Findings of the AmericasNLP 2023 Shared Task on Machine Translation into Indigenous Languages

Abteen Ebrahimi[♢] Manuel Mager[♠] Shruti Rijhwani[♡]
Enora Rice[♢] Arturo Oncevay[▽] Claudia Garcia Baltazar
María Elena Méndez Cortés Cynthia Montaño[♣] John E. Ortega[ψ]
Rolando Coto-Solano[Ω] Hilaria Cruz[♯] Alexis Palmer[♢] Katharina Kann[♢]
[♢]University of Colorado Boulder [♠]AWS AI Labs [♡]Google DeepMind
[▽]University of Edinburgh [♣]University of California, Berkeley [ψ]Notheastern University
[Ω]Dartmouth College [♯]University of Louisville

Abstract

In this work, we present the results of the AmericasNLP 2023 Shared Task on Machine Translation into Indigenous Languages. This edition of the shared task features eleven language pairs, one of which – Chatino–Spanish – uses a newly collected evaluation dataset, consisting of professionally translated text from the legal domain. Seven teams participated in the shared task, with a total of 181 submissions. Additionally, we conduct a human evaluation of the best system outputs and compare them to the best submissions from the 2021 shared task. We find that this analysis agrees with the quantitative measure we use to rank submissions, ChrF, which itself shows an improvement of 9.64 points on average across all languages, compared to the prior winning system.

1 Introduction

The majority of Indigenous languages, including those native to the Americas, are under-represented in modern natural language processing (NLP), as technological advances are often concentrated on the small set of languages that have large amounts of easily available data (Joshi et al., 2020). Beyond the lack of data, linguistic factors like morphological complexity, non-standard orthographies, and language isolates make it even more challenging to adapt existing NLP methods to Indigenous languages (Mager et al., 2018; Schwartz et al., 2020).

However, there are multiple benefits of developing technologies that support Indigenous languages – building NLP models for under-represented languages can bring equitable access to information and technology to speakers of these languages (Mager et al., 2018). Additionally, several Indigenous languages in the Americas are endangered, and language technologies have proven to be beneficial to Indigenous communities and linguistic researchers in the documentation, preservation, and revitalization of endangered languages (Galla, 2016; Anastasopoulos, 2019; Zhang et al., 2022; Rijhwani, 2023). The AmericasNLP workshop seeks to highlight NLP and linguistic research on Indigenous languages spoken across the Americas, and promote the development of computational approaches which work well for these languages. The AmericasNLP Shared Task on Machine Translation into Indigenous Languages is hosted as part of the workshop to specifically focus on improvements in machine translation (MT) systems for these languages. In this work, we describe the third edition of the shared task. For this year, a new gold-standard parallel dataset for translation evaluation, between Spanish and Chatino, was developed. This dataset uses text from the legal domain, with source sentences taken from press releases of the Supreme Court of Mexico. This allows for evaluation on technical and challenging text, which are likely to be relevant to speakers of the language.

Language	ISO	Family	Train	Dev	Test
Asháninka	cni	Arawak	3883	883	1002
Aymara	aym	Aymaran	6531	996	1003
Bribri	bzd	Chibchan	7508	996	1003
Chatino	**ctp**	**Oto-Manguean**	**357**	**499**	**1000**
Guarani	gn	Tupi-Guarani	26032	995	1003
Nahuatl	nah	Uto-Aztecan	16145	672	996
Otomí	oto	Oto-Manguean	4889	599	1001
Quechua	quy	Quechuan	125008	996	1003
Rarámuri	tar	Uto-Aztecan	14721	995	1002
Shipibo-Konibo	shp	Panoan	14592	996	1002
Wixarika	hch	Uto-Aztecan	8966	994	1003

Table 1: The languages in the AmericasNLP 2023 shared task. Chatino (bolded) is the new language for this edition of the competition.

This work is structured as follows: in Section 2, we present a brief overview of related work on MT and Indigenous languages; in Section 3 and 4, we provide details on the shared task rules, and newly collected data; in Section 5, we summarize the submitted systems; and, in Sections 6 and 7, we provide an analysis of the main results and further

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 206–219
July 14, 2023 ©2023 Association for Computational Linguistics

	Team	Andes	CIC-NLP	Helsinki-NLP*	LCT-EHU	LTLAmsterdam	Playground	Sheffield*
	Langs	1	11	11	1	11	10	11
	Subs	1	33	66	5	33	10	33
Data	Crawl			✓	✓			
	Ext. Bilingual	✓		✓	✓			
	Opus			✓				
	Religous			✓	✓			✓
	Wikipedia			✓				
	Prior Year			✓	✓		✓	✓
	No Addtl.		✓					
	Monolingual Trans			✓	✓		✓	✓
	Pivot Trans.			✓				
	Cleaning/Norm			✓		✓		✓
Pretraining	ChatGPT					✓		
	Encoder-Decoder		✓	✓	✓	✓		
	M2M-100		✓			✓		
	mBART		✓					
	mT5	✓						
	NLLB						✓	✓
Train	Ensemble							✓
	Multistage		✓	✓			✓	✓
	Multilingual		✓	✓			✓	✓

Table 2: Participating teams (*Team*) with system description paper. The information contained in this table is as follows: number of languages with a corresponding submission (*Langs.*), total number of submissions (*Sub.*). (*Data*) presents a summary of any external data collection, or *No Add.* if no external data was used, as well as if preprocessing steps are described. The *Pretraining* section describes if a pretrained translation model, or from-scratch encoder-decoder architecture was used. The *Train* section provides a summary of the training process for submissions. For more details we refer to the system description paper of each system, and note that certain external datasets or preprocessing steps may have been used within a system and not described in the description paper. We describe how each feature is defined in Appendix A.2.

experiments.

2 Related Work

2.1 NLP for Indigenous Languages

Low-resource languages are often referred to as 'less studied', 'resource-scarce', 'less computerized', 'less privileged', 'less commonly taught', or 'low-density' (Magueresse et al., 2020). Indigenous languages are largely included under this umbrella term, and they represent a unique challenge when dealing with NLP tasks.

