Leveraging Large Language Models for Spanish-Indigenous Language Machine Translation at AmericasNLP 2025

Mahshar Yahan, Dr. Mohammad Amanul Islam
Department of Computer Science and Engineering
Uttara University, Bangladesh
mahshar@uttara.ac.bd, amanul.islam@uttarauniversity.edu.bd

Abstract

This paper presents our approach to machine translation between Spanish and 13 Indigenous languages of the Americas as part of the AmericasNLP 2025 shared task. Addressing the challenges of low-resource translation, we fine-tuned advanced multilingual models, including NLLB-200 (Distilled-600M), Llama 3.1 (8B-Instruct) and XGLM 1.7B, using techniques such as dynamic batching, token adjustments, and embedding initialization. Data preprocessing steps like punctuation removal and tokenization refinements were employed to achieve data generalization. While our models demonstrated strong performance for Awajun and Quechua translations, they struggled with morphologically complex languages like Nahuatl and Otomí. Our approach achieved competitive ChrF++ scores for Awajun (35.16) and Quechua (31.01) in the Spanish-to-Indigenous translation track (Es→Xx). Similarly, in the Indigenous-to-Spanish track (Xx→Es), we obtained ChrF++ scores of 33.70 for Awajun and 31.71 for Quechua. These results underscore the potential of tailored methodologies in preserving linguistic diversity while advancing machine translation for endangered languages.

1 Introduction

Nearly half of the world's 7,000 languages are currently endangered[1]. Experts predict that around 1,500 of these languages could vanish by the end of this century due to factors like globalization, economic growth, and insufficient support for Indigenous languages[2]. Indigenous languages are not just cultural gems but also hold unique perspectives and knowledge. The United Nations has declared 2022–2032 as the International Decade of Indigenous Languages, highlighting the urgency of this issue (Boodeea et al., 2025).

Machine Translation (MT) presents significant challenges, particularly in low-resource settings. Limited data availability, the presence of diverse dialects, and complex linguistic structures such as polysynthesis significantly increase the challenges. However, recent improvements in neural machine translation (NMT) and multilingual learning have shown promise. For example, models like Meta's NLLB-200 (Distilled-600M)(Costa-Jussà et al., 2022) and fine-tuned methods using Low-Rank Adaptation(LoRA) (Hu et al., 2022) have worked well in low-resource settings, improving translation accuracy while helping preserve languages with the involvement of Indigenous communities.

The AmericasNLP 2025 Shared Task focuses on translating between Spanish and 13 Indigenous languages, such as Quechua, Guarani, and Wayuunaiki. This project uses advanced MT techniques and works closely with Indigenous communities to create accurate and culturally respectful translation models. By using advanced techniques like improved tokenization and batching, the initiative aims to build strong MT systems that respect linguistic diversity while pushing forward the field of computational linguistics.

This task is an important step towards using technology to bridge cultural gaps, ensuring that Indigenous voices are heard and preserved for future generations.

The implementation details have been provided in a GitHub repository[3].

2 Related Work

MT has emerged as a promising solution for low-resource languages. Fine-tuning large language models and innovative tokenization strategies have played a big role in these improvements. However, challenges such as limited training data, linguistic

[1] https://www.science.org/content/article/languages-are-being-wiped-out-economic-growth

[2] https://www.anu.edu.au/news/all-news/1500-endangered-languages-at-high-risk

[3] https://github.com/mahshar-yahan/AmericansNLP-2025/tree/main/Shared%20Task-1

Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 126–133
May 4, 2025 ©2025 Association for Computational Linguistics

diversity, and issues like overgeneration continue to hinder the development of robust systems.

