# Experiments in Mamba Sequence Modeling and NLLB-200 Fine-Tuning for Low Resource Multilingual Machine Translation

**Dan DeGenaro***
Department of Linguistics,
Georgetown University
37th and O Sts NW, Washington, D.C.
drd92@georgetown.edu

**Tom Lupicki***
Department of Computer Science,
Georgetown University
37th and O Sts NW, Washington, D.C.
tml89@georgetown.edu

## Abstract

This paper presents DC_DMV's submission to the AmericasNLP 2024 Shared Task 1: Machine Translation Systems for Indigenous Languages. Our submission consists of two multilingual approaches to building machine translation systems from Spanish to eleven Indigenous languages: fine-tuning the 600M distilled variant of NLLB-200, and an experiment in training from scratch a neural network using the Mamba State Space Modeling architecture. We achieve the best results on the test set for a total of 4 of the language pairs between two checkpoints by fine-tuning NLLB-200, and outperform the baseline score on the test set for 2 languages.

## 1 Introduction

The 2024 AmericasNLP Shared Task on machine translation (MT) for Indigenous languages consists of developing an MT system (or systems) for the purpose of translating Spanish to 11 Indigenous languages of the Americas: Aymara (aym), Bribri (bzd), Asháninka (cni), Chatino (ctp), Guaraní (gn), Wixarika (hch), Nahuatl (nah), Hñähñu/Otomí (oto), Quechua (quy), Shipibo-Konibo (shp), and Rarámuri (tar). We take two approaches in parallel, namely finetuning NLLB-200 (Team et al., 2022) and training a Mamba architecture-based neural network (Gu and Dao, 2023) from scratch.[1]

## 2 Data

### 2.1 Data Sources

We utilize data from a number of sources, namely the training and development sets provided by the task organizers, data gathered as part of last year's HelsinkiNLP submission (De Gibert et al., 2023), parallel data from Tatoeba[2] released under a CC-BY 2.0 FR., and pivot translations generated from non-Spanish-to-target language parallel data from the Tatoeba Translation Challenge (Tiedemann, 2020). We include additional data to try to compensate for the sparseness of data available in the target languages more generally.

**Organizer-provided Data** Training and development data for the 11 target languages included in the shared task were released by task organizers[3]. The provided data includes data explicitly denoted as the training set, supplemental translation data from Spanish, and supplemental translation data from English. An overview of the organizer-provided data we used can be found in Table 1.

**HelsinkiNLP** Data collected for the 2023 HelsinkiNLP submission to the shared task (De Gibert et al., 2023) was also provided by the task organizers. This data is sourced from the OPUS corpus collection (Tiedemann, 2012), the FLORES-200 corpus (Team et al., 2022), the JHU Bible corpus (McCarthy et al., 2020), and various other texts spanning legal, educational, and news domains.

**Tatoeba Translation Challenge** Spanish-to-target-language parallel data is available from the Tatoeba website[2] for Guarani, Nahuatl, and Quechua.

**Pivot Translations** The Tatoeba Translation Challenge (Tiedemann, 2020) provides non-Spanish parallel data for Guarani, Nahuatl, and Quechua. We utilize machine translation systems to construct additional parallel language data. Data in English, Esperanto, French, German, Hebrew, Japanese, Macedonian, Polish, Russian, and Ukrainian was translated using bilingual Opus-MT

*Both authors contributed equally to this work.

[1]Code for both of our models is available here: https://github.com/tomlup/americasnlp-2024-st1-dc_dmv

[2]Tatoeba website.