First, most of the Indigenous languages worldwide are generally understudied, which means that even though we can grasp some of their general grammatical features based on other previously studied languages from the same linguistic families, there are still particular traits which haven't been described. Second, Indigenous languages are typologically different: some of them are polysynthetic, such as the languages belonging to Uto-Aztecan family (e.g. Nahuatl, Wixarika) with rich morphophonemics and a large number of inflections (Mithun, 2001). Other languages are highly analytic with simpler morphology, but with complex tonal systems such as Chatino and Chinantec, from the Oto-Manguean family. Due to the lack of prior study, it becomes challenging to even define what constitutes a language versus a language variety among Indigenous languages.

Finally, another major challenge is the diversification of orthographies and the scarcity of written corpora in such languages. However, in lieu of these challenges, there has been a substantial increase in NLP applications for Indigenous languages (Mohanty et al., 2023). For example, Hedderich et al. (2020) survey common methods used in low-resource scenarios, such as data augmentation, distant supervision, and cross-lingual language models. Mager et al. (2018) provide an overview of research in NLP related to the Indigenous languages of the Americas, with an accompanying, and continually-updated,repository of research works and other resources for Indigenous languages. Recently, ACL 2022 featured a theme track on *Language Diversity: from Low-Resource to Endangered Languages*, which highlights papers

Rank	Team	Version	Count	Tot. ChrF	Tot. Bleu	Avg. BLEU	Avg. ChrF	Avg. BLEU ALL	Avg. ChrF ALL
1	Sheffield	1	11	335.04	61.29	5.57	30.46	5.57	30.46
2	Sheffield	2	11	333.57	60.35	5.49	30.32	5.49	30.32
3	Sheffield	3	11	325.51	57.59	5.24	29.59	5.24	29.59
4	Helsinki-NLP	6	11	317.09	56.17	5.11	28.83	5.11	28.83
5	Helsinki-NLP	2	11	284.11	43.60	3.96	25.83	3.96	25.83
6	Helsinki-NLP	3	11	283.62	40.57	3.69	25.78	3.69	25.78
7	Helsinki-NLP	4	11	283.09	47.19	4.29	25.74	4.29	25.74
8	Helsinki-NLP	1	11	277.71	44.22	4.02	25.25	4.02	25.25
10	LTLAmsterdam	3	11	261.83	35.53	3.23	23.80	3.23	23.80
11	PlayGround	1	10	249.71	30.52	3.05	24.97	2.77	22.70
12	CIC-NLP	2	11	222.50	17.38	1.58	20.23	1.58	20.23
13	CIC-NLP	1	11	207.80	18.49	1.68	18.89	1.68	18.89
14	Helsinki-NLP	5	11	205.96	15.63	1.42	18.72	1.42	18.72
15	CIC-NLP	3	11	197.69	14.46	1.31	17.97	1.31	17.97
16	LTLAmsterdam	2	11	171.11	18.70	1.70	15.56	1.70	15.56
17	LTLAmsterdam	1	10	160.42	12.68	1.27	16.04	1.15	14.58
18	LCT-EHU	3	1	38.59	3.45	3.45	38.59	0.31	3.51
19	LCT-EHU	1	1	38.40	3.08	3.08	38.40	0.28	3.49
20	LCT-EHU	2	1	38.21	3.11	3.11	38.21	0.28	3.47
21	LCT-EHU	4	1	37.71	3.47	3.47	37.71	0.32	3.43
22	LCT-EHU	5	1	37.26	3.06	3.06	37.26	0.28	3.39
23	Andes	1	1	9.22	0.12	0.12	9.22	0.01	0.84

Table 3: Ranking of the submissions to the shared task. For each team and submission version, Count represents the number of languages supported with Tot. ChrF and Tot. BLEU representing the sum ChrF and BLEU scores over all supported languages by a submission. While Avg. BLEU and Avg. ChrF represent the average of all supported languages by a submission, the Avg*ALL columns represent the average over all 11 shared task languages, with Avg. ChrF ALL determining the final ranking of the submissions.

focusing on Indigenous languages, and featured a keynote discussion on how to best support linguistic diversity (Muresan et al., 2022).

2.2 Low-Resource MT

Low-Resource MT (LRMT) tackles the challenge of developing translation systems for language pairs with limited parallel data. Traditional neural machine translation approaches struggle in such scenarios due to data scarcity.

Multilingual transfer learning has been successful in enhancing translation quality in LRMT by leveraging knowledge from related languages (Zoph et al., 2016; Nguyen and Chiang, 2017; Aharoni et al., 2019). By utilizing shared representations across languages, multilingual models can generalize well to unseen language pairs with limited data.

One effective LRMT approach using transfer learning is finetuning large multilingual language models on specific language pairs. This involves adapting pretrained models like mBART, M2M-100, and NLLB-200 to target specific language pairs or domains of interest (Liu et al., 2020; Fan et al., 2020; Team et al., 2022). Refining the model's parameters through this technique enhances translation quality for low-resource languages (Thillainathan et al., 2021; Liu et al., 2020).

Back-translation is another effective technique employed in LRMT, which generates synthetic parallel data by translating and re-translating monolingual data (Sennrich et al., 2016; Feldman and Coto-Solano, 2020; Lample et al., 2018). By incorporating this technique, LRMT systems can benefit from additional training examples, leading to improved translation performance.

3 Task and Evaluation

The shared task focuses on *open* machine translation: outside of the development set and any prohibited datasets, teams are allowed to collect and train on an unlimited amount of external data. As translation performance for low-resource Indigenous languages is generally low, we choose this setting to allow models to achieve the best possible performance, in hopes that usable translation models become more quickly developed.