Recent Advancements
Recent advancements in multilingual models have significantly improved translation quality for low-resource languages. (Costa-Jussà et al., 2022) introduced NLLB-200 (Distilled-600M), a massively multilingual model trained on 200 languages, demonstrating the effectiveness of fine-tuning for low-resource settings. A recent study further highlighted the potential of NLLB-200 (Distilled-600M) by showing that fine-tuning this model can substantially improve translation quality for specific language pairs, such as Spanish to Quechua and Spanish to Guarani (Gilabert et al., 2024). Additionally, LoRA-based approaches (Hu et al., 2022) have shown promise by enabling efficient parameter updates in large language models without requiring extensive computational resources. Notably, leveraging LoRA has led to a performance improvement of 14.2%.

Tokenization Strategies
Indigenous languages often exhibit agglutinative or polysynthetic structures that challenge standard tokenization methods. (Attieh et al., 2024) compared various tokenization strategies, including SentencePiece and BPE-MR. They found that BPE-MR performs better for morphologically rich languages by preserving meaningful subword units. Our approach inspired upon these findings by tailoring tokenization strategies to the linguistic characteristics of AmericasNLP languages.

Overgeneration issues
Overgeneration is a well-documented issue in machine translation systems, where models produce excessively long or redundant outputs that compromise translation quality. Prior work has addressed this problem through evaluation metrics and architectural modifications. For instance, LAAL (Length-Adaptive Average Lagging) provides unbiased metrics to measure overgeneration during simultaneous translation tasks (Papi et al., 2022). Additionally, methods such as beam search optimization (Cohen and Beck, 2019) have been proposed to mitigate excessive output length.

Addressing Similar Challenges
MMTAfrica (Emezue and Dossou, 2022) employs backtranslation and reconstruction techniques to enhance multilingual translations for African languages. Similarly, we have utilized backtranslation in our system, enabling each of our models to translate between Spanish and Indigenous languages bidirectionally. On the other hand, ModeLing (Chi et al., 2024) is a benchmark dataset designed to evaluate linguistic reasoning in low-resource settings. This work focused on phenomena such as possessive morphology and word order variation. ModeLing provides insights into linguistic challenges similar to those faced in AmericasNLP.

3 Dataset

The dataset provided by AmericasNLP 2025 in Shared Task 1 (de Gibert et al., 2025) focuses on MT between Spanish and 13 Indigenous languages of the Americas: Awajun (agr), Aymara (ayr), Bribri (bzd), Asháninka (cni), Chatino (ctp), Guarani (grn), Wayuunaiki (guc), Wixarika (hch), Nahuatl (nah), Otomí (oto), Quechua (quy), Raramuri (tar) and Shipibo-Konibo (shp). It is divided into training, development, and test sets. Training samples vary from 3,883 (Asháninka) to 125,008 (Quechua), while development sets contain between 599 and 6,635 samples per language. The test set is mostly balanced, with 1,003 samples per language, except for Awajun (358) and Wayuunaiki (498). The dataset supports two translation sub tasks: Spanish to Indigenous languages (Es→Xx) and Indigenous languages to Spanish (Xx→Es). Across all datasets, we identified an average of approximately 765 new words per language that were not present in the initial vocabulary of the NLLB-200(Distilled-600M) tokenizer (Costa-Jussà et al., 2022), which we used for this task. Among the provided datasets, we have utilized all except Chatino and Rarámuri. Here the number of train, development, and test datasets for different subtasks is shown in the table 1.

4 Methodology

In this section, we explain the process of translating a sentence into a specific language. Here, we will discuss both sub-tracks of AmericasNLP 2025 Shared Task 1, where Spanish is translated to Indigenous languages and vice versa. Additionally, we will see how to handle unknown words while training the model for a new language. Also we explore how sentence length can help reduce translation errors.

[5] https://en.wikipedia.org/wiki/Wayuu_language
[6] https://en.wikipedia.org/wiki/Aymara_language

Language	Train	Dev	Test
agr	21,964	1,018	358
ayr	6,531	996	1,003
bzd	7,508	996	1,003
cni	3,883	883	1,003
ctp	357	499	1,000
grn	26,032	995	1,003
guc	59,715	6,635	498
hch	8,966	994	1,003
nah	16,145	672	1,003
oto	4,889	599	1,003
quy	125,008	996	1,003
shp	14,592	996	1,003
tar	14,720	995	1,003

Table 1: Language Data Across Stages

4.1 Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for MT. In this step, we have cleaned and standardized text to improve model performance and ensure consistency across languages.

