66.37	***69.33***	**69.29**	**68.76**
Random (BP)	74.86	**75.46**	***75.52***	**75.17**

Table 4: Results for fine-tuning. BL=baseline, and we indicate the number of sentence pairs used for fine-tuning. Boldface shows improvement over baseline; best results for each test bucket are italicised.

in untranslated sentences. Thus, we conclude that while reranking is an interesting choice in that it does not require any modifications to the MT system, its effectiveness is very limited.

5.3 Inference Biasing

Table 3 shows the results for inference biasing using the BoWs defined in Section 4.3. Since inference biasing does not affect coverage, evaluation results are again on the complete test buckets. For convenience, we repeat the baseline results from Table 1.

We find that domain adaptation at inference time can help push translations closer to the intended variant. Improvements are visible for EP in both the entity and lexical buckets, even if they are mild. This highlights the potential of this method as a lightweight and effective intervention, particularly for the underrepresented variant. When the goal is to increase lexical fidelity to specific language variants in a resource-efficient manner, inference biasing appears to be an interesting alternative. However, the definition of suitable BoWs for the variants introduces a hyperparameter that we do not experiment with.

5.4 Fine-tuning

Table 4 highlights different fine-tuning strategies (2.8K, 10K, and 50K sentence pairs) and compares them against the baseline results.

In contrast to the other methods, all fine-tuned systems outperform the baselines on all EP and BP datasets, and do so by a substantial margin: for EP, performance improves by 3-6 points chrF, with the highest gain for the Lexical bucket. For BP, we still see improvements by 1-2 points chrF. Interestingly, the best results are obtained for either the 2.8K condition (EP) or the 10K (most BP buckets). This indicates that going beyond a small amount of data for fine-tuning runs the risk of interference between the original and the fine-tuned model (Shaham et al., 2023). Framed positively, a small amount of target variant data appears to be enough to obtain a substantial positive effect.

5.5 LLM-based Translation with Prompting

Finally, we present the results for translating with an LLM, where we specify the target variant in the prompt. Table 5 indicates that the Gemini model consistently outperforms the baseline, on average by about 2.5 points chrF for European Portuguese and about 1.5 points chrF for Brazilian Portuguese. Taken at face value, this indicates that appropriately prompted LLMs can effectively enhance translation quality, and that the improvement is greater for the minority variant, thus reducing the performance gap, even if not closing it: even for Gemini, there is still a 5-6 point chrF gap between EP and BP.

However, there are two caveats. First, it is well established that LLMs are extremely good at memorising data that they have seen during training, and

Test Bucket	chrF (baseline)	chrF (Gemini 2.5)
Entity (EP)	69.43	**74.49**
Entity (BP)	76.54	**79.06**
Lexical (EP)	61.06	**68.83**
Lexical (BP)	73.30	**75.98**
Random (EP)	66.37	**72.44**
Random (BP)	74.86	**78.71**

Table 5: Results for LLM-based translation. Boldface shows improvements for the LLM over the baseline.

over-perform on such data (Michel et al., 2025). Given that FRMT is a publicly available dataset from 2023, there is a fair chance that the test data are not completely unseen for Gemini, and the high performance would decrease at least to some extent on completely novel data. Second, it has to be taken into account that Gemini is several orders of magnitude more complex than the OPUS MT model and correspondingly vastly more costly in terms of training and inference. From this perspective, a higher performance is all but expected.

6 Related Work

MT for low-resource languages and variants. A popular approach to low-resource MT is transfer learning, where the key idea is to first train a high-resource language pair (the parent model), then transfer some of the learned parameters to the low-resource pair (the child model) to initialise and constrain training (Zoph et al., 2016). Another option is to use monolingual corpora, which are easier to obtain than parallel corpora, to derive unsupervised and semi-supervised MT techniques (Ranathunga et al., 2023).

Strategies such as data augmentation also play a key role in low-resource and variant translation. Back-translation leverages monolingual data in the target variant by automatically generating synthetic parallel data, which can be added to the training set (Sennrich et al., 2016). He et al. (2020) revisit self-training strategies and propose methods to augment the original labelled dataset with unlabelled data. Other works target low-frequency words by generating new sentence pairs containing rare words in synthetically created contexts (Fadaee et al., 2017).