Metrics Translation evaluation is done with ChrF (Popović, 2015), as implemented in SCAREBLEU (Post, 2018), as the target languages are morphologically rich. While teams are not required to submit a system for all languages, the final score for each submission is calculated by taking an average over all eleven languages; if there is no model output for a given language, the score is taken as 0.

4 Languages and Data

For development and evaluation, the AmericasNLP 2021 shared task used multi-way parallel translations of the Spanish XNLI test set across 10 languages: Asháninka, Aymara, Bribri, Guarani, Nahuatl, Otomí, Quechua, Rarámuri, Shipibo-Konibo and Wixarika (Ebrahimi et al., 2022). For this edition of the shared task, we use the same evaluation set and additionally introduce a new evaluation dataset, created from Mexican court proceedings, for Spanish–Chatino. This set was released as a surprise language near the end of the competition, along with a small amount of Spanish–Chatino and English–Chatino data for training. In this section, we describe the Chatino language, Spanish source data, and translation process. For a detailed overview of the ten other evaluation languages, we refer the reader to Ebrahimi et al. (2022) and Mager et al. (2021).

4.1 Chatino

San Juan Quiahije Chatino (SJQ, ISO 639-3 ctp), spoken by about 5000 people, is an Oto-Manguean language spoken in Oaxaca, Mexico and by Chatinos who live in many cities throughout the United States, with a high concentration in the Southeastern United States in the states of North Carolina, Alabama, and Georgia. The Chatino languages are some of the most complex tonal languages in the world. SJQ has 10 tonemes and 15 morphological tonal categories. In the created corpus, tones are represented as superscripts.

4.2 Evaluation Dataset

Source Data A main motivation for this dataset is to create a resource which could be more directly applicable to the real life needs of the communities involved, while at the same time limiting negative ethical implications (Mager et al., 2023). As such, we choose to use legal text as the source domain. The Mexican Constitution and the General Law of Linguistic Rights of Indigenous Peoples (*Ley General De Derechos Lingüísticos de los Pueblos Indígenas*[1]) states that the 68 Indigenous languages spoken in the country before the Spanish conquest are National Languages. This gives all people the right to perform bureaucratic and legal actions in their native language. As a first approximation of this text, we gather press releases from the Mexican Supreme Court.[2] This allows us to avoid the potential harms of directly generating low-quality translations of written laws and court decisions, while still allowing for insights into the issues and challenges of translating legal terms and text. Furthermore, the text generated by the Mexican Supreme Court is public domain, allowing for free usage.

Translation Process To create the dataset, we crawl 10,000 instances from the Supreme Court press releases, and randomly select a subset for translation. Translations are jointly done by two professional translators, who are native San Juan Quiahije Chatino speakers. Legal terms in Spanish are translated into Chatino, in order to reduce code-switching and borrowed words. This translation of domain-specific terms represents the most challenging aspect of the translation process, with translators investigating the context and meaning of specific words in order to create accurate translations. For more difficult cases, translators consulted with lawyers to clarify the meaning of certain texts. For all translations, both translators worked together to reach an agreement on the translated text. Examples of difficult to translate words and entities include "dismissal, approval, jurisprudence, regulations among others and Chamber of Deputies, the nation's Supreme Court of Justice and Magistrate."

5 Baseline and Submitted Systems

In this section, we describe the 2023 baseline system and each team's approach. We present a summary of all approaches in Table 2.

5.1 Baseline

The AmericasNLP 2021 shared task used a transformer encoder–decoder model (Vaswani et al., 2017) along with hyperparameters shown to work well for low-resource settings (Guzmán et al., 2019). For this year's edition of the shared task, we use the winning 2021 system (Vázquez et al., 2021) as the baseline, as it greatly outperformed the previous baseline and other submissions on all languages.

5.2 Andes

The Andes team (Gillin and Gummibaerhausen, 2023) submitted a translation system for Spanish–Aymara. The system is based on mT5 (Xue et al.,

[1] https://www.diputados.gob.mx/LeyesBiblio/pdf/LGDLPI.pdf

[2] https://www.scjn.gob.mx/multimedia/comunicados

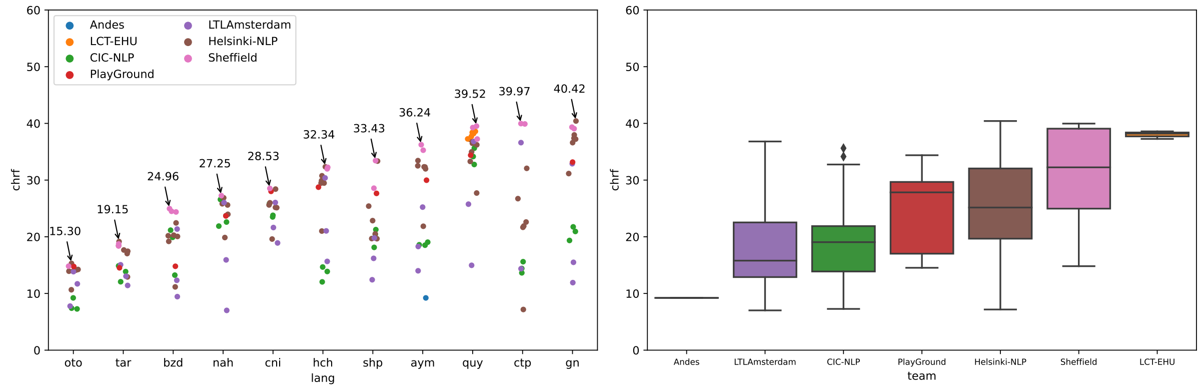

Figure 1: Main results of the shared task, in ChrF. In the left chart, we plot the performance of every submission, for each language. On the right, we show the distribution of per-team performance, across all submissions and languages. We note that distributions may not be directly comparable depending on the number of submissions from each team.