4.1.1 Punctuation Removal

In this step, we remove punctuation marks to ensure uniformity across the dataset. The removal of punctuation helps in the tokenization process as it reduces unnecessary symbols. We used the *MosesPunctNormalizer* (Koehn et al., 2007) function from the *sacremoses* (Face, 2018) library for normalization. For example,
Before Removal: Tujash, senchi nampekaju, nunik jiyanitan nagkamawag, senchi maninau.
After Removal: Tujash senchi nampekaju nunik jiyanitan nagkamawag senchi maninau.

4.1.2 Whitespace and Character Cleaning

Whitespace inconsistencies were addressed by removing extra spaces and ensuring proper formatting. Leading and trailing spaces were trimmed, and multiple spaces were condensed into one. Additionally, invalid characters were identified and removed to avoid errors during tokenization. In the following example an unnecessary extra space before a fullstop is removed,
Before Cleaning: Nuniamuik pishak najaneaku .
After Cleaning: Nuniamuik pishak najaneaku.

4.1.3 Lowercasing

All text was converted to lowercase for consistency unless case sensitivity was required. However, sometimes capitalization is important, like for proper nouns, acronyms, or special terms. In those cases, we keep the original case instead of converting everything to lowercase. To ensure accurate handling of case-sensitive words, we utilized the SpaCy library (Honnibal et al., 2020) for Spanish text processing. SpaCy's built-in Named Entity Recognition (NER) capabilities allowed us to identify and retain the original case for entities like names, locations, and other significant terms. For instance,
Before: Etsa wantintuk yumijau
After: etsa wantintuk yumijau

4.1.4 Handling Unknown Tokens

Unknown tokens are words or symbols not present in the tokenizer's vocabulary. To address this, we introduced <unk> tokens to represent out-of-vocabulary items. During preprocessing, texts containing unknown tokens were flagged for review, allowing us to refine the vocabulary or handle these cases systematically. For instance, rare Indigenous words were either added to the tokenizer or mapped to <unk> during training. This strategy minimized disruptions caused by unseen words while maintaining translation quality.

4.2 Token Adjustment

Since some languages are new to the model, we need to adjust the tokenization process to fit them. This step is essential for helping the model generalize and properly understand Indigenous languages. By doing this, we can improve translation quality and ensure the model handles these languages more effectively.

4.2.1 Adding New Language Tokens

To add new languages in the translation model, we introduced special language tokens. These tokens help the model recognize the source and target languages during both training and inference. The token addition process involved updating the tokenizer's vocabulary and mappings to integrate these new tokens seamlessly. Each language was assigned a unique token, such as <agr_Latn> for Awajun and <spa_Latn> for Spanish. These tokens were added to sentences during training to clearly specify the language. For example:
Before: Yama nagkamchamunmak Chijajai, Timantim, Sukuyá.
After: <agr_Latn>Yama nagkamchamunmak Chijajai, Timantim, Sukuyá.

Language	Closest Supported Language	Basis for Similarity
agr_Latn (Awajun)	quy_Latn (Quechua)	Geographic proximity in Peru and shared agglutinative morphology (Goulder, 2005).
bzd_Latn (Bribri)	grn_Latn (Guarani)	Both are polysynthetic languages with tonal systems in Central and South America (Kann et al., 2022).
cni_Latn (Asháninka)	quy_Latn (Quechua)	Regional proximity in Peru and shared syntactic traits (Goulder, 2005; Bustamante et al., 2020).
guc_Latn (Wayuunaiki)	grn_Latn (Guarani)	Polysynthetic structure and noun incorporation in northern South America.[5]
hch_Latn (Wixarika)	quy_Latn (Quechua)	Shared agglutinative features despite different language families (Goulder, 2005).
nah_Latn (Nahuatl)	ayr_Latn (Aymara)	Typological similarities like agglutination and SOV word order due to historical interactions.[6]
oto_Latn (Otomí)	ayr_Latn (Aymara)	Borrowing from Nahuatl and typological resemblance to Aymara.[6]
shp_Latn (Shipibo-Konibo)	quy_Latn (Quechua)	Shared Amazonian influences and agglutinative morphology (Goulder, 2005; Bustamante et al., 2020).