While data augmentation strategies can improve translation quality via regularisation and domain adaptation, parallel training signals are limited (Dabre et al., 2020). An approach aimed at addressing this issue is Multilingual NMT, where models are trained simultaneously on multiple language pairs, allowing for parameter sharing across related languages and enabling cross-lingual transfer (Aharoni et al., 2019; Arivazhagan et al., 2019).

A further refinement has been the introduction of language tags or variant control tokens, which explicitly indicate the desired target language or dialect during training and inference, an approach pioneered by Johnson et al. (2017). Conditioning a model on such tags can steer it toward the correct variant even if the training data is unbalanced.

Variants of Portuguese. Zampieri and Gebre (2012) investigate methods for automatic language identification of written texts in European and Brazilian Portuguese. Preda et al. (2024) review methodologies to distinguish between EP and BP, concluding that transformer-based models seem to be robust for out-of-domain data, but the best performance was obtained using simple representation techniques and a traditional classifier. Aepli et al. (2023) provide an overview of recent submissions for the shared tasks organized under the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), including work on Portuguese.

Early approaches to address the challenges pertaining to translation of BP and EP include the work of Marujo et al. (2011), who presented a rule-based system designed to transform Brazilian Portuguese texts into European Portuguese, highlighting the potential of leveraging linguistic similarities between these varieties to improve machine translation performance. Several years later, Costa-jussà et al. (2018) presented the first neural-based machine translation system (RNN with attention mechanism) trained to translate between EP and BP, and compared the performance of this method to a phrase-based SMT system.

Cortes et al. (2024) recently highlighted the challenge of localisation, considering MT approaches for EP and BP. In this study, generative GPT-4 delivered superior performance, opening doors for novel generative LLM techniques with prompts. In a similar line, Sanches (2024) proposed the development of a BP to EP translation system by fine-tuning existing Large Language Models, noting that ChatGPT 3.5 yielded the best results concerning out-of-domain generalisation.

7 Conclusions

This paper has started out from the observation that the performance of state-of-the-art dedicated neural machine translation models for translating into Portuguese tends to conform more to the Brazilian standard (BP) than the European one (EP). This is unsurprising, given the predominance of BP data compared to EP data. We have asked the question of how the performance of such a model for the minority variant, EP, can be improved.

We have assessed three methods for this purpose: reranking of the n-best list output with a variant classifier, which does not require any adaptation; inference biasing based on variant-specific bags-of-words, which requires changes to the decoding algorithm but keeps the model constant; and fine-tuning with variant-specific data. Unfortunately, we found a strong correlation between the degree of intervention and the effectiveness of the methods: only fine-tuning leads to robust and substantial improvements over the baseline MT performance, once again confirming the old adage that "there is no data like more data". On the upside, however, we found that 2.8K sentence pairs were sufficient to obtain most of the benefit of fine-tuning, indicating that (at least in our LoRA-based setup) a small amount of variant-specific data can go a long way.

We also addressed the task with a prompted LLM and obtained results that improved over the fine-tuned MT model, even without further prompt tuning. This is in line with the observation that current multilingual LLMs are extremely good at generating specific types of language, cf. the success of the so-called "persona prompting" (Kambhatla et al., 2025). While this development is very promising for the many underresourced languages and variants, our study could not exclude the possibility that memorisation of the test dataset is behind the good results of the LLM.

Going beyond our concrete experiments, our study illustrates the range of methods available to bias machine translation toward specific language variants or dialects. Future work could extend the analysis to other variants of Portuguese, and beyond, to languages that underwent similar processes, such as European and Latin American Spanish. In this vein, a deeper linguistic analysis of variant-specific phenomena could provide better insights into how MT systems capture variant-specific phenomena. Moving forward, research should focus on ensuring that these technologies not only produce accurate translations, but also respect cultural nuances, preserve regional and minority language features, and reinforce the identities embedded in language (Lin, 2025). By embracing both technical innovation and cultural sensitivity, future work can help build language technologies that respect linguistic richness.

8 Limitations

Our study has a number of limitations. First, our small-scale evaluation of the automatic Portuguese variant classifier indicates that the classifier, although mostly correct, shows a bias toward Brazilian Portuguese when sentences do not contain variant-specific markers (cf. Section 3.3). Having native speakers rank the outputs of the original models could resolve ambiguities where sentences could belong to either of the variants, as well as address issues such as untranslated sentences.

