Improving Indigenous Language Machine Translation with Synthetic Data and Language-Specific Preprocessing

Aashish Dhawan
University of Florida
aashish.dhawan@ufl.edu

Christopher Driggers-Ellis
University of Florida
driggersellis.cw@ufl.edu

Christan Grant
University of Florida
christan@ufl.edu

Daisy Wang
University of Florida
daisyw@cise.ufl.edu

Abstract

Low-resource indigenous languages often lack the parallel corpora required for effective neural machine translation (NMT). Synthetic data generation offers a practical strategy for mitigating this limitation in data-scarce settings. In this work, we augment curated parallel datasets for indigenous languages of the Americas with synthetic sentence pairs generated using a high-capacity multilingual translation model. We fine-tune a multilingual mBART model on curated-only and synthetically augmented data and evaluate translation quality using chrF++, the primary metric used in recent AmericasNLP shared tasks for agglutinative languages. We further apply language-specific preprocessing, including orthographic normalization and noise-aware filtering, to reduce corpus artifacts. Experiments on Guarani–Spanish and Quechua–Spanish translation show consistent chrF++ improvements from synthetic data augmentation, while diagnostic experiments on Aymara highlight the limitations of generic preprocessing for highly agglutinative languages. All code is publicly released(will be released on submission).

1 Introduction

Many indigenous languages face increasing risk of endangerment due to limited digital presence and scarce linguistic resources, posing significant challenges for the development of robust machine translation (MT) systems (Mager et al., 2023; Woodbury, 2012). Languages such as Aymara and Guarani exemplify this challenge: despite being spoken by sizable communities in the Americas, their complex morphological structures and lack of large-scale parallel corpora hinder effective MT development (Rodríguez et al., 2022). As a result, conventional MT approaches that rely on abundant supervised data often perform poorly in these settings.

Recent advances in neural machine translation (NMT), particularly multilingual pretraining and data augmentation, have opened new opportunities for low-resource MT. Community-driven efforts such as AmericasNLP have demonstrated that multilingual models combined with synthetic data generation can substantially improve translation quality for indigenous languages (Ebrahimi et al., 2023a). Motivated by these findings, this work investigates the impact of synthetic parallel data augmentation on MT performance for Aymara–Spanish and Guarani–Spanish translation (Woodbury, 2012). Specifically, we augment curated datasets with synthetic sentence pairs generated using a high-capacity multilingual MT system, following the data-centric paradigm explored in projects such as (Driggers-Ellis et al., 2025). In addition, we include Spanish–Quechua experiments to assess whether language-specific orthographic normalization can further improve performance for morphologically rich indigenous languages.

We fine-tune the multilingual mBART model on both curated-only and synthetically augmented datasets and evaluate performance using chrF++ as the primary metric. Through controlled experiments and comparative analysis, this study aims to assess whether forward-translated synthetic data can reliably improve translation quality in low-resource indigenous language settings, contributing to ongoing efforts within the AmericasNLP initiative to support language preservation and accessibility (Ebrahimi et al., 2023a).

2 Background and Related Work

Machine translation (MT) for low-resource and indigenous languages remains a persistent challenge due to limited parallel corpora, orthographic variation, under-documented grammar, and high linguistic diversity (Ebrahimi et al., 2023a; Mager et al., 2021). These constraints often render conventional supervised MT approaches ineffective, motivating research into data-efficient and transfer-based techniques.

A prominent direction in low-resource MT has been the use of data augmentation to mitigate data scarcity. Back-translation (Sennrich et al., 2016), which generates synthetic parallel data from monolingual corpora, has been shown to improve translation quality across a range of low-resource settings. In parallel, multilingual pre-trained models such as mBART50 (Tang and et al., 2020) and NLLB-200 (Costa-Jussà et al., 2022) have enabled effective cross-lingual transfer by leveraging shared representations across many languages, achieving strong performance even with limited task-specific supervision.

Community-driven initiatives such as Americas-NLP have played a central role in advancing MT research for indigenous languages of the Americas by providing benchmark datasets, standardized evaluation protocols, and shared tasks (Ebrahimi et al., 2023a; Mager et al., 2021; Ebrahimi et al., 2024). Results from recent AmericasNLP shared tasks indicate that the most successful systems consistently combine multilingual pre-trained models with synthetic data generation. For example, strong submissions have employed NLLB-200 with back-translation (Gow-Smith and Sánchez Villegas, 2023) or explored multilingual transfer using models such as mBART50 and M2M-100 (Tonja et al., 2023). Earlier work, including IndT5 (Nagoudi et al., 2021), further demonstrated the benefits of training models directly on indigenous language corpora.