Table 2: Mapping of Embedding Initialization for Unsupported Languages Based on Linguistic Similarity using NLLB-200 (Distilled-600M)

4.2.2 Embedding Initialization

The NLLB-200 (Distilled-600M) (Costa-Jussà et al., 2022) model directly supports three Indigenous languages: Aymara (ayr_Latn), Guarani (grn_Latn), and Quechua (quy_Latn). However, when extending the model to new languages that are not explicitly supported, embeddings are initialized using representations from linguistically similar languages. For example, Awajun (agr_Latn) uses Quechua(quy_Latn) embeddings due to linguistic similarities. This approach leverages existing knowledge, reducing training time and improving convergence. Using PyTorch, the embedding layer is resized, and new token IDs are mapped to pre-trained embeddings, ensuring compatibility while preserving prior representations. This method enables efficient extension to low-resource languages.

In comparison, models like LLaMA 3.1 (Touvron et al., 2023) and XGLM (Lin et al., 2021) offer multilingual capabilities but do not directly support Indigenous languages. LLaMA 3.1 focuses on eight high-resource languages, such as Spanish and Hindi. XGLM uses a balanced multilingual corpus but lacks direct support for low-resource Indigenous languages.

4.3 Fine Tuning Process

The fine-tuning process was conducted separately for Task 1 (Es→Xx) and Task 2 (Xx→Es) using NLLB-200(Distilled-600) (Costa-Jussà et al., 2022), LLaMA 3.1 (Touvron et al., 2023), and XGLM (Lin et al., 2021) models. Each model was adapted to the specific translation direction by leveraging its pre-trained multilingual capabilities.

For NLLB, the training process involved freezing encoder layers to reduce computational overhead while updating decoder layers for task-specific adaptation. The model was fine-tuned using a custom training loop with Adafactor optimizer and a constant learning rate scheduler with warm-up steps. Training batches were dynamically generated, ensuring source-target alignment through language-specific tokens. Periodic checkpoints were saved, and the best-performing model was selected based on ChrF++ scores on the development set. Language-specific tokens (e.g., spa_Latn for Spanish and agr_Latn for Awajun) were used to guide the model during training and evaluation.

For LLaMA 3.1 and XGLM, we followed a similar fine-tuning strategy but incorporated the parameter-efficient technique LoRA. This method allowed us to train adapter layers in self-attention blocks while freezing most of the model's parameters. Dynamic batching was employed, where language pairs were randomly selected for each batch. It allowed the model to learn from diverse linguistic contexts and improve generalization across languages. Mixed-precision training was employed to further optimize GPU utilization. Both models were fine-tuned using the same bilingual datasets but with task-specific configurations for each translation direction.

4.4 Post Processing

To ensure the translated text remains concise and relevant, we first determined the length of the original sentence and compared it to the length of the translated output. If the translated text was more than twice the length of the original, we retained only the first 1.25 times the original length. Since we used a causal learning model, it sometimes generated extra information. This method helped control excessive output while maintaining translation quality.

5 Results and Analysis

The evaluation of our system in the AmericasNLP 2025 Shared Task on MT revealed mixed results across languages for both Track 1 (Spanish to Indigenous languages) and Track 2 (Indigenous languages to Spanish) will be discussed in this section. Our experiments utilized fine-tuned versions of NLLB-200 (Distilled-600M) (Costa-Jussà et al., 2022), XGLM 1.7B (Lin et al., 2021), and Llama 3.1(8B-Instruct) (Touvron et al., 2023), focusing on multilingual setups to optimize performance across diverse linguistic structures. The test results of the submitted system using NLLB-200 (Distilled-600M) are presented in Table 6.