2021) and is further finetuned on English–Aymara data, in addition to the provided Spanish–Aymara data. The English parallel data consists of a lexicon, collected from books meant for language learning (Wexler and Programs, 1967; Parker, 2008)

5.3 CIC-NLP

The CIC-NLP team (Tonja et al., 2023) submitted three different models across all languages, based on either mBART50 (Tang et al., 2021) and M2M100 (Fan et al., 2020) or a publicly released English–Spanish translation model.[3] The multilingual models were first optionally finetuned on a concatenation of the es-XX training data across all languages. Language-specific models were then created by further finetuning on data for a specific target language. The English–Spanish model was only finetuned on data for a specific language pair.

5.4 Helsinki-NLP

The Helsinki-NLP team (Vázquez et al., 2023) submitted six different models across all languages, following four main modeling approaches. Model B is a copy of the team's winning multilingual one-to-many 2021 model, and Model C is a re-implementation of this approach using OpusTrainer and a language specific-finetuning step. Model A focuses on knowledge distillation and transfer learning: a parent English–Spanish model is distilled from the NLLB model, and is then further finetuned on target-language data. Model D uses language-specific decoders as part of a modular architecture: a specified number of decoder layers are shared across languages, while others are trained separately per language. The team also focused heavily on data collection and cleaning. In addition to the data provided by the shared task, the team collected data from OPUS (Tiedemann, 2012), the FLORES-200 (Team et al., 2022) evaluation sets, the Bible (McCarthy et al., 2020), the Universal Declaration of Human Rights, and various texts extracted from websites or PDFs of educational materials and news. MT was also used to leverage monolingual Wikipedia data as well as parallel data between the target languages and English. Texts were detokenized and whitespace normalized if necessary. Data from all sources was concatenated and deduplicated to create the final training data, and special tags denoting the quality and language variety of the source material were added to each example.

5.5 LCT-EHU

The LCT-EHU team (Ahmed et al., 2023) focused on the Spanish–Quechua language pair and submitted five different models to the competition. Among their contributions, they collected new parallel corpora, experimented with high-resource bilingual systems as pretrained models, such as Spanish–English and Spanish–Finnish, and generated synthetic parallel data from monolingual texts using back-translation and the copied corpus technique (Currey et al., 2017). The best result on the test set was obtained by using a model pretrained on Spanish–Finnish and by including new parallel data from the literature and legal domains, despite originating from different variants of Quechua Ayacucho.

[3] https:huggingface.co/Helsinki-NLP/opus-mt-es-en

Team	AYM	BZD	CNI	CTP	GN	HCH	NAH	OTO	QUY	SHP	TAR
2021 Baseline	15.70	6.80	10.20	-	19.30	12.60	15.70	5.40	30.40	12.10	3.90
2021 Best	28.30	16.50	25.80	-	33.60	30.40	26.60	14.70	34.30	32.90	18.40
Andes	9.22	-	-	-	-	-	-	-	-	-	-
CIC-NLP	19.05	21.17	25.85	15.61	21.75	14.67	26.57	9.22	35.62	21.26	14.87
Helsinki-NLP	33.44	22.45	28.41	32.07	**40.42**	**32.34**	26.87	**15.30**	37.19	33.35	**19.15**
LCT-EHU	-	-	-	-	-	-	-	-	38.59	-	-
LTLAmsterdam	25.23	21.36	26.04	36.61	32.89	30.38	26.03	13.85	36.81	19.8	15.06
PlayGround	29.98	14.80	28.01	-	33.17	28.75	23.68	14.75	34.38	27.66	14.53
Sheffield	**36.24**	**24.96**	**28.53**	**39.97**	39.34	32.25	**27.25**	14.81	**39.52**	**33.43**	18.74
↑ 2021	12.60	9.70	15.60	-	14.30	17.80	10.90	9.30	3.90	20.80	14.50
↑ 2023	7.94	8.46	2.73	-	6.82	1.94	0.73	0.60	5.22	0.53	0.75

Table 4: Summary of best performing submission from each team per language. Note that values can come from multiple submissions, making these scores different than what is used to calculate the overall shared task ranking. ↑2021 marks the difference between the 2021 Baseline and 2021 winning system. ↑2023 marks the difference between the 2021 best (i.e., 2023 baseline) system and the best 2023 system.

5.6 LTLAmsterdam

The LTLAmsterdam team (Stap and Araabi, 2023) submitted four different models for all language pairs. Their approaches included a bilingual system, an off-the-shelf commercial large language model used for translation, and a finetuned multilingual model with additional adaptation. The bilingual systems were trained using transformer models with parameters specifically tailored for low-resource languages (Araabi and Monz, 2020). For the large language model, they utilized the ChatGPT API[4] and followed the prompts proposed by Jiao et al. (2023). Additionally, they finetuned the M2M100 multilingual model (Fan et al., 2021), specifically choosing the 418M parameter version and training a model for each language pair. It is important to highlight that none of the target languages in the shared task were originally included in the set of languages of M2M100. Finally, they augmented the finetuned M2M100 model with a k-nearest neighbor (kNN) datastore for inference (Khandelwal et al., 2021), effectively creating a semi-parametric model that combines the parametric M2M100 model with a nearest neighbor retrieval mechanism.

5.7 PlayGround

The PlayGround team (Gu et al., 2023) submitted one model for each language pair, except for Spanish–Chatino. Their approach focused on utilizing the pretrained NLBB-200 model (Team et al., 2022), which they finetuned using the available monolingual and parallel data for the shared task. They conducted a comparison between bilingual and multilingual finetuned models, incorporating back-translated data through finetuning the NLBB-200 model with Spanish as the target language. Additionally, they adopted a weight-averaging approach (Wortsman et al., 2022).