More recently, synthetic data generation via forward translation has gained attention as a scalable alternative to back-translation for languages with scarce monolingual resources. The MultiScript30k dataset, for instance, translated Spanish captions from Multi30k into indigenous languages such as Aymara and Guarani using NLLB-200(Driggers-Ellis et al., 2025), illustrating the potential of high-quality synthetic parallel data to improve MT performance when paired with robust multilingual models.

Building on these advances, our work investigates the effectiveness of synthetic parallel data augmentation for Aymara–Spanish and Guarani–Spanish translation. We situate our results within the context of the AmericasNLP 2023 shared task, comparing against established benchmarks to assess the impact of forward-translated synthetic data on low-resource indigenous language MT.

Lang	Setting	Split	Total	Valid	Drop %
aym	Curated	Train	6,531	6,092	5.54
		Dev	996	996	0.00
	+Synthetic	Train	35,531	33,712	5.12
		Dev	996	945	5.12
gn	Curated	Train	26,032	25,417	2.36
		Dev	995	995	5.93
	+Synthetic	Train	53,083	52,929	0.29
		Dev	995	995	5.93
quy	Curated	Train	154,008	138,786	9.88
		Dev	996	996	0.20
	+Synthetic	Train	163,651	147,607	9.80
		Dev	996	996	0.20

Table 1: Dataset statistics for Aymara (aym), Guarani (gn), and Quechua (quy) from the AmericasNLP 2023 shared task. "Total" denotes raw sentence pairs; "Valid" denotes pairs retained after preprocessing and filtering. Synthetic data is added only to the training split.

3 Dataset and Methodology

3.1 Curated Datasets

We use curated parallel datasets released as part of the AmericasNLP 2023 Shared Task (Ebrahimi et al., 2023a) for Aymara–Spanish (aym–es), Guarani–Spanish (gn–es), and Quechua–Spanish (quy–es) translation. The Aymara dataset consists of 6,531 training and 996 development sentence pairs, while the Guarani dataset contains 26,032 training and 995 development pairs. For Quechua–Spanish, we use the largest curated dataset in the shared task, comprising 154,008 training sentence pairs and a 996-sentence development set.

These raw datasets are drawn from diverse sources, including governmental documents, educational materials, community-driven projects, and linguistic corpora, and exhibit substantial variation in domain, style, and orthography. As summarized in Table 1, we apply language-specific preprocessing and filtering to remove misaligned, duplicate, or noisy sentence pairs, resulting in moderate reductions in dataset size while preserving the majority of valid training examples.

The curated data spans multiple domains—particularly for Guarani, which includes legal, health, and educational content—leading to substantial lexical and structural diversity. These characteristics reflect well-known challenges for machine translation of Indigenous languages, which often lack standardized orthographies and exhibit rich morphological variation (Mager et al., 2021). The curated datasets serve as the baseline for all experiments, with synthetic data added only to the training splits as described in Section 3.2.

Lang	Setting	Split	Avg Src	Avg Tgt	Tgt/Src
aym	Curated-only	Train	19.62	14.89	0.85
		Dev	11.20	7.09	0.66
	Curated+Synthetic	Train	14.07	10.35	0.74
		Dev	11.23	7.10	0.63
gn	Curated-only	Train	23.23	15.57	0.67
		Dev	11.18	7.21	0.64
	Curated+Synthetic	Train	17.42	12.35	0.71
		Dev	11.18	7.21	0.64
quy	Curated-only	Train	15.17	9.33	0.62
		Dev	11.17	7.44	0.67
	Curated+Synthetic	Train	14.72	9.38	0.64
		Dev	11.17	7.43	0.67

Table 2: Average sentence length statistics for Aymara (aym), Guarani (gn), and Quechua (quy) before and after synthetic data augmentation. Tgt/Src denotes the ratio of average target length to average source length.

Tables 1 and 2 summarize the effect of preprocessing and filtering on dataset size and sentence length statistics for Guarani–Spanish. Filtering removes between approximately 2–10% of sentence pairs across languages, primarily due to extreme length mismatches, duplication, or clear alignment errors, while preserving the majority of valid training examples.