5.1 Hyper Parameter Setting

Table 5 shows parameter settings for different models.

In Table 5, *lr*, *optim*,*la* and *l4* represents *learning_rate*,*optimizer*, *lora_alpha* and *load_in_4bit* and respectively.

5.2 Evaluation Metrics

The performance of various models has been evaluated using the Bilingual Evaluation Understudy (BLEU) score, the Character-level F-score (ChrF), and the Character-level F-score++ (ChrF++) metrics on the development and test dataset.

5.3 Comparative Analysis

In this subsection, we provide a detailed analysis of the performance of different models across both development and test datasets for the submitted languages. Using Table 3 and Table 4, which present development results, and Table 6, summarizing test results, we analyze the performance of submitted models across languages. This comparison helps identify trends and determine which models perform better for specific languages in both tracks.

5.3.1 Track 1 (Es→Xx)

NLLB-200 (Distilled-600M) consistently outperformed LLaMA 3.1 and XGLM across all languages on both development and test datasets. While all models performed below baseline, notable trends were observed in Awajun (agr) and Quechua (quy), where results approached the baseline. For the test data, NLLB-200 achieved the highest ChrF++ scores, with 35.16 for agr and 31.01 for quy, demonstrating its ability to handle low-resource Indigenous languages. On the development data, agr and quy also performed well, with ChrF++ scores of 31.55 and 40.01, respectively, showing consistency across datasets.
LLaMA 3.1 exhibited moderate performance for agr on development data (25.17 ChrF++) but struggled with other languages, including quy (13.74 ChrF++). XGLM performed the weakest overall, with ChrF++ scores of 20.44 for agr and only 9.45 for quy on development data, indicating significant challenges in adapting to low-resource settings. However, even in NLLB-200 (Distilled-600M), the best-performed model also showed poor performance relative to the baseline, particularly for morphologically complex languages like Nahuatl (ChrF++: 13.88 vs. baseline 26.36) and Wayuunaiki (ChrF++: 14.40 vs. baseline 24.74) on test results. These results highlight challenges in handling linguistic diversity despite leveraging advanced models.

5.3.2 Track 2 (Xx→Es)

The performance of NLLB-200, LLaMA 3.1, and XGLM in Track 2 was evaluated using ChrF++

Language	NLLB-600M		Llama 3.1 (8B-Instruct)		XGLM 1.7B	
	BLEU	**ChrF++**	**BLEU**	**ChrF++**	**BLEU**	**ChrF++**
	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++
agr	5.97	31.55	5.11	25.17	3.25	20.44
aym	4.03	30.11	4.09	28.13	2.51	22.45
bzd	3.63	16.25	2.72	15.19	1.85	12.37
cni	2.35	24.24	2.02	22.46	1.45	18.92
grn	3.44	19.53	2.57	20.13	1.83	16.24
guc	1.11	17.56	0.56	11.44	0.32	8.76
hch	8.66	28.17	6.79	24.21	4.32	19.87
nah	1.13	14.64	0.93	10.29	0.61	7.85
oto	0.62	15.12	0.23	6.43	0.15	4.21
quy	2.43	40.01	1.16	13.74	0.78	9.45
shp	1.30	18.12	1.01	9.76	0.67	6.32

Table 3: Comparison of BLEU and ChrF++ scores of development data across different models and languages of Es to Xx(Track 1).