5.8 Sheffield

The Sheffield team (Gow-Smith and Villegas) submitted three models for all languages. Approaches were based off various versions of the NLLB-200 model (Team et al., 2022). In addition to the provided training data, the team used data from teams which participated in prior editions of the shared task (Moreno, 2021; Vázquez et al., 2021). Data from other sources, such as the Bible (McCarthy et al., 2020) and NLLB project were also considered, however the authors found that Bible data did not improve performance on the development set, and did not include it in the final systems. Back-translation was also used to create additional parallel data. The submissions include specific preprocessing steps to prepare the data, such as detokenization and replacement of tone markings for Chatino. The team experimented with the distilled 600M, 1.3B and 3.3B versions of NLLB, and models were first finetuned on a concatenation of all available training data. The checkpoint with best average ChrF across all languages was considered as Submission 3. For Submission 2, the best check-

[4] https://platform.openai.com/docs/api-reference/chat

point per language was used. Submission 1 consists of ensembles of the various NLLB models. As NLLB relies on specific tags to denote the target languages, the embedding matrix was extended and new languages tags were created for the shared task languages which are unsupported.

6 Results

We present the overall ranking of submissions to the shared task in Table 3 and the best score per language for each team across all submissions in Table 4.

The overall winner of the shared task, the Sheffield Submission 1, achieves the best performance for 7 languages: Aymara, Bribri, Asháninka, Chatino, Nahuatl, Quechua, and Shipibo-Konibo. The Helsinki Submission 6 (i.e., Model B) has the highest performance for 4 languages: Guarani, Wixarika, Otomí, and Rarámuri. Systems are much more competitive than prior competitions, achieving extremely close ChrF scores for many languages, such as Asháninka, Guarani, Wixarika, and Shipibo-Konibo. The Sheffield and Helsinki teams both collect additional data, and train models in a multilingual and multi-stage fashion. Both also mention data cleaning and preprocessing in their pipeline, and we hypothesize that this step is likely vital for good performance, due to noise, domain mismatch, and differences in variants between the training and evaluation sets. For all languages except for Aymara, all teams have at least one submission which improves (often by a large margin) over the original 2021 baseline.

Comparison with Prior Years As the evaluation set for 10 of the languages is the same as for 2021, we can analyze the performance of submitted MT systems over time. In this year's shared task, we see improvements over the best 2021 system, the 2021 Helsinki submission (Vázquez et al., 2021), for all languages, but to varying degree. The largest improvements are for Bribri, Aymara, Guarani and Quechua. We also see small improvements for Asháninka and Wixarika. However, improvements for Nahuatl, Otomí, and Shipibo-Konibo are marginal. Overall, the improvements over Vázquez et al. (2021) are smaller in magnitude, compared to the improvements in 2021. This can be expected, however, as the baseline for this year's shared task represents a much stronger lower bound. Of the four languages with largest improvement, three are achieved by a Sheffield submission: Aymara, Bribri, and Quechua. This may be attributed, in part, to the use of the NLLB model by the team, which supports Aymara and Quechua in its original set of pretraining languages. On average across the 10 shared languages, we see a further 9.63 improvement in ChrF over 2021 results by the best submitted systems.

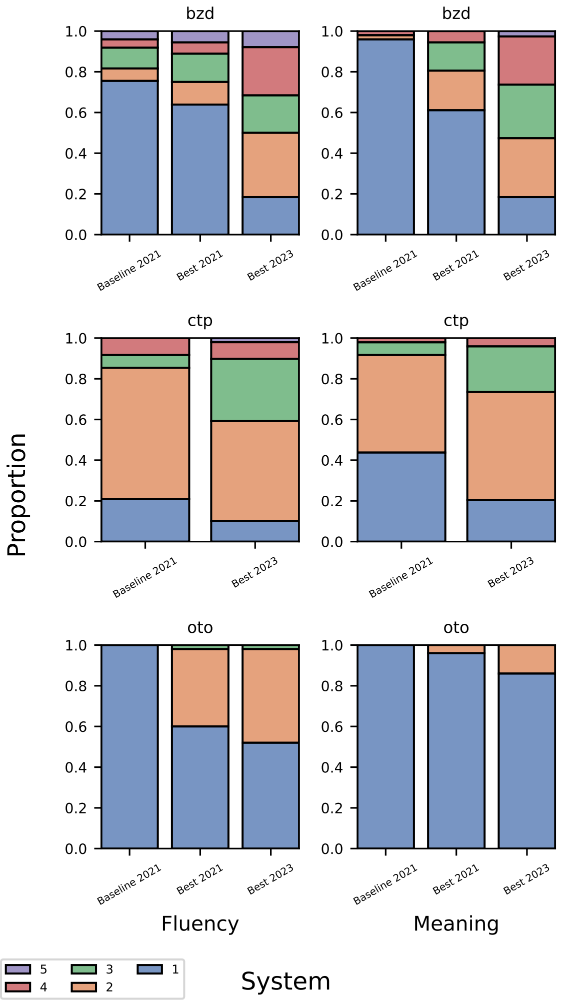

Figure 2: Results of the qualitative human evaluation. Ratings of *fluency* are displayed in the left column, and *meaning* in the right. Results are shown as a proportion of all evaluated sentences.

7 Additional Experiments

7.1 Qualitative Analysis

As quantitative measures of translation performance do not paint a complete picture, we also conduct a qualitative analysis of the system outputs for Bribri, Chatino, and Otomí. We randomly sample 50 parallel examples across the 2021 baseline, the 2021 winning system (Vázquez et al., 2021), and the 2023 submission with best performance for each language: Sheffield Submission 1 for Bribri and Chatino, and Helsinki Submission 6 for Otomí. Examples are shuffled and presented to a native speaker of each language, along with the Spanish source and gold reference. Annotations are done across two dimensions: *meaning* and *fluency*, using a categorical 1-5 scale. The guidelines given to annotators can be found in Appendix A.1.

