Sheffield's Submission to the AmericasNLP Shared Task on Machine Translation into Indigenous Languages

Edward Gow-Smith, Danae Sánchez Villegas

Computer Science Department, University of Sheffield, UK

{egow-smith1, dsanchezvillegas1}@sheffield.ac.uk

Abstract

In this paper we describe the University of Sheffield's submission to the AmericasNLP 2023 Shared Task on Machine Translation into Indigenous Languages which comprises the translation from Spanish to eleven indigenous languages. Our approach consists of extending, training, and ensembling different variations of NLLB-200. We use data provided by the organizers and data from various other sources such as constitutions, handbooks, news articles, and backtranslations generated from monolingual data. On the dev set, our best submission outperforms the baseline by 11% average chrF across all languages, with substantial improvements particularly for Aymara, Guarani and Quechua. On the test set, we achieve the highest average chrF of all the submissions, we rank first in four of the eleven languages, and at least one of our submissions ranks in the top 3 for all languages.[1]

1 Introduction

The 2023 AmericasNLP Shared Task (Ebrahimi et al., 2023) involves developing machine translation systems for translating from Spanish to eleven low resource indigenous languages: Aymara (*aym*), Bribri (*bzd*), Asháninka (*cni*), Chatino (*czn*), Guarani (*gn*), Wixarika (*hch*), Nahuatl (*nah*), Hñähñu (*oto*), Quechua (*quy*), Shipibo-Konibo (*shp*), and Rarámuri (*tar*). Developing machine translation systems for these languages is challenging since many of them are polysynthetic (i.e., words are composed of several morphemes) and word boundaries are not standardized; they present different orthographic variations (e.g., classical vs. modern Nahuatl variations); presence of code-switching is common, among other difficulties of low resource settings.

Previous work has explored the effectiveness of pretrained machine translation models in low resource settings (Haddow et al., 2022) showing their impact on improving translation quality and addressing data scarcity challenges. Following this approach, our submissions to the 2023 AmericasNLP shared task consist of extending and finetuning various versions of NLLB-200 (Costa-jussà et al., 2022), a state-of-the-art machine translation model specifically designed for low resource settings. NLLB-200 is trained on 202 languages across 1 220 language pairs, including three of the languages present in the AmericasNLP shared task: *aym*, *gn*, and *quy*.[2] We further train our models on data from various sources such as constitutions and news articles, and we leverage multilingual training and ensembling to improve their performance. Models are evaluated using chrF (Popović, 2015), the official metric of the task. On the test set, we achieve the highest average chrF across all languages, and the best chrF for four of the languages.

The rest of the paper is organised as follows: Section 2 describes the data sources for training our models, Section 3 explains our three submissions in detail, Section 4 presents the results on the dev and test sets, Section 5 analyses the impact of different factors to the model's performance, Section 6 looks at zero-shot capabilities, and we draw conclusions in Section 7.

2 Data

2.1 Data Collection

We collect data from a variety of data sources, including training data provided by the organisers (AmericasNLP 2023), data from prior submissions to the AmericasNLP shared task (Helsinski and REPUcs) and relevant datasets specific to the in-

[1] We release code for training our models here: https://github.com/edwardgowsmith/americasnlp-2023-sheffield

[2] We present inference results on the dev set for these models in Table 4.

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 192–199
July 14, 2023

Language	AmericasNLP 2023	Helsinki	REPUcs	NLLB	Train Total	Backtranslations	Bibles
aym	15,586	149,225	10,729	8,809	173,620	16,750	154,520
bzd	7,508				7,508		38,502
cni	3,883				3,883	13,192	38,846
czn	3,118				3,118		
gn	26,032	1,713		7,906	33,938	40,515	39,457
hch	8,966	2,404			11,370	510	39,756
nah	16,145				19,993	8,703	39,772
oto	4,889	3,834			8,723	537	39,726
quy	542,914	3,634			557,277		154,825
shp	14,592	14,656			29,248	23,592	79,341
tar	14,721	3,856			18,577		39,444

Table 1: Amount of parallel data collected for each language. AmericasNLP 2023: parallel training data provided by the organizers, Helsinki: data taken from Vázquez et al. (2021), REPUcs: data taken from Moreno (2021), NLLB: data from Costa-jussà et al. (2022), Backtranslations: back-translations created from monolingual data, Bibles: data from The JHU Bible corpus (McCarthy et al., 2020).

digenous languages included in the task (NLLB). Table 1 shows the size of the training data for each language. The total amount of training data is unevenly distributed among datasets, with Quechua (557 277), Aymara (173 620), and Guarani (33 938) having the greatest amount of training data.