[3]AmericasNLP 2024 Shared Task GitHub



*Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 188–194

| Target Language | Data Source(s) |
|---|---|
| aym | Global Voices (Tiedemann, 2012) |
| bzd | (Feldman and Coto-Solano, 2020) |
| cni | AshanikaMT (Ortega et al., 2020; Cushimariano Romano and Sebastián Q., 2008; Mihas, 2011) |
| ctp | https://scholarworks.iu.edu/dspace/handle/2022/21028 |
| gn | (Chiruzzo et al., 2020) |
| hch | (Mager et al., 2018) |
| nah | Axolotl (Gutierrez-Vasques et al., 2016) |
| oto | https://tsunkua.elotl.mx/about/ |
| quy | JW300 (Agić and Vulić, 2019), Global Voices (Tiedemann, 2012) |
| shp | (Montoya et al., 2019), (Galarreta et al., 2017), https://www.sil.org/resources/archives/30143 |
| tar | (Brambila, 1976) |

Table 1: Sources of data provided by task organizers.

systems (Tiedemann and Thottingal, 2020). Data in Chinese, Javanese, and Portuguese was translated into Spanish using NLLB-200 (Team et al., 2022). Additionally, English-Indigenous language data that was provided as supplemental data by task organizers were also translated using Opus-MT. We make use of pivot translations only in the Mamba model.

### 2.2 Data Organization

For the purposes of training, we organize our collected data into three stages. Stage 1 includes all synthetic parallel texts created by means of pivot translation and synthetic data provided by task organizers. Stage 2 includes the supplemental data sourced from the 2023 HelsinkiNLP submission, as well as other Spanish-source supplemental data provided by task organizers. Stage 3 includes training data provided by the shared task organizers.

### 2.3 Duplicate Filtering

After all training data was organized into stages, all data for each target language was then filtered to remove duplicates using OpusFilter (Aulamo et al., 2020). The pipeline for filtering was as follows: All duplicates within Stage 3 data were removed. Then, all duplicates within Stage 2 and overlap with Stage 3 were removed from Stage 2. Finally, all duplicates within Stage 1 and any overlap with Stage 2 and Stage 3 were removed from Stage 1. The total number of training examples from each stage is shown in Table 2.

| Language | Stage 1 | Stage 2 | Stage 3 |
|---|---|---|---|
| aym | 16,338 | 17,679 | 6,453 |
| bzd | 0 | 0 | 7,303 |
| cni | 13,018 | 0 | 3,860 |
| ctp | 2,762 | 2,246 | 357 |
| gn | 617,894 | 42,184 | 14,500 |
| hch | 505 | 2,628 | 6,587 |
| nah | 9,279 | 2,493 | 15,450 |
| oto | 0 | 9,012 | 4,531 |
| quy | 64,337 | 16,112 | 119,471 |
| shp | 23,125 | 16,719 | 14,511 |
| tar | 0 | 2,254 | 14,658 |
| Total | 747,258 | 110,787 | 207,681 |

Table 2: Overview of data organization by number of examples.

## 3 Methods

### 3.1 Finetuning NLLB-200

Our first method involves fine-tuning the NLLB-200 model (Team et al., 2022). We use the distilled 600M parameter variant, and leave all parameters trainable. We motivate this decision as follows. Given that we are tokenizing previously unseen languages using an already-trained tokenizer, the distribution and linear ordering of tokens in our fine-tuning data will differ vastly from the distribution and linear ordering in the languages previously seen by the model. As such, it is sensible to retrain the entire model, including the embeddings, to model this very different distribution. To that end, we introduce additional language tokens for the eight target languages in the shared task not already represented in the model (all except for Ay-

mara, Guarani, and Quechua), which are randomly initialized.

We finetune on padded mini-batches of size 4 with a maximum sequence length of 384, in which all 4 training examples in a given batch have the same target language. However, batches from all 11 target languages are shuffled together. We optimize using AdamW, with a learning rate of $1 \cdot 10^{-5}$ and a weight decay of $1 \cdot 10^{-4}$.

With regard to training stages, we do not use the Stage 1 data to fine-tune NLLB. The number of epochs through each stage for each of our fine-tuned NLLB models are presented in Table 3.

The generation process for producing translations for evaluation uses a maximum sequence length of 384 and beam search with 4 beams and early stopping.