Also, all of our methods can be further refined. Our reranking strategy (cf. Section 4.2) could arguably benefit from search procedures that focus on promoting dissimilar translation alternatives, such as diverse beam search (Vijayakumar et al., 2016). In turn, our inference biasing approach is limited by its exclusive focus on lexical cues (cf. Section 4.3). As shown in Annex A, differences across EP and BP span multiple linguistic and contextual dimensions, and incorporating those could improve the effectiveness of this approach. The practical challenge is to compute these features from incomplete translation hypotheses. Finally, our fine-tuning approach uses a union of FRMT and Tatoeba sentences to optimise the MT model, but was evaluated on FRMT data only. This offers a possible explanation for the lack of further improvement we see for larger fine-tuning datasets: it may be due to the differences between the corpora, or due to Tatoeba containing many short and potentially noisy sentences, which hinder rather than support model biasing. We leave these observations to future work.

References

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial Evaluation Campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively Multilingual Neural Machine Translation. 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 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

N. Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The Missing Ingredient in Zero-Shot Neural Machine Translation. *ArXiv*, abs/1903.07091.

Anabela Barreiro, Luzia Wittmann, and Maria Pereira. 1996. Lexical differences between European and Brazilian Portuguese. *The INESC Journal of Research & Development*, 5.2:75–101.

Eduardo G. Cortes, Ana Luiza Vianna, Mikaela Martins, Sandro Rigo, and Rafael Kunst. 2024. LLMs and Translation: different approaches to localization between Brazilian Portuguese and European Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 45–55, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Marta R. Costa-jussà, Marcos Zampieri, and Santanu Pal. 2018. A Neural Approach to Language Variety Translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 275–282, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Survey of Multilingual Neural Machine Translation. *ACM Comput. Surv.*, 53(5).

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *Proceedings of the International Conference on Learning Representations*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data Augmentation for Low-Resource Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Carlos Gonçalves, Graça Rio-Torto, and João Tavares. 2021. Morphological studies on portuguese: a representative sample in Brazil and Portugal. *Revista Diadorim*, 23:153–158.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting Self-Training for Neural Sequence Generation. In *Proceedings of the International Conference on Learning Representations*.

Wan-hua Her and Udo Kruschwitz. 2024. Investigating Neural Machine Translation for Low-Resource Languages: Using Bavarian as a Case Study. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 155–167, Torino, Italia. ELRA and ICCL.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations*.

Catherine Ikae and Mascha Kurpicz-Briki. 2024. Current State-of-the-Art of Bias Detection and Mitigation in Machine Translation for African and European Languages: a Review. *arXiv preprint*, arXiv:2410.21126.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural Language Processing for Dialects of a Language: A Survey. *ACM Comput. Surv.*, 57(6).

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Emīls Kadiķis. 2023. Plug-and-Play Domain Adaptation for Neural Machine Translation. Master's thesis, University of Stuttgart, Institut für Maschinelle Sprachverarbeitung, September. Available at `http://elib.uni-stuttgart.de/handle/11682/14249`.

Gauri Kambhatla, Chantal Shaib, and Venkata S Govindarajan. 2025. Measuring Lexical Diversity of Synthetic Data Generated through Fine-Grained Persona Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21024–21033, Suzhou, China. Association for Computational Linguistics.

Mary A. Kato and Ana Maria Martins. 2016. European Portuguese and Brazilian Portuguese: An Overview on Word Order. In Leo Wetzels, Sergio Menuzzi, and João Costa, editors, *The Handbook of Portuguese Linguistics*, pages 15–40. Wiley-Blackwell, Hoboken, NJ.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Zi-Xiang Lin. 2025. Digital Tongues: Internet Language, Collective Identity, and Implications for Human-Computer Interaction. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 1–6, Suzhou, China. Association for Computational Linguistics.

Luis Marujo, Nuno Grazina, Tiago Luis, Wang Ling, Luisa Coheur, and Isabel Trancoso. 2011. BP2EP - Adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, Leuven, Belgium. European Association for Machine Translation.

Eduardo Medeiros, Leonel Corado, Luís Rato, Paulo Quaresma, and Pedro Salgueiro. 2023. Domain Adaptation Speech-to-Text for Low-Resource European Portuguese Using Deep Learning. *Future Internet*, 15(5).