The results of this analysis are shown in Figure 2. Similar to the trend of improvement in ChrF, we also see improvements in the rating of meaning and fluency across the three systems in this analysis. For Bribri, a strong majority of translations from the original 2021 baseline has a score of 1 across both dimensions. While we see some improvements from the Helsinki 2021 system, the 2023 system provides a considerable increase in translation quality; ratings of between 2-4 are now assigned to the majority of examples. For Chatino, the baseline system is stronger than for Bribri, and the improvement between the two systems is smaller when considering the proportion of examples rated as 1. For the 2023 system, we see the largest increase in quantity for ratings of 3. Otomí sees the worst performance of the three languages, with the majority of examples being rated as 1, across all three systems. Fluency does improve slightly, with an increase in the number of 2 ratings. However, examples with higher ratings are effectively non-existent. We also see a difference in improvement across *fluency* and *meaning*, with the former showing higher improvement. For all languages, even if we see an increase in the proportion of higher rated examples, the number of near-perfect (i.e., rating of 5) remains consistently small.

7.2 Impact of In-domain Data

The LTLAmsterdam team (Stap and Araabi, 2023) describes systems which make use of kNN and an external data store (Khandelwal et al., 2021) during decoding. It was jointly decided in a discussion between the organizers and team that submissions which use this approach – Submissions 4,5,6,7, and 8 – fall in a grey area with respect to the competition rules and would not be included in the main results, due to the fact that development set examples were included in the data store. However, these submissions can give insights into the potential improvements one can expect if there is access to parallel examples which are in-domain with respect to an expected test set. If we consider these submissions, they achieve the best performance for three languages: Bribri, Asháninka, and Nahuatl. Improvements over the next best team submission is 0.88 ChrF on average over the three languages. As such, given that systems still struggle with producing outputs with the highest qualitative rating (§7.1), this approach may be beneficial for producing more constrained and higher-quality outputs, given that access to high-quality parallel data is available.

8 Conclusion

In this paper we present the results of the AmericasNLP 2023 shared task. For this iteration, we collect a new dataset for translation evaluation between Spanish and Chatino, consisting of legal text from court press releases. Additionally, we keep the prior 10 evaluation languages used in 2021. Overall, 7 teams participated in the shared task. For all languages, multiple submissions improve over the previous best ChrF, but the magnitude varies per language. The best results were achieved by either finetuned versions of NLLB or a from-scratch transformer encoder–decoder model. To confirm the improvement in ChrF from the previous shared task, we conduct a human evaluation of system outputs, which, although it supports the quantitative improvement, highlights the fact that systems are still not able to produce translations of the highest quality. Furthermore, there is still variability in the absolute performance across languages. As such, while the results of the shared task mark a promising trend in increasing translation quality for Indigenous languages, there are still improvements which can be made in order to create usable translation systems for Indigenous languages.

Acknowledgments

We would like to thank all teams that submitted systems to this year's AmericasNLP shared task! We would also like to thank Eric Ramos Aguilar for their help with the qualitative annotation of system outputs.

References

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.

Nouman Ahmed, Natalia Flechas Manrique, and Antonije Petrović. 2023. Enhancing Spanish-Quechua Machine Translation with Pre-Trained Models and Diverse Data Sources: LCT-EHU at AmericasNLP Shared Task. In *"Proceedings of the Third Workshop on Natural Language Processing for Indigenous Languages of the Americas"*. Association for Computational Linguistics.

Antonios Anastasopoulos. 2019. *Computational tools for endangered language documentation*. University of Notre Dame.

Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating Zero-shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-resource Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation.

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*, page 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Candace Kaleimamoowahinekapu Galla. 2016. Indigenous language revitalization, promotion, and education: Function of digital technology. *Computer Assisted Language Learning*, 29(7):1137–1151.

Nat Gillin and Brian Gummibaerhausen. 2023. Few-shot Spanish-Aymara Machine Translation Using English-Aymara Lexicon.

Edward Gow-Smith and Danae Sánchez Villegas. Sheffield's Submission to the AmericasNLP Shared Task on Machine Translation into Indigenous Languages.

Tianrui Gu, Kaie Chen, Siqi Ouyang, and Lei Li. 2023. PlayGround Low Resource Machine Translation System for the 2023 AmericasNLP Shared Task. In *"Proceedings of the Third Workshop on Natural Language Processing for Indigenous Languages of the Americas"*. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based &amp neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

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.

Marianne Mithun. 2001. *The languages of native North America*. Cambridge University Press.

Sushree Mohanty, Shantipriya Parida, and Satya Dash. 2023. Role of nlp for corpus development of endangered languages.

Oscar Moreno. 2021. The REPU CS' Spanish–Quechua Submission to the AmericasNLP 2021 Shared Task on Open Machine Translation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 241–247, Online. Association for Computational Linguistics.

Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors. 2022. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Philip M. Parker. 2008. *Webster's Aymara - English Thesaurus Dictionary*. ICON Group International, Inc.

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.

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.

Shruti Rijhwani. 2023. Improving Optical Character Recognition for Endangered Languages.

Lane Schwartz, Francis Tyers, Lori Levin, Christo Kirov, Patrick Littell, Chi-kiu Lo, Emily Prud'hommeaux, Hyunji Hayley Park, Kenneth Steimel, Rebecca Knowles, et al. 2020. Neural polysynthetic language modelling. *arXiv preprint arXiv:2005.05477*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data.

David Stap and Ali Araabi. 2023. Chatgpt is not a good indigenous translator. In *"Proceedings of the Third Workshop on Natural Language Processing for Indigenous Languages of the Americas"*. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual Translation from Denoising Pre-Training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. 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.

Sarubi Thillainathan, Surangika Ranathunga, and Sanath Jayasena. 2021. Fine-tuning self-supervised multilingual sequence-to-sequence models for extremely low-resource nmt. In *2021 Moratuwa Engineering Research Conference (MERCon)*, page 432–437.

Jorg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS.

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.

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

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online. Association for Computational Linguistics.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2023. The Helsinki submission to the AmericasNLP shared task. pages 255–264. Association for Computational Linguistics.

Paul Wexler and Washington State University Peace Corps Training Programs. 1967. *Beginning Aymara: A Course for English Speakers*. The Programs.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

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*, page 1568–1575, Austin, Texas. Association for Computational Linguistics.