AmericasNLP 2023 Data provided by the organisers of the 2023 AmericasNLP Shared Task includes parallel datasets for training the eleven languages. Table 8 contains all datasets and references.

Helsinki We take data from OPUS (Tiedemann, 2012) and other sources (including constitutions) provided by the University of Helsinki's submission (Vázquez et al., 2021) to the AmericasNLP 2021 Shared Task (Mager et al., 2021). The collected data from constitutions includes translations of the Mexican constitution into Hñähñu, Nahuatl, Raramuri and Wixarika, of the Bolivian constitution into Aymara and Quechua, and of the Peruvian constitution into Quechua.

REPUcs We use data collected for the REPUcs' submission to the 2021 AmericasNLP shared task (Moreno, 2021). They introduce a new parallel corpus with Quechua data from three sources: (1) Duran (2010), which contains poems, stories, riddles, songs, phrases and a vocabulary for Quechua; (2) Lyrics translate (2008) which provides different lyrics of poems and songs; and (3) a Quechua handbook (Iter and Ortiz Cárdenas, 2019).

NLLB We use two datasets introduced by Costa-jussà et al. (2022) as part of the training and evaluation for NLLB-200: (1) the NLLB Multi-Domain dataset, which provides 8 809 English-Aymara examples in the news, health, and unscripted chat domains and (2) the NLLB Seed dataset, which contains 6 193 English-Guarani examples consisting of professionally-translated sentences.

Bibles We also collect translations from the JHU Bible corpus (McCarthy et al., 2020), which provides translations of the bible for all languages of the Shared Task except for Chatino. However, we do not observe performance improvements from using this data in our experiments (Section 5).

2.2 Backtranslations

We generate backtranslations using the monolingual data sourced by Vázquez et al. (2021) for seven languages. This data comes from Bustamante et al. (2020), Tiedemann (2020), Mager et al. (2018), Tiedemann (2012), and Agić and Vulić (2019). We train NLLB-200 3.3B on X-*es* for all 11 languages, X, in the task. We take two checkpoints of this model at different stages of training (**backtrans 1** and **backtrans 2**). We find this data improve performance for two of the languages in the task (*gn* and *shp*, see Section 4).

2.3 Data Overlap

We note that NLLB-200, the pretrained machine translation model we base our experiments on (see Section 3) is trained on a portion of the collected data. Specifically, Spanish-Aymara and English-Aymara data from GlobalVoices, and Spanish-Quechua data from Tatoeba, both as part of OPUS. We believe that the inclusion of this data will still be beneficial to the model, since NLLB-200 is not optimised for the languages we are interested in as part of this task.

Model	aym	bzd	cni	czn	gn	hch	nah	oto	quy	shp	tar	mean
Baseline												
(Vázquez et al., 2021)	32.7	23.8	26.8	-	31.1	29.9	29.8	14.7	33.8	31.7	19.6	27.4
Submission 3												
NLLB-1.3B (single best)	39.1	24.5	30.5	40.1	35.5	31.8	30.1	14.7	35.8	32.2	19.4	29.4
Submission 2												
NLLB-1.3B (best per lang)		24.6										
NLLB-3.3B									38.8			
NLLB-1.3B (- NLLB Seed)	41.1											30.3
NLLB-1.3B (+ backtrans 1)					36.9							
NLLB-1.3B (+ backtrans 2)										35.4		
Submission 1												
Ensemble 1		25.1										
Ensemble 2				40.2								
Ensemble 3						31.8						30.5
Ensemble 4									39.1			
Ensemble 5											20.0	

Table 2: Dev set chrF scores for our three submissions. Here, the mean excludes *czn*.

2.4 Data Processing

The training data provided by the organisers is tokenised for *nah* and *oto*. We detokenise it to put it in line with the rest of the training data. We replace punctuation not included in NLLB-200's vocabulary. For *oto*, we find that 7% of the dev set contains characters not in the vocabulary, since these characters do not occur in the training sets, we don't take steps to handle them. For *czn*, we replace all superscript tone markings at the end of words with their standard counterparts, and then replace them naively back at inference.

3 Models

To tackle the 2023 AmericasNLP task on automatic translation of eleven low resource indigenous languages, we use NLLB-200 (Costa-jussà et al., 2022), a state-of-the-art machine translation model specifically designed for low resource settings. We experiment with different distilled versions of NLLB-200 with 600M and 1.3B parameters, and the version with 3.3B parameters. Although inference results on three languages[3] show that the largest version, NLLB-3.3B, performs better than smaller versions (see Table 4), due to the large computational cost of using NLLB-3.3B we run most of our experiments with the 1.3B distilled version. Models are fine-tuned on all the training data (Train Total), i.e. all data sources in Section 2 excluding Bibles and backtranslations, unless indicated. Moreover, we look at ensembling as an approach to improve the overall performance.