Hanımnur Mercan, Yaşar Akgün, and Mehmet Cem Odacıoğlu. 2024. The Evolution of Machine Translation: A Review Study. *Uluslararası Dil ve Çeviri Çalışmaları Dergisi*, 4(1):104–116.

Gaspard Michel, Elena V. Epure, Romain Hennequin, and Christophe Cerisara. 2025. Evaluating LLMs for Quotation Attribution in Literary Texts: A Case Study of LLaMa3. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 742–755, Albuquerque, New Mexico. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Proceedings of Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

David Preda, Tomás Osório, and Henrique Lopes Cardoso. 2024. Across the Atlantic: Distinguishing Between European and Brazilian Portuguese Dialects. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 353–363, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural Machine Translation for Low-resource Languages: A Survey. *ACM Comput. Surv.*, 55(11).

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. FRMT: A Benchmark for Few-Shot Region-Aware Machine Translation. *Transactions of the Association for Computational Linguistics*, 11:671–685.

João Filipe Silva Lopes Seguro Sanches. 2024. From Brazilian Portuguese to European Portuguese. Master's thesis, Instituto Superior Técnico da Universidade de Lisboa, June.

Danielle Saunders and Bill Byrne. 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Beatrice Savoldi, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025. A decade of gender bias in machine translation. *Patterns*, 6(6):101257.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2023. Causes and Cures for Interference in Multilingual Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15849–15863, Toronto, Canada. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, 58:1–43.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in Back-Translation: Emotion Preservation in Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. *arXiv preprint arXiv:1610.02424*.

Marcos Zampieri and Benjamin G. Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. In *Proceedings of the Conference on Natural Language Processing 2012*, pages 233–237, Vienna. Österreichische Gesellschaft für Artificial Intelligence (ÖGAI).

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. In *Proceedings of Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

EN	BP	EP
cellphone	celular	telemóvel
breakfast	café da manhã	pequeno-almoço
train	trem	comboio
bus	ônibus	autocarro
juice	suco	sumo
toilet	banheiro	casa de banho
fridge	geladeira	frigorífico
grass	grama	relva
sport	esporte	desporto
team	time	equipa

Table 6: Examples of differences in vocabulary for Brazilian Portuguese and European Portuguese, and the English translation.

A Differences between EP and BP

The differences between EP and BP span multiple linguistic dimensions, including phonetics, syntax, lexicon, and orthography (Barreiro et al., 1996), of which we provide an overview below (note that this does not encompass regional linguistic varieties).

In what concerns lexical differences, we can have different words for the same referent, as shown in Table 6.

There are also words that are available in both variants, but that are not used with the same frequency; for example, both *chávena* and *xícara*, meaning teacup, are used in EP and BP, but the former is preferred in EP and the latter in BP. In addition, each variant has different dialects with unique slang and regional terms, as well as words that have no equivalent in the other variant (such as the names of certain plants, fruits, or animals), which would be important to consider for future research. Another important difference in lexicon refers to institutional contrastive words (Barreiro et al., 1996), since official and administrative systems are not organised in the same way in Brazil and in Portugal (as an example, Brazil is divided into *Estados* [States] and Portugal into *Distritos* [Districts]; the Ministry of Foreign Affairs is named *Ministério dos Negócios Estrangeiros* in Portugal and *Ministério das Relações Exteriores* in Brazil).

Both in Brazil and Portugal, works in morphology are still a minority, compared to studies in syntax and in semantics (Gonçalves et al., 2021). However, it is important to point out two morphological phenomena that distinguish BP and EP: different derivation (prefixes and suffixes)—the word

for PhD being *doutoramento* in EP and *doutorado* in BP—and different inflexion, such as the existence of two different past participles for the verb "to accept": *aceite* in EP and *aceito* in BP (Barreiro et al., 1996).

In what concerns syntax, there are several major differences between the two variants. One of the most well-known is the use of the gerund in Brazilian Portuguese in place of the structure "a + infinitive", more common in European Portuguese.

Regarding verbal reflexivity and prepositional phrases, clitic placement is also contrasting in BP and EP (Kato and Martins, 2016). In constructions involving a verb and a reflexive pronoun (me, te, se, nos), EP places the pronoun after the verb (enclisis), while in BP the reflexive pronoun often comes before the infinitive (proclisis).

There are exceptions, however, where in EP the reflexive pronoun will also follow a proclitic placement, such as in the use of the negative ("They didn't hug us" would become "Elas *não* **nos** abraçaram"), or when the reflexive verb follows an adverb, or a question word (interrogative adverbs and interrogative pronouns).