### 3.2 Mamba State Space Model

Our second method involves training a neural network using repeating multiple Mamba architecture layers and a language model head. We submit results for a model containing 3 Mamba layers and a final linear layer with 256 dimensions, and a vocabulary of 16,000 subword tokens trained on all data using SentencePiece (Kudo and Richardson, 2018) using a unigram language model algorithm (Kudo, 2018).

For the purposes of training our Mamba model, we modify our training data by appending a target language token to the beginning of each source sentence. We additionally append a start of sentence token and end of sentence token to the start and end of each sentence, respectively.

We train our model on padded mini-batches of size 128 with a maximum sequence length of 512. Each mini-batch contains shuffled data taken from all languages and all data used for training during an epoch. We optimize the model using AdamW using a learning rate of $1 \cdot 10^{-3}$ and a weight decay of $1 \cdot 10^{-4}$. The model is trained for 5 epochs through all data (Stage 1, Stage 2, and Stage 3), followed by an additional 25 epochs on combined Stage 2 and Stage 3 data. We motivate our decision to include Stage 1 data only in early training by our belief that our synthetic pivot translations are noisier than original Spanish-source translation data, but find it important to train our model on a wide range of data early on. In this regard, we view our later stages of training on Stage 2 and Stage 3 data as tuning our model on higher quality data.

## 4 Results

We present our results in Tables 5 and 6, alongside results for the two baseline systems. The reported scores are calculated using the chrF++ metric (Popović, 2017), as stipulated by the shared task.

Our NLLB+FT(v2) model beats both baseline systems on the development set for Aymara and Quechua, and both baseline systems on the test set for Quechua and Rarámuri. Additionally, several of our models beat at least one baseline system on the development set for Bribri, Nahuatl, Quechua, and Shipibo-Konibo.

Of all submissions this year, our NLLB+FT(v2) model achieves the best result for Aymara, Shipibo-Konibo, and Rarámuri, and our NLLB+FT(v4) model achieves the best result for Bribri, as evaluated on the test set. Our NLLB+FT(v2) and NLLB+FT(v4) models achieve average chrF++ scores across all languages of 22.17 and 23.32 respectively, with NLLB+FT(v4) representing the second best overall submission.

Interestingly, while our models did not achieve the best result on the test set for Asháninka, Hñähñu, and Quechua as measured by the official metric, at least one of our NLLB+FT models outperformed the best submission in BLEU score (Post, 2018). We report these scores in Table 4.

Our Mamba model shows poor performance at the stage in training at time of submission. However, we believe much of this to be due to undertraining given that our model is trained from scratch. With this in mind, we believe continued training may lead to success of our Mamba model, and plan to continue experiments with this architecture.

## 5 Conclusion

In this paper, we presented our submission to the AmericasNLP 2024 shared task on machine translation systems for Indigenous languages. Our submissions included six versions of a fine-tuned 600M parameter distilled variant of NLLB-200, and one Mamba-based model trained from scratch. We trained all of our models on multilingual data to translate from Spanish to 11 target Indigenous languages. We achieve the best chrF++ scores on 4 languages with our fine-tuned NLLB-200 models, improving upon the baseline systems for two languages and setting a new highest score for Rarámuri. Additionally, we find our Mamba-based

| Version | # Epochs Stage 2 | # Epochs Stage 3 | # Epochs Addl. Stage 2 | # Epochs Addl. Stage 3 |
|---|---|---|---|---|
| v1 | 3 | 10 | 0 | 10 |
| v2 | 3 | 10 | 3 | 0 |
| v3 | 3 | 10 | 3 | 3 |
| v4 | 3 | 10 | 3 | 4 |
| v5 | 3 | 10 | 0 | 6 |
| v6 | 3 | 10 | 3 | 8 |