Appendix

Lang.	Team	Ver.	ChrF	BLEU
aym	Sheffield	1	36.24	4.45
aym	Sheffield	3	35.27	4.03
aym	Helsinki-NLP	6	33.44	3.37
aym	Helsinki-NLP	4	32.52	3.15
aym	Helsinki-NLP	3	32.34	3.04
aym	Helsinki-NLP	1	32.31	3.30
aym	Helsinki-NLP	2	31.98	2.44
aym	PlayGround	1	29.98	1.96
aym	LTLAmsterdam	3	25.23	1.68
aym	Helsinki-NLP	5	21.86	1.10
aym	CIC-NLP	1	19.05	1.13
aym	CIC-NLP	3	18.59	0.56
aym	CIC-NLP	2	18.52	0.84
aym	LTLAmsterdam	1	18.28	0.96
aym	LTLAmsterdam	2	14.00	0.09
aym	Andes	1	9.22	0.12
bzd	Sheffield	1	24.96	6.35
bzd	Sheffield	3	24.49	6.21
bzd	Sheffield	2	24.38	6.18
bzd	Helsinki-NLP	6	22.45	5.64
bzd	LTLAmsterdam	3	21.36	5.23
bzd	CIC-NLP	2	21.17	4.72
bzd	Helsinki-NLP	4	20.28	5.02
bzd	Helsinki-NLP	1	20.18	4.66
bzd	Helsinki-NLP	3	20.06	4.44
bzd	CIC-NLP	1	19.90	3.92
bzd	Helsinki-NLP	2	19.19	4.36
bzd	PlayGround	1	14.80	2.04
bzd	CIC-NLP	3	13.24	1.66
bzd	LTLAmsterdam	2	12.32	0.97
bzd	Helsinki-NLP	5	11.16	1.10
bzd	LTLAmsterdam	1	9.44	1.38
cni	Sheffield	1	28.53	3.23
cni	Helsinki-NLP	6	28.41	4.45
cni	PlayGround	1	28.01	3.53
cni	LTLAmsterdam	3	26.04	3.03
cni	Helsinki-NLP	2	25.99	3.39
cni	CIC-NLP	2	25.85	2.72
cni	Helsinki-NLP	3	25.62	2.31
cni	Helsinki-NLP	1	25.18	3.40
cni	Helsinki-NLP	4	25.14	3.44
cni	CIC-NLP	3	23.79	3.28
cni	CIC-NLP	1	23.50	2.84
cni	LTLAmsterdam	2	21.63	0.59
cni	Helsinki-NLP	5	19.60	0.13
cni	LTLAmsterdam	1	18.91	2.35
ctp	Sheffield	1	39.97	12.33
ctp	Sheffield	3	39.90	12.26
ctp	LTLAmsterdam	2	36.61	8.45
ctp	Helsinki-NLP	6	32.07	8.59
ctp	Helsinki-NLP	3	26.73	3.75
ctp	Helsinki-NLP	4	22.61	4.01
ctp	Helsinki-NLP	1	21.89	3.49
ctp	Helsinki-NLP	2	21.67	3.73
ctp	CIC-NLP	2	15.61	1.20
ctp	CIC-NLP	1	14.41	1.09
ctp	LTLAmsterdam	3	14.37	0.98
ctp	CIC-NLP	3	13.64	0.87
ctp	Helsinki-NLP	5	7.17	0.00
gn	Helsinki-NLP	6	40.42	8.40
gn	Sheffield	1	39.34	6.96
gn	Sheffield	3	39.07	7.18
gn	Helsinki-NLP	4	37.97	7.99
gn	Helsinki-NLP	3	37.38	7.49
gn	Helsinki-NLP	1	37.23	7.55
gn	Helsinki-NLP	2	36.60	6.90
gn	PlayGround	1	33.17	5.56
gn	LTLAmsterdam	3	32.89	5.43
gn	Helsinki-NLP	5	31.15	4.69
gn	CIC-NLP	2	21.75	1.84
gn	CIC-NLP	3	20.94	1.54
gn	CIC-NLP	1	19.35	1.34
gn	LTLAmsterdam	1	15.50	1.21
gn	LTLAmsterdam	2	11.91	0.10
hch	Helsinki-NLP	6	32.34	11.49
hch	Sheffield	1	32.25	12.04
hch	Sheffield	2	31.98	11.43
hch	Helsinki-NLP	3	30.76	10.98
hch	LTLAmsterdam	3	30.38	11.56
hch	Helsinki-NLP	4	29.90	12.59
hch	Helsinki-NLP	2	29.48	11.30
hch	Helsinki-NLP	1	29.47	12.30
hch	PlayGround	1	28.75	9.90
hch	LTLAmsterdam	2	21.04	7.69
hch	Helsinki-NLP	5	21.01	6.24
hch	LTLAmsterdam	1	15.66	0.71
hch	CIC-NLP	3	14.67	1.46
hch	CIC-NLP	2	13.88	0.08
hch	CIC-NLP	1	12.05	1.58
nah	Sheffield	1	27.25	2.33
nah	Helsinki-NLP	6	26.87	2.05
nah	CIC-NLP	2	26.57	1.36
nah	LTLAmsterdam	3	26.03	1.33
nah	Helsinki-NLP	4	25.82	1.75
nah	Helsinki-NLP	2	25.61	2.00
nah	Helsinki-NLP	1	23.96	1.41
nah	Helsinki-NLP	3	23.72	1.75
nah	PlayGround	1	23.68	0.90