Different usage of pronouns in the two variants doesn't end with reflexive pronouns. Direct object pronouns (in Portuguese, **o**, **a**, **os** and **as**) are usually preferred in EP, while BP replaces them with personal pronouns (**ele**, **ela**, **eles**, **elas**). Therefore, while saying "I saw her yesterday", EP speakers would prefer "*Eu vi-**a** ontem*", and BP "*Eu vi **ela** ontem*".

In Portuguese, the preposition *em* (meaning "at") can be combined with definite articles **o**, **a**, **os** and **as** to form the contractions **no**, **na**, **nos**, and **nas**, which BP uses in a way that EP doesn't. For example, "I went to his house" would turn into "*Fui **a** casa dele*" in EP and "*Fui **na** casa dele*" in BP.

A final major difference in pronoun usage involves the use of second-person pronouns. In European Portuguese, as in, for example, German, "tu" (German "du") is used with closer relationships, and "você" ("Sie") when speaking to elderly relatives or acquaintances. In Brazilian Portuguese, "você" is the preferred pronoun even for the informal or closer relationships.

Concerning orthography, differences in accentuation and spelling persist despite the 1990 Orthographic Agreement[10], in full implementation since 2009 in Portugal and Brazil. These distinctions are sometimes very subtle (see Table 7), posing challenges for Machine Translation (MT) systems, which must account for such linguistic variability to generate high-quality translations tailored to each variant.

EN	BP	EP
baby	bebê	bebé
demon	demônio	demónio
with us	conosco	connosco
amnesty	anistia	amnistia
fact	fato	facto
academic	acadêmico	académico
economic	econômico	económico

Table 7: Orthographic differences persisting despite the 1990 Orthographic Agreement.

European and Brazilian Portuguese differ across multiple linguistic levels. While some distinctions are subtle, others can significantly affect communication. Acknowledging these variant-specific features is crucial, as they directly impact the analysis of written data and the development of variant-sensitive linguistic or computational methods.

B BoW Features for Inference Biasing

BP_BoW: ["ônibus", "os ônibus", "trem", "você", "legal", "cara", "a gente", "cachorro", "celular", "se tornou", "ensino médio", "prêmio", "cerimônia", "canônico", "grade curricular", "acadêmico", "esporte", "recepção", "irônico", "ganhador", "comitê", "planejar", "controle", "cancêr", "em um", "em uma", "sacada", "quilômetro", "prefeitura", "abacaxi", "mamão", "shorts", "econômico", "equipe", "gol", "caminhão", "bebê", "banheiro", "garoto", "garota", "bachelorado", "gênero", "higiêne", "marrom", "terno", "geladeira", "freezer", "ozônio", "presunto", "boliche", "mouse", "suco", "rodovias", "estresse", "turnê", "registro", "xícara", "açougue", "camiseta", "sorvete"].

EP_BoW: ["autocarro", "os autocarros", "comboio", "tu", "fixe", "rapaz", "telemóvel", "cão", "tornou-se", "ensino secundário", "prémio", "cerimónia", "canónico", "currículo escolar", "académico", "desporto", "receção", "irónico", "vencedor", "comité", "planear", "controlo", "cancro", "num", "numa", "varanda", "quilómetro", "câmara municipal", "ananás", "papaia", "calções",

[10]The specifics of this Orthographic Agreement can be checked at the Portal da Língua Portuguesa.

"económico", "equipa", "golo", "camião", "bebé", "casa de banho", "rapaz", "rapariga", "licenciatura", "género", "higiene", "castanho", "fato", "frigorífico", "congelador", "ozono", "fiambre", "bowling", "rato", "sumo", "autoestradas", "stress", "digressão", "registo", "chávena", "talho", "t-shirt", "gelado"].

English translation: bus, the buses, train, you, cool, guy, we, dog, cellphone, became, high school, award, ceremony, canonical, curriculum, academic, sport, reception, ironic, winner, committee, to plan, control, cancer, in a, in one, balcony, kilometre, city hall, pineapple, papaya, shorts, economic, team, goal, truck, baby, bathroom, boy, girl, bachelor's degree, gender, hygiene, brown, suit, refrigerator, freezer, ozone, ham, bowling, mouse, juice, highways, stress, tour, record, cup, butcher shop, t-shirt, ice cream.