Table 3: Our six fine-tuned NLLB submissions differ solely in the number of epochs through each fine-tuning stage. All models were trained for 3 epochs on the Stage 2 data (# Epochs Stage 2), followed by 10 epochs on the stage 3 data (# Epochs Stage 3). We then experiment with training the models on the Stage 2 data again (# Epochs Addtl. Stage 2), on the Stage 3 data again (# Epochs Addtl. Stage 2), or both. The order in which this process occurs is laid out left-to-right in the table. For instance, NLLB+FT(v6) was trained, in order, for 3 epochs through Stage 2, followed by 10 epochs through Stage 3, followed by 3 more epochs through stage 2, and finally 8 epochs through Stage 3.

| Language | v1 | v2 | v3 | v4 | v5 | v6 |
|---|---|---|---|---|---|---|
| cni | **3.56** | 3.52* | 3.56* | 3.51* | 3.41* | 3.49* |
| oto | 1.55* | 1.46* | **1.66** | 1.49* | 1.52* | 1.36 |
| quy | 4.01 | **5.41** | 4.13 | 4.32 | 3.91 | 4.05 |

Table 4: BLEU scores for our six NLLB+FT submissions for the languages on which we achieve a higher BLEU score than the winning submission. The highest score for each language is bolded. All other results that achieve a higher BLEU score than the submission with the highest chrF++ score for that language are denoted with an asterisk.

model to perform poorly given its training, but plan to continue training and experimentation with this architecture.

## Limitations

Due to dialectal and orthographic variation of the Indigenous languages included in this shared task, it is unclear how our systems would perform on language data that spans such variation not represented in the task data. For example, the provided data for Quechua belongs to the Quechua Ayacucho variant of the Southern Quechua dialect[4]. It is unclear how performance would vary for different varieties of Quechua.

## Ethics Statement

To our knowledge, our work on this project adheres to the principles set forth in Schwartz, 2022.

## Acknowledgements

We would like to acknowledge Dr. Kenton Murray from Johns Hopkins University for guiding us as we undertook this project.

[4]Datasets Information

| **Language** | Helsinki | Sheffield | NLLB+FT(v1) | NLLB+FT(v2) | NLLB+FT(v3) | NLLB+FT(v4) | NLLB+FT(v5) | NLLB+FT(v6) | Mamba |
|---|---|---|---|---|---|---|---|---|---|
| aym | 32.63 | 34.28 | 30.83 | **34.38** | 31.95 | 31.90 | 31.12 | 31.56 | 9.98 |
| bzd | 22.65 | **25.03** | 22.82* | 19.18 | 23.59* | 22.98* | 23.36* | 23.25* | 5.01 |
| cni | 25.68 | **26.34** | 23.30 | 20.08 | 24.05 | 23.42 | 23.22 | 23.69 | 11.12 |
| ctp | 30.06 | **37.33** | 16.73 | 8.25 | 16.17 | 15.83 | 16.41 | 16.29 | 2.90 |
| gn | **34.74** | 32.17 | 29.28 | 31.67 | 30.08 | 30.10 | 29.18 | 29.99 | 9.25 |
| hch | **27.98** | **27.98** | 26.16 | 20.25 | 25.65 | 26.16 | 25.90 | 25.92 | 8.56 |
| nah | 22.78 | **25.58** | 23.90* | 19.26 | 22.96* | 23.50* | 23.56* | 23.79* | 11.11 |
| oto | 13.10 | 12.69 | 12.17 | 11.20 | 12.11 | 12.33 | 12.18 | 12.12 | 4.05 |
| quy | 28.78 | 30.22 | 32.49* | **35.46** | 32.64* | 33.13* | 32.54* | 32.61* | 10.78 |
| shp | **30.59** | 28.39 | 24.39 | 29.94* | 26.35 | 26.70 | 25.48 | 25.37 | 10.70 |
| tar | **17.58** | 16.91 | 14.75 | 17.46 | 15.56 | 15.28 | 14.79 | 15.11 | 7.34 |

Table 5: Comparison of chrF++ scores for our models versus this year's baseline scores for the development set. The top score on each language is bolded. We denote with an asterisk any other result that beats one or both baselines. The differences in the NLLB versions are specified in Table 3.