Lang.	Team	Ver.	ChrF	BLEU
nah	CIC-NLP	3	22.60	1.22
nah	CIC-NLP	1	21.88	1.07
nah	Helsinki-NLP	5	19.87	0.14
nah	LTLAmsterdam	1	15.93	0.96
nah	LTLAmsterdam	2	7.02	0.03
oto	Helsinki-NLP	6	15.30	1.95
oto	Sheffield	1	14.81	1.71
oto	PlayGround	1	14.75	1.07
oto	Helsinki-NLP	2	14.23	1.45
oto	Helsinki-NLP	4	14.11	1.51
oto	Helsinki-NLP	1	13.93	1.41
oto	Helsinki-NLP	3	13.92	1.43
oto	LTLAmsterdam	3	13.85	1.25
oto	LTLAmsterdam	1	11.70	1.34
oto	Helsinki-NLP	5	10.66	0.12
oto	CIC-NLP	1	9.22	0.26
oto	LTLAmsterdam	2	7.77	0.02
oto	CIC-NLP	2	7.40	0.07
oto	CIC-NLP	3	7.28	0.05
quy	Sheffield	1	39.52	4.61
quy	Sheffield	2	39.26	4.54
quy	LCT-EHU	3	38.59	3.45
quy	LCT-EHU	1	38.40	3.08
quy	LCT-EHU	2	38.21	3.11
quy	LCT-EHU	4	37.71	3.47
quy	LCT-EHU	5	37.26	3.06
quy	Sheffield	3	37.24	4.33
quy	Helsinki-NLP	4	37.19	4.28
quy	LTLAmsterdam	3	36.81	3.00
quy	Helsinki-NLP	2	36.49	3.77
quy	Helsinki-NLP	1	36.22	3.49
quy	CIC-NLP	2	35.62	2.55
quy	Helsinki-NLP	3	34.97	2.74
quy	PlayGround	1	34.38	2.53
quy	CIC-NLP	1	34.15	2.59
quy	Helsinki-NLP	6	33.29	2.99
quy	CIC-NLP	3	32.75	2.05
quy	Helsinki-NLP	5	27.72	0.91
quy	LTLAmsterdam	1	25.75	1.47
quy	LTLAmsterdam	2	14.97	0.33
shp	Sheffield	1	33.43	6.32
shp	Helsinki-NLP	6	33.35	6.10
shp	Sheffield	3	28.57	4.00
shp	PlayGround	1	27.66	2.81
shp	Helsinki-NLP	2	25.41	3.13
shp	Helsinki-NLP	5	22.85	1.05
shp	CIC-NLP	2	21.26	1.83
shp	Helsinki-NLP	4	20.51	2.25
shp	CIC-NLP	1	20.43	2.28
shp	LTLAmsterdam	3	19.80	1.83
shp	Helsinki-NLP	3	19.68	2.04
shp	Helsinki-NLP	1	19.66	2.03
shp	CIC-NLP	3	18.13	1.66
shp	LTLAmsterdam	1	16.20	1.59
shp	LTLAmsterdam	2	12.42	0.34
tar	Helsinki-NLP	6	19.15	1.16
tar	Sheffield	1	18.74	0.95
tar	Helsinki-NLP	3	18.43	0.60
tar	Sheffield	2	18.39	0.88
tar	Helsinki-NLP	1	17.67	1.18
tar	Helsinki-NLP	2	17.45	1.13
tar	Helsinki-NLP	4	17.04	1.21
tar	LTLAmsterdam	3	15.06	0.22
tar	CIC-NLP	2	14.87	0.17
tar	PlayGround	1	14.53	0.23
tar	CIC-NLP	1	13.86	0.38
tar	LTLAmsterdam	1	13.04	0.72
tar	Helsinki-NLP	5	12.92	0.14
tar	CIC-NLP	3	12.07	0.09
tar	LTLAmsterdam	2	11.42	0.09

Table 5: Main results of the AmericasNLP 2023 shared task.

A Annotation and Table Guidelines

A.1 Human Evaluation Guidelines

Annotators were given the following guidelines for their evaluation:

Fluency: Is the output sentence easily readable and similar to a human-produced text?

1. *Extremely bad*: The output contains mainly repetitions or hallucinations [> 80%], and is largely illegible. The text is clearly not produced by a human.
2. *Bad*: The output may contain repetitions or erroneous characters [> 60%], but also some correct words or phrases.
3. *Acceptable*: The output does not contain a significant number of repetitions, and mainly contains correct words, however may still have grammatical errors.
4. *Sufficiently good*: The output seems like a human-produced text in the target language, without repetitions or erroneous characters, but may still contain some grammatical errors.
5. *Excellent*: The output seems like a human produced text in the target language, and is readable without issues.

Meaning: How well does the translation reflect the meaning of the reference?

1. *Extremely bad*: The meaning of the source sentence can not be inferred at all.
2. *Bad*: A small number of words or phrases allow the reader to guess the meaning or semantic content of the sentence
3. *Acceptable*: A larger number of correctly translated phrases and words allow a stronger understanding of the meaning.
4. *Sufficiently good*: The general meaning of the source sentence is conveyed, while some details may be missing.
5. *Excellent*: The meaning of the source sentence, along with all relevant details, is conveyed completely.

A.2 Guidelines for System Summary

Data

- Crawl: Does the team collect additional data from websites, PDFs, documents, books, etc.
- External Bilingual: Does the team leverage existing parallel data for language pairs not used for evaluation?
- Opus/Religious/Wikipedia: Does the team use additional data from the respective resource?
- Prior Year: Does the team use data collected from the 2021 or 2022 Shared Tasks?
- Monolingual Translation: Does the team create synthetic training data by translating a monolingual dataset?
- Pivot Translation: Does the team leverage exiting parallel data, between an unsupported language pair, through translation?
- Cleaning/Normalization: Does the team specifically describe any cleaning or normalization steps?
- No Additional: Does the team solely use the data provided from the competition?

Pretraining: A check is given if the team describes a submission which uses one of the pretrained systems. Encoder-Decoder represents a vanilla encoder-decoder transformer model trained from scratch.

Train

- Ensemble: Does the team describe a submission which makes use of multiple models for translation?
- Multistage: Does the team describe the training procedure as multiple stages, with variations in hyperparameters or training data?
- Multilingual: Does the team describe the training as multilingual, or create models which are trained on multiple language pairs?