3.2 Synthetic Data Generation

To mitigate data scarcity, we augment the curated parallel datasets with synthetic sentence pairs generated using the MultiScript30k dataset. In this pipeline, the Spanish portion of the Multi30k dataset is forward-translated into Aymara and Guarani using the NLLB-200 (3.3B) multilingual machine translation model (Costa-Jussà et al., 2022). This forward-translation approach produces synthetic parallel data that preserves the semantic content of the original captions while increasing linguistic diversity on the target side.

Unlike conventional back-translation approaches (Sennrich et al., 2016), our method uses Spanish as a high-resource pivot language. This design choice is consistent with recent findings from the AmericasNLP shared tasks, where synthetic data generation via multilingual pretrained models has been shown to improve translation quality for indigenous and low-resource languages (Ebrahimi et al., 2023a; Tonja et al., 2023; Gow-Smith and Sánchez Villegas, 2023). The resulting augmented datasets increase both the volume and variability of training examples, supporting improved generalization in low-resource settings.

3.3 Preprocessing

We apply language-specific preprocessing pipelines to account for the distinct linguistic and orthographic properties of Aymara and Guarani while maintaining a consistent overall workflow.

Guarani. For Guarani, we employ a more specialized preprocessing pipeline to address orthographic variability and phonological representation. All text is normalized using Unicode **NFKC** to canonicalize visually equivalent and compatibility characters, followed by lowercasing and whitespace normalization. We remove non-linguistic symbols while explicitly preserving Guarani-specific diacritics and nasal vowels (e.g., ã, ẽ, ĩ, õ, ũ) as well as essential punctuation.

To reduce orthographic inconsistency across sources, we standardize frequent Guarani digraphs by merging space-separated realizations into single units (e.g., "c h"→"ch", "m b"→"mb", "n g"→"ng"). This step improves token consistency and reduces fragmentation during subword tokenization. Finally, to mitigate parallel corpus noise, we apply a length-ratio filter based on word counts. Given a source sentence of length L_s and a target sentence of length L_t, a sentence pair is retained only if $1/\tau \le L_t/L_s \le \tau$, with $\tau = 2.5$. This heuristic removes highly misaligned sentence pairs while preserving the majority of valid training examples.

Quechua. For Spanish–Quechua (quy), we observe systematic intra-word spacing artifacts on the target side (e.g., ch aypiqa, sin ch i, uma ll iqniy, ch u), which introduce tokenization noise and disproportionately affect string-based evaluation metrics. To address this, we apply a deterministic orthographic normalization procedure that targets frequent split patterns without relying on external linguistic resources. Specifically, we (i) merge ch or ll followed by vowel-initial fragments (e.g.,ch a...→cha...), (ii) merge three-token se-

quences such as `sin ch i`→`sinchi`, (iii) normalize isolated sequences such as `ch u`→`chu`, and (iv) merge single-character alphabetic fragments with adjacent tokens when they form valid Quechua word patterns. In addition to orthographic normalization, we apply conservative corpus filtering to remove empty or punctuation-only lines, boilerplate or URL-like content, exact duplicates, severe numeric mismatches, extreme length-ratio outliers, and excessively long sentences. We optionally append bilingual dictionary entries as short parallel pairs to the training data to improve lexical coverage during fine-tuning. For evaluation, the same Quechua normalization is applied to both system outputs and reference translations, ensuring that metric computation reflects orthographic equivalence rather than spacing artifacts.

Aymara. For Aymara–Spanish (aym), we apply a deliberately conservative preprocessing pipeline designed to improve data cleanliness while preserving surface orthographic structure. This includes Unicode normalization (NFKC), normalization of apostrophe variants, whitespace normalization, and removal of empty or misaligned sentence pairs. To reduce obvious alignment noise, we filter sentence pairs exhibiting extreme length mismatches using the same length-ratio heuristic applied to Guarani, and remove exact duplicate pairs.

In addition, we address a systematic corpus artifact in which apostrophes— used in common Aymara orthographies to mark ejective or glottalized consonants— are separated from surrounding characters by spurious whitespace (e.g., `jach 'a`, `t 'äw`, `qilqt 'am`). We apply a deterministic orthographic normalization rule that merges intra-word `letter ' letter` patterns into a single token (e.g., `jach'a`, `t'äw`), without introducing any additional linguistic rewriting.