Submission 3 We train NLLB-200 1.3B distilled on the training data[4] and we choose the best checkpoint based on average chrF across all languages. We submit translations for all languages using this model (**NLLB-1.3B (single best)**).

Submission 2 We take the best-performing single model per language, excluding ensembles. We find that for the majority of languages, the best single model (by dev chrF) is the same as Submission 3, so we only submit additional translations for five languages:

- **NLLB-1.3B (- NLLB Seed) - *aym*** NLLB-1.3B trained on all data (Train Total) except for NLLB Seed.
- **NLLB-1.3B (best per lang) - *bzd*** NLLB-1.3B trained on all data.
- **NLLB-1.3B (+ backtrans 1) - *gn*** NLLB-1.3B trained on all data plus backtranslations from checkpoint 1.
- **NLLB-3.3B - *quy*** NLLB-3.3B trained on all data.
- **NLLB-1.3B (+ backtrans 2) - *shp*** NLLB-1.3B trained on all data plus backtranslations from checkpoint 2.

Submission 1 We experiment with various ensembles of models in attempt to improve performance further – we only find improvements over Submission 2 through ensembling for five of the

[3] NLLB-200 training data includes *aym*, *gn* and *quy*.

[4] We exclude Bibles data and backtranslations.

Submission	aym	bzd	cni	czn	gn	hch	nah	oto	quy	shp	tar	mean
3	35.3	24.5	28.5	39.9	39.1	32.0	27.3	14.8	37.2	28.6	18.4	29.6
2	36.2	24.4			39.3				39.3	33.4		30.3
1		25.0		40.0		32.3			39.5		18.7	30.5

Table 3: Test set chrF scores for our three submissions. Here, the mean includes all languages.

Model	quy	aym	gn
Baseline			
(Vázquez et al., 2021)	33.8	32.7	31.1
Inference			
600M distilled	30.0	34.2	32.5
1.3B distilled	31.0	35.2	35.2
1.3B	31.2	34.5	34.3
3.3B	32.9	35.4	35.6
Submission			
3	35.8	39.1	35.5
2	38.8	41.1	36.9
1	39.1	-	-

Table 4: Dev set chrF results for various NLLB-200 models, compared to the baseline and our submissions.

languages in the task. These selected ensembles are as follows:

- **Ensemble 1 - *bzd*** The best NLLB-1.3B model for *bzd* and an NLLB-600M model trained on all languages.
- **Ensemble 2 - *czn*** The best average NLLB-1.3B model and an NLLB-3.3B model trained on all languages.
- **Ensemble 3 - *hch*** The best average NLLB-1.3B model and an NLLB-600M model trained on all languages.
- **Ensemble 4 - *quy*** NLLB-3.3B trained on all languages, NLLB-3.3B trained on just the three supported languages (*aym*, *gn*, and *quy*), and NLLB-1.3B trained on all languages.
- **Ensemble 5 - *tar*** NLLB-1.3B trained on all languages, NLLB-600M trained on all languages, and NLLB-1.3B trained on all languages with a label smoothing of 0.2 (rather than 0.1).

3.1 Experimental Setup

We train the models in a multilingual fashion across all 11 language pairs present in the task, extending the embedding matrix to cover the tags for the new languages. We experiment with freezing various parameters, but find best results from training everything. We run our experiments on a single A100 GPU with batch sizes of 64, 16, and 2 for the 600M-, 1.3B-, and 3.3B-parameter models, respectively. We run our experiments in fairseq (Ott et al., 2019). Full hyperparameters for all of our runs are provided in Table 7. To evaluate our models, following the official evaluation, we use chrF (Popović, 2015) computed using SacreBLEU (Post, 2018) with signature: `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0`.

4 Results

4.1 Dev Set Results

Table 2 presents the results of our models on the dev set. We observe that for all languages, at least one of our models outperforms the baseline (Vázquez et al., 2021), with the exception of *oto* where we obtain comparable performance. The greatest improvements over the baseline model are on the three NLLB supported languages: *aym* (41.1 compared to 32.7), *gn* (36.9 compared to 31.1) and *quy* (39.1 compared to 33.8). We note that backtranslations only lead to improved performance on *gn* and *shp*, which are the two languages with the greatest amount of available monolingual data.