C Full Results

C.1 Reranking EP

Test Bucket	# Unt.	BLEU	chrF	BLEURT
Entity	318	42.39	66.34	74.15
Entity (BL)	-	45.63	68.56	76.74
Lexical	312	**35.19**	**61.76**	70.02
Lexical (BL)	-	34.18	60.66	70.98
Random	295	40.15	66.79	71.91
Random (BL)	-	42.70	68.27	74.39

Table 8: Full results for reranking (**EP**). Boldface shows improvement against baseline. We also show the number of untranslated sentences.

C.2 Reranking BP

Test Bucket	# Unt.	BLEU	chrF	BLEURT
Entity	265	49.36	71.68	77.45
Entity (BL)	-	54.84	74.70	79.76
Lexical	70	47.78	70.68	74.85
Lexical (BL)	-	51.98	72.87	76.41
Random	69	50.58	72.78	74.71
Random (BL)	-	56.83	76.02	76.94

Table 9: Full results for reranking (**BP**). No result was better than the baseline. We also show the number of untranslated sentences.

C.3 Inference Biasing

Test Bucket	BLEU	chrF	BLEURT
Entity (EP)	**46.26**	**69.52**	76.62
Entity (BP)	56.60	76.20	79.58*
Lexical (EP)	**34.56**	**61.57****	**70.31**
Lexical (BP)	51.88	72.84*	75.78
Random (EP)	40.24	66.14	72.27**
Random (BP)	54.03*	74.16**	75.48**

Table 10: Evaluation scores after performing domain adaptation at inference time, with scores that improved when compared to the baseline boldfaced.

C.4 Fine-Tuning: 2.8K

Test Bucket	BLEU	chrF	BLEURT
Entity (EP)	**52.61****	**72.89****	**77.55****
Entity (BP)	**60.11****	**78.21****	79.84
Lexical (EP)	**42.77****	**66.96****	**72.59****
Lexical (BP)	**55.52****	**75.13****	**76.66****
Random (EP)	**46.59****	**69.33****	**73.62****
Random (BP)	**56.10**	**75.46**	**76.14**

Table 11: Results after fine-tuning on 2.8K FRMT sentences (exemplars + dev). Boldface results represent scores higher than baseline.

C.5 Fine-Tuning: 10K

Some of 10K sentences are drawn from the Tatoeba dataset, less similar to the evaluation test set, but the additional informal, general-domain data still contributes to better results.

Test Bucket	BLEU	chrF	BLEURT
Entity (EP)	**52.85****	**72.79****	**77.72****
Entity (BP)	**59.97****	**78.27****	**80.10**
Lexical (EP)	**42.48****	**66.88****	**72.58****
Lexical (BP)	**54.97****	**75.00****	**76.57***
Random (EP)	**46.38****	**69.29****	**73.57****
Random (BP)	**56.07**	**75.52**	**76.40**

Table 12: Results after fine-tuning on 10K training sentences. Boldface: scores higher than the baseline, highlights: improvement against 2.8K fine-tuning.

C.6 Fine-Tuning: 50K

Test Bucket	BLEU	chrF	BLEURT
Entity (EP)	**51.30****	**72.07****	**77.46****
Entity (BP)	**59.58****	**78.00****	**80.14**
Lexical (EP)	**41.81****	**66.36****	**72.59****
Lexical (BP)	**54.40****	**74.55****	**76.40**
Random (EP)	**45.82****	**68.76****	**73.51****
Random (BP)	**55.50**	**75.17**	75.97

Table 13: Results after fine-tuning on 50K training sentences, with scores higher than the baseline boldfaced. Performance drops across all buckets when compared to the 10K setting, showing that simply increasing the volume of data can introduce noise and negatively impact results.

C.7 LLM-prompting

Test Bucket	BLEU	chrF	BLEURT
Entity (EP)	**55.29****	**74.49****	**80.47****
Entity (BP)	**62.04****	**79.06****	**83.02****
Lexical (EP)	**45.40****	**68.83****	**75.84****
Lexical (BP)	**55.83****	**75.98****	**79.47****
Random (EP)	**51.23****	**72.44****	**77.10****
Random (BP)	**61.07****	**78.71****	**80.19****

Table 14: Results for Gemini 2.5 Pro translations. All scores are better than all the other models (and therefore boldfaced).