Despite improving overall data consistency, these preprocessing steps yield limited gains in translation quality for Aymara. We hypothesize that this is due to the language's strongly agglutinative morphology, in which grammatical information is encoded through productive suffix chains and phonemic contrasts. Generic normalization combined with subword tokenization may fragment these morphemes, motivating morpheme-aware preprocessing or segmentation strategies as future work (Mager et al., 2021).

3.4 Model Fine-tuning

We fine-tune the multilingual mBART model (`facebook/mbart-large-50`) using the Hugging-Face Transformers library. mBART is pretrained using a denoising autoencoding objective across 50 languages, enabling effective cross-lingual transfer in low-resource translation scenarios (Tang and et al., 2020). We use the `MBart50Tokenizer` with Spanish (`es_XX`) as the source language and the appropriate target language tags (`aym_XX` for Aymara and `gn_XX` for Guarani).

To better support Guarani orthography, we extend the tokenizer vocabulary with Guarani-specific characters and frequent multi-character units, including nasal vowels and diacritics ($ã, ẽ, ĩ, õ, ũ$), the combining tilde, and common digraphs (*ch, mb, ng*). After extending the tokenizer, the model's embedding matrix is resized accordingly.

For Aymara–Spanish translation, we train using a learning rate of 2×10^{-5}, batch size 16, gradient accumulation of 4, and 20 epochs. For Guarani–Spanish translation, we use a learning rate of 3×10^{-5}, batch size 8, gradient accumulation of 4, and 15 epochs. All models are optimized using AdamW with a cosine learning rate schedule and warm-up. Early stopping with a patience of three evaluation intervals based on validation BLEU is applied to prevent overfitting. These configurations are consistent with recent AmericasNLP submissions (Gow-Smith and Sánchez Villegas, 2023; Tonja et al., 2023).

3.5 Evaluation Metrics

We evaluate translation quality primarily using chrF++, computed with the official AmericasNLP 2023 evaluation script. While BLEU is monitored during training for early stopping, we do not report BLEU scores in our main results due to its known limitations for morphologically rich and agglutinative languages such as Guarani, Quechua, and Aymara(Popović, 2015, 2017).

4 Experimental Results

4.1 Aymara–Spanish Translation

For Aymara–Spanish translation, mBART fine-tuning yields limited improvements under both curated-only and synthetically augmented settings. While chrF++ increases modestly with synthetic data, overall performance remains low, consistent with prior findings for highly agglutinative languages under subword tokenization. These results

suggest that generic normalization and data augmentation are insufficient for Aymara, motivating morpheme-aware approaches discussed in Section 6.

4.2 Guarani–Spanish Translation

A similar trend is observed for Guarani–Spanish translation. The curated-only model achieves a chrF++ score of 42.00 on the development set. Incorporating synthetic data improves performance to 44.00 chrF++, yielding a gain of +2.00 points. This result indicates that forward-translated synthetic data provides consistent benefits even when training data is limited.

As with Aymara–Spanish, training dynamics for Guarani–Spanish exhibit stable optimization behavior, with improved validation performance across epochs when synthetic data is included. Together, these results demonstrate that the benefits of synthetic augmentation extend beyond a single language pair.

4.3 Quechua–Spanish Translation

For Spanish–Quechua translation, applying deterministic orthographic normalization yields substantial improvements in chrF++. In particular, resolving systematic intra-word spacing artifacts (e.g., *ch aypiqa*, *sin ch i*) reduces token fragmentation and improves character-level alignment. On the AmericasNLP 2023 development set, our model achieves a chrF++ score of 36.6, outperforming both the shared-task baseline and matching the best reported systems from the Sheffield submission. As in prior work, performance under n-gram-based metrics remains challenging due to Quechua's agglutinative morphology, and we therefore emphasize chrF++ as the primary evaluation metric.

4.4 Effect of Synthetic Data Augmentation

Across both language pairs, synthetic data augmentation consistently improves chrF++ scores relative to curated-only training. These findings align with prior work showing that synthetic parallel data generated using high-capacity multilingual models can significantly enhance MT performance for indigenous and other low-resource languages. In line with results reported in recent AmericasNLP shared tasks, our experiments confirm that forward-translated synthetic data can serve as an effective and scalable strategy for improving NMT quality in data-scarce scenarios.