| **Language** | Helsinki | Sheffield | NLLB+FT(v1) | NLLB+FT(v2) | NLLB+FT(v3) | NLLB+FT(v4) | NLLB+FT(v5) | NLLB+FT(v6) | Mamba |
|---|---|---|---|---|---|---|---|---|---|
| aym | 29.36 | **31.84** | 27.36 | 30.97* | 28.12 | 28.32 | 27.09 | 27.48 | 8.69 |
| bzd | 23.47 | **25.58** | 23.19 | 19.60 | 23.32 | 23.47 | 23.41 | 23.15 | 4.72 |
| cni | **24.92** | 24.76 | 22.44 | 19.89 | 22.87 | 22.46 | 22.53 | 22.98 | 9.81 |
| ctp | 29.84 | **37.05** | 16.52 | 8.06 | 16.17 | 15.78 | 16.11 | 16.04 | 0.91 |
| gn | **37.02** | 35.76 | 31.58 | 33.31 | 32.58 | 32.44 | 31.22 | 31.66 | 8.91 |
| hch | **28.67** | 28.28 | 26.46 | 19.56 | 25.60 | 26.23 | 25.97 | 25.66 | 7.12 |
| nah | 22.78 | **23.28** | 21.63 | 18.52 | 21.07 | 21.44 | 21.43 | 21.41 | 10.46 |
| oto | **13.32** | 12.87 | 12.63 | 11.50 | 12.42 | 12.42 | 12.40 | 12.20 | 3.63 |
| quy | 28.81 | 34.01 | 33.91* | **36.02** | 34.08* | 34.29* | 33.94* | 33.56* | 11.42 |
| shp | **30.21** | 30.06 | 22.25 | 29.37 | 23.84 | 24.74 | 23.59 | 23.05 | 9.67 |
| tar | 16.98 | 16.25 | 14.39 | **17.03** | 15.42 | 14.92 | 14.51 | 14.57 | 5.29 |

Table 6: Comparison of chrF++ scores for our models versus this year's baseline scores for the test set. The top score on each language is bolded. We denote with an asterisk any other result that beats one or both baselines. The differences in the NLLB versions are specified in Table 3.

## References


Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online. Association for Computational Linguistics.

David Brambila. 1976. Diccionario raramuri – castellano (tarahumara).

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Rubén Cushimariano Romano and Richer C. Sebastián Q. 2008. Ñaantsipeta asháninkaki birakochaki. diccionario asháninka-castellano. versión preliminar. http://www.lengamer.org/publicaciones/diccionarios/.

Ona De Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2023. Four Approaches to Low-Resource Multilingual NMT: The Helsinki Submission to the AmericasNLP 2023 Shared Task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 177–191, Toronto, Canada. Association for Computational Linguistics.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ana-Paula Galarreta, Andrés Melgar, and Arturo Oncevay. 2017. Corpus creation and initial SMT experiments between Spanish and Shipibo-konibo. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 238–244, Varna, Bulgaria. INCOMA Ltd.

Albert Gu and Tri Dao. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint*. ArXiv:2312.00752 [cs].

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. *arXiv preprint*. ArXiv:1804.10959 [cs].

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.

Manuel Mager, Carrillo Dionico, and Ivan Meza. 2018. The wixarika-spanish parallel corpus.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Elena Mihas. 2011. *Añaani katonkosatzi parenini, El idioma del alto Perené*. Milwaukee, WI: Clarks Graphics.

Héctor Erasmo Gómez Montoya, Kervy Dante Rivas Rojas, and Arturo Oncevay. 2019. A continuous improvement framework of machine translation for Shipibo-konibo. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 17–23, Dublin, Ireland. European Association for Machine Translation.

John Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020. Overcoming resistance: The normalization of an Amazonian tribal language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13, Suzhou, China. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages

724–731, Dublin, Ireland. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint*. ArXiv:2207.04672 [cs].

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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.