Language	NLLB-600M		Llama 3.1 (8B-Instruct)		XGLM 1.7B	
	BLEU	**ChrF++**	**BLEU**	**ChrF++**	**BLEU**	**ChrF++**
agr	11.12	32.80	9.45	28.17	6.73	23.54
aym	8.82	31.72	7.21	26.85	5.34	22.16
bzd	4.31	26.74	3.52	22.18	2.65	18.72
cni	2.85	21.20	2.31	17.65	1.74	14.84
grn	8.62	32.07	7.15	27.26	5.17	22.45
guc	2.22	12.58	1.78	10.46	1.33	8.81
hch	3.69	23.36	3.05	19.48	2.21	16.35
nah	7.22	26.89	5.86	22.41	4.33	18.82
oto	1.50	19.01	1.23	15.84	0.90	13.31
quy	8.76	33.83	7.18	28.76	5.26	23.68
shp	7.22	27.33	5.87	23.23	4.33	19.13

Table 4: Comparison of BLEU and ChrF++ scores of development data across different models and languages of Xx to Es(Track 2).

Model	**lr**	**optim**	**la**	**l4**
NLLB-200 (Distilled-600M)	$2e^{-4}$	Ada Factor	-	-
Llama 3.1 (8B-Instruct)	$3e^{-3}$	Paged Adamw	4	8
XGLM 1.7B	$3e^{-3}$	Adam	4	8

Table 5: Parameter settings for different models

scores on both development and test datasets. Similarly, as track 1 Awajun (agr) and Quechua (quy) showed results approaching the baseline, demonstrating better adaptability compared to other languages.

On the development data, NLLB-200 outperformed the other models across all languages. It achieved ChrF++ scores of 32.80 for agr and 33.83 for quy, showcasing its strong multilingual capabilities. LLaMA 3.1 followed with moderate performance, scoring 28.17 ChrF++ for agr and 22.86 ChrF++ for quy, indicating some adaptability to low-resource languages in this track. XGLM exhibited weaker performance overall, with ChrF++ scores of 23.54 for agr and 20.36 for quy, reflecting its challenges in handling complex linguistic diversity.

On the test data, NLLB-200 maintained its dominance, achieving ChrF++ scores of 33.70 for

Language	Es to Xx (Track 1)			Xx to Es (Track 2)		
	BLEU	**ChrF**	**ChrF++**	**BLEU**	**ChrF**	**ChrF++**
agr	7.82	40.10	35.16[1]	13.21	36.11	33.70[2]
aym	1.96	31.61	27.72[1]	5.89	27.53	25.78[1]
bzd	4.55	21.68	22.77[1]	5.87	27.53	26.22[2]
cni	2.43	26.96	23.17[1]	3.06	21.34	20.13[2]
grn	3.46	17.84	16.21[2]	15.14	26.15	24.70[2]
guc	0.11	15.86	12.83[2]	3.14	16.19	14.40[2]
hch	11.07	30.47	26.77[1]	3.98	23.69	22.02[2]
nah	0.65	15.73	12.64[2]	4.00	15.40	13.88[2]
oto	0.76	14.16	12.02[1]	1.50	19.91	17.80[1]
quy	3.07	36.14	31.01[2]	10.60	33.26	31.71[2]
shp	0.37	14.94	12.76[2]	8.94	32.58	30.83[2]

Table 6: Translation Evaluation Metrics for submitted test languages using NLLB-200 (distilled-600M)

agr and 31.71 for quy, coming close to the baseline scores of 38.39 (agr) and 37.18 (quy). These results highlight NLLB-200's ability to generalize well across datasets. However, even NLLB-200 struggled with morphologically complex languages like Nahuatl (nah), scoring only 13.88 ChrF++, which is below its baseline of 26.36 ChrF++.

Overall, NLLB-200 delivered solid results in both tracks for Awajun (agr), indicating that the token adjustments effectively compensated for the model's lack of direct understanding of the language. This demonstrates the adaptability of NLLB-200 in handling low-resource languages through fine-tuning. LLaMA 3.1 exhibited moderate potential, particularly for Awajun (agr) and Quechua (quy), suggesting that further fine-tuning could enhance its performance in these languages. However, all models, including NLLB-200, showed relatively poor performance compared to the baseline for morphologically complex languages like Nahuatl (nah) and Otomí (oto), highlighting the challenges posed by such linguistic diversity.