4.5 Comparison with AmericasNLP 2023

We compare our systems against the AmericasNLP 2023 baseline and the best per-language results reported by the University of Sheffield on the official development sets. Following prior work, we focus on chrF++ due to BLEU's known limitations for agglutinative languages(Popović, 2015)(Popović, 2017).

Top-performing submissions primarily relied on large multilingual models such as NLLB-200(Costa-Jussà et al., 2022), often paired with back-translation or other forms of synthetic data generation (Gow-Smith and Sánchez Villegas, 2023; Tonja et al., 2023). In particular, the University of Sheffield system employed NLLB-200 with extensive back-translation and diverse parallel sources, achieving strong results across multiple language pairs (Gow-Smith and Sánchez Villegas, 2023). These findings highlight the effectiveness of data-centric approaches in low-resource indigenous MT.

Our method aligns with this paradigm by leveraging synthetic parallel data generated via forward translation using NLLB-200, while fine-tuning a multilingual mBART model. Although our approach differs from prior work in its exclusive use of forward-translated synthetic data from the MultiScript30k pipeline, our results are consistent with trends observed in the shared task: well-curated synthetic data, when paired with robust multilingual models, can substantially improve translation quality for indigenous languages.

5 Baseline Comparison

To contextualize our results, we compare our system against the best-performing submission reported in the AmericasNLP 2023 shared task for Guarani–Spanish translation (Ebrahimi et al., 2023a). While differences in training data and augmentation strategies prevent a strictly controlled comparison, this provides a meaningful reference point within the established evaluation framework. Our results indicate that synthetic data augmentation improves performance over a curated-only baseline and yields competitive chrF++ scores relative to reported shared-task systems.Table 3 compares our dev-set chrF++ scores with the AmericasNLP 2023 baseline and the best per-language results reported by the University of Sheffield submission. Table 3 shows that synthetic data augmentation improves chrF++ across all three languages.

Team / System (dev chrF)	aym	gn	quy
2021 Baseline (Vázquez et al., 2021)	15.70	19.30	30.40
2021 Best System	28.30	33.60	34.30
Andes	9.22	–	–
CIC-NLP	19.05	21.75	35.62
Helsinki-NLP	33.44	40.42	37.19
LCT-EHU	–	–	38.59
LTLAmsterdam	25.23	32.89	36.81
PlayGround	29.98	33.17	34.38
Sheffield	36.24	39.34	39.52
↑ 2021 (Best – Baseline)	+12.60	+14.30	+3.90
↑ 2023 (Best – 2021 Best)	+7.94	+6.82	+5.22
Ours (Curated-only)	26.85	42.00	37.12
Ours (Curated + Synthetic)	30.82	44.00	37.83

Table 3: Development-set chrF++ comparison for Aymara (aym), Guarani (gn), and Quechua (quy) on the AmericasNLP benchmarks(Ebrahimi et al., 2023b). Prior results are taken from the official shared-task reports.

Gains are largest for Aymara (+3.97 chrF++), followed by Guarani (+2.00), while Quechua shows smaller but consistent improvements (+0.71), reflecting its already large curated training set. Although Aymara shows the largest relative chrF++ gain, absolute performance remains low, reinforcing the need for morphology-aware modeling. While direct comparison is limited by differences in model architecture and training configurations, this provides context for the effectiveness of our synthetic data augmentation approach.

6 Discussion

Our results demonstrate that fine-tuning a multilingual mBART model with synthetic data augmentation is an effective strategy for improving MT performance for low-resource indigenous languages such as Aymara and Guarani. The observed gains in chrF++ indicate that forward-translated synthetic data can meaningfully complement small curated corpora, supporting better generalization in data-scarce settings. Language-specific preprocessing further contributes to these improvements by reducing noise while preserving important morphological and phonological structure.

The effectiveness of our approach is consistent with findings from recent AmericasNLP shared tasks, where multilingual pre-trained models combined with synthetic data generation have emerged as strong baselines for indigenous language MT. While some systems—such as the University of Sheffield submission—achieved strong performance using larger NLLB-200 models and back-translation (Gow-Smith and Sánchez Villegas, 2023), our results suggest that competitive improvements can also be obtained using mBART paired with forward-translated synthetic data. This high-

lights the flexibility of data-centric augmentation strategies across different model architectures.