6 Conclusion

This research work on MT provided valuable insights into the challenges and potential of translating between Spanish and Indigenous languages. Our approach incorporated techniques like token adjustments and dynamic batching to address linguistic diversity and complex grammatical structures. The results highlighted both the strengths and limitations of our models. While Awajun and Quechua showed decent performance, most other languages underperformed against the baseline, revealing gaps in handling morphosyntactic complexities. This study shows the importance of developing tailored strategies for Indigenous languages, which often feature unique linguistic phenomena such as polysynthesis and agglutination.

7 Limitations

Our models struggled to consistently outperform the baseline in most languages, likely due to difficulties in handling complex grammar and sentence structures. Training large models like NLLB-200 (Distilled-600M) and Llama required powerful GPUs, which were not fully available. This constraint impacted critical processes such as hyperparameter tuning and token adjustments, which are essential for optimizing performance. Additionally, the reduced training duration (limited to 5 epochs) further hindered the models' ability to fully adapt to the linguistic intricacies of the target languages.

8 Future Work

Future efforts will focus on addressing the challenges identified in this study to improve translation quality for Indigenous languages. First, increasing training epochs and leveraging more powerful computational resources will allow for better fine-tuning of large models. Exploring transfer learning from linguistically similar languages may also enhance performance for underperforming cases like Guarani and Nahuatl. Another key area for improvement is the development of specialized architectures or fine-tuning strategies tailored to polysynthetic and agglutinative languages. Finally, expanding the dataset with diverse linguistic phenomena and experimenting with ensemble methods could further enhance translation accuracy and robustness across all languages.

References

Joseph Attieh, Zachary Hopton, Yves Scherrer, and Tanja Samardzic. 2024. System description of the nordicsalps submission to the americasnlp 2024 machine translation shared task. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 150–158.

Zaheenah Beebee Jameela Boodeea, Sameerchand Pudaruth, Nitish Chooramun, and Aneerav Sukhoo. 2025. Automatic translation between kreol morisien and english using the marian machine translation framework. In *Informatics*, volume 12, page 16. MDPI.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from pdf files of truly low-resource languages in peru. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2914–2923.

Nathan A Chi, Teodor Malchev, Riley Kong, Ryan A Chi, Lucas Huang, Ethan A Chi, R Thomas McCoy, and Dragomir Radev. 2024. Modeling: A novel dataset for testing linguistic reasoning in language models. *arXiv preprint arXiv:2406.17038*.

Eldan Cohen and Christopher Beck. 2019. Empirical analysis of beam search performance degradation in neural sequence models. In *International Conference on Machine Learning*, pages 1290–1299. PMLR.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ona de Gibert, Raul Vazquez, Robert Pugh, Abteen Ebrahimi, Pavel Denisov, Ali Marashian, Enora Rice, Edward Gow-Smith, Juan C. Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno Veliz, Ángel Lino Campos, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2025. Findings of the AmericasNLP shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages. In *Proceedings of the 5th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2025)*, Albuquerque, New Mexico. Association for Computational Linguistics.

Chris C Emezue and Bonaventure FP Dossou. 2022. Mmtafrica: Multilingual machine translation for african languages. *arXiv preprint arXiv:2204.04306*.

Hugging Face. 2018. Sacremoses: A python port of moses tokenizer.

Javier García Gilabert, Aleix Sant, Carlos Escolano, Francesca De Luca Fornaciari, Audrey Mash, and Maite Melero. 2024. Bsc submission to the americasnlp 2024 shared task. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 143–149.

Paul Goulder. 2005. The languages of peru: Their past, present, and future survival. *EnterText*, 2(2).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. Version 3.0.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Katharina Kann, Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, John E Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo A Giménez-Lugo, et al. 2022. Americasnli: Machine translation and natural language inference systems for indigenous languages of the americas. *Frontiers in Artificial Intelligence*, 5:995667.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022. Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation. *arXiv preprint arXiv:2206.05807*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.