Despite these encouraging results, several limitations remain. Our evaluation relies exclusively on automatic metrics, which may not fully capture cultural nuance or idiomatic correctness in indigenous languages (Mager et al., 2023). Future work should therefore incorporate human evaluation involving proficient speakers. Additionally, further investigation into synthetic data quality, diversity, and filtering strategies—as well as alternative augmentation methods such as back-translation—may yield additional gains. Finally, incorporating explicit linguistic knowledge of Aymara and Guarani into model design or preprocessing pipelines (Garvin and Mathiot, 1972) represents a promising direction for improving translation accuracy while ensuring ethical and community-centered deployment of MT systems (Mager et al., 2023).

7 Conclusion and Future Work

In this work, we investigated synthetic data augmentation and language-specific preprocessing for low-resource indigenous language machine translation, focusing on Guarani–Spanish and Quechua–Spanish, with diagnostic experiments on Aymara.

Future work will extend this framework in several directions. First, we plan to conduct human evaluations to better assess translation quality and cultural adequacy. Second, we aim to explore additional indigenous languages, including Aymara, with more targeted preprocessing and data augmentation strategies. A promising direction for future work is the incorporation of visual context for indigenous language translation. Resources such as MultiScript30k provide aligned

image–caption–translation triples for languages including Guarani, enabling investigation of multimodal machine translation in low-resource settings. Future experiments will explore whether visual grounding can improve translation robustness, reduce ambiguity, or accelerate model convergence when parallel text is scarce. This direction aligns with recent interest in data-efficient and multimodal approaches for low-resource MT.

8 Limitations

This study has several limitations. First, evaluation relies exclusively on automatic metrics (chrF++), which may not fully capture translation adequacy, fluency, or culturally grounded meaning in indigenous languages. Human evaluation by proficient speakers is therefore necessary to assess real-world translation quality. While orthographic normalization improves Quechua translation, it relies on deterministic rules and may not generalize to dialectal variation or other Quechua varieties. Second, synthetic data is generated automatically and may introduce systematic biases or translation artifacts, despite filtering and preprocessing. Finally, while we report detailed experiments for Guarani–Spanish, results for other indigenous languages such as Aymara remain limited and are not fully explored in this work.

References

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*.

Christopher Driggers-Ellis, Detravious Brinkley, Ray Chen, Aashish Dhawan, Daisy Zhe Wang, and Christan Grant. 2025. Multiscript30k: Leveraging multilingual embeddings to extend cross script parallel data.

Abteen Ebrahimi, Manuel Mager, and et al. 2023a. Findings of the americasnlp 2023 shared task on machine translation into indigenous languages. In *Proceedings of the AmericasNLP Workshop*.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023b. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on*

Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Abteen Ebrahimi et al. 2024. Findings of the americasnlp 2024 shared task on machine translation into indigenous languages. In *Proceedings of the AmericasNLP Workshop*.

Paul L Garvin and Madeleine Mathiot. 1972. *The urbanization of the Guarani language: a problem in language and culture*. State University of New York et Buffalo.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the americasnlp shared task on machine translation into indigenous languages. In *Proceedings of the Third Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. *arXiv preprint arXiv:2305.19474*.

Manuel Mager, Arturo Oncevay, et al. 2021. Findings of the americasnlp 2021 shared task on open machine translation for indigenous languages of the americas. In *Proceedings of the AmericasNLP Workshop*.

El Moatez Billah Nagoudi, Yustin Gutiérrez-Vásquez, Arturo Oncevay, Manuel Mager, Xavier Tannier, and Djamé Seddah. 2021. Indt5: A text-to-text transformer for 10 indigenous languages. In *Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*.

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

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics.

Yliana Rodríguez, Luis Chiruzzo, and Santiago Góngora. 2022. The challenges of creating a corpus of minority languages and its dialects in natural language processing: the case of the south american indigenous language guarani. Presented at the 32nd Meeting of Computational Linguistics in The Netherlands (CLIN32).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of ACL*.

Yinhan Tang and et al. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. In *Proceedings of EMNLP*.

Atnafu Lambebo Tonja, Hellina Hailu Nigatu, Olga Kolesnikova, Grigori Sidorov, Alexander Gelbukh, and Jugal Kalita. 2023. Enhancing translation for indigenous languages: Experiments with multilingual models. In *Proceedings of the Third Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*.

Anthony C Woodbury. 2012. What is an endangered language. *Linguistic society of America*.