Inference results NLLB-200 is trained on data from three of the languages in this shared task: *quy*, *aym*, *gn*. Table 4 shows the inference results for these languages on the dev set for different variations of NLLB-200 models, along with our submissions. We observe a considerable improvement from the distilled 600M to 1.3B distilled models, with the greatest improvement over the baseline model for *gn*. We note that the 1.3B and 3.3B models outperform the baseline model for *aym* and *gn*. For *quy*, the inference results are worse than the baseline, likely due to the large amount of training data available in the task. We are able to improve substantially upon the inference results for *quy* and *aym*, but much less so for *gn* – this may be due to much less training data being available for *gn* compared to the other two languages.

Model	aym	bzd	cni	czn	gn	hch	nah	oto	quy	shp	tar	mean
NLLB-1.3B (single best)	39.1	24.5	30.5	40.1	35.5	31.8	30.1	14.7	35.8	32.2	19.4	30.3
NLLB-3.3B only quy									35.3			
NLLB-3.3B all langs									38.3			
1.3B random initialisation	21.9	17.6	24.2	33.7	22.8	25.1	24.3	13.7	22.9	22.2	16.9	22.3
NLLB-1.3B + bibles	38.3	24.1	30.0	38.0	35.5	30.0	28.0	14.7	35.2	31.9	18.9	28.7

Table 5: Dev set chrF scores for our additional experiments. For comparison, we reproduce the best single model as the first row.

4.2 Test Set Results

Results on the test set are shown in Table 3. Overall, our best submission achieves the highest average chrF across all languages from all submissions to the task (the second-best average is 29.4, compared to our 30.5). We also rank first for four of the eleven languages: *aym*, *czn*, *quy*, and *shp*. Our biggest improvement upon the second-place team is for *czn*, where we achieve 40.0 compared to 36.6. Submissions 1 and 2 rank in the top 3 for all languages. Surprisingly, the best chrF score was obtained on *czn* (40.0), the language with the least amount of training data (3 118 examples), followed by *quy* (39.5), and *aym* (36.5).

5 Additional Experiments

We provide the results of additional experiments to better understand the impact of various factors to our model's performance. The results of these experiments are shown in Table 5.

Multilingual training We look into whether multilingual training is beneficial to the model. For this, we train a 3.3B-parameter model on the *quy* data only, and compare this version (NLLB-3.3B only quy) to the one trained on all languages (NLLB-3.3B all langs) at the same number of updates (480 000). We find that multilingual training greatly improves the performance on *quy*, suggesting the model benefits from transfer learning across the languages. We suspect the benefit of the multilingual approach is related to the fact that although the languages included in the task are from different linguistic families, they share linguistic properties (e.g., polysynthetic or agglutinative).

Random initialization To analyse the benefit of starting from NLLB-200, we train an equivalent model to the 1.3B parameter version with randomly-initialised parameters. We see that this model performs much worse than the equivalent NLLB-200 model. As expected, we observe the greatest differences on the languages supported by NLLB-200 (*aym*, *gn*, *quy*).

Bibles data Similar to findings of Vázquez et al. (2021), we observe a drop in average performance through training on the Bibles data for the majority of languages except for *gn* and *oto*, where we obtain comparable performance.

	es-shp	quy-shp	aym-shp	gn-shp
NLLB-1.3B (+ backtrans 2)	35.4	30.5	29.3	26.7

Table 6: Dev set chrF scores for three zero-shot translation directions with our best model for *es-shp*.

6 Zero-shot Performance

We investigate whether our models have any zero-shot capabilities, i.e. translating a language pair for which the model has not seen any training data. For this, we take the best-performing model for *es-shp* (NLLB-1.3B + backtrans 2), and evaluate it on translating *quy-shp*, *aym-shp*, and *gn-shp*.[5] The results of these experiments are shown in Table 6. We find that our model is able to retain decent performance for these three zero-shot directions (maximum 25% drop in chrF), despite training all of the parameters of the machine translation model.

7 Conclusions

In this paper we describe our submissions to the AmericasNLP 2023 Shared Task. We participated with three submissions which consist of training different versions of the NLLB-200 model on publicly available data from different sources. Models are trained in a multilingual fashion and we experiment with different ensembles of models to further improve performance. We improve upon the inference scores for NLLB-200 3.3B for its three supported languages, and our best submission achieved the highest average chrF across all languages of any submission to the task.

[5]This is possible due to multiparallel dev sets across all languages.

Acknowledgments

This work is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by the UK Research and Innovation grant EP/S023062/1.

References

Constenla Umaña Adolfo, F Elizondo Figueroa, and F Pereira Mora. 1998. Curso básico de bribri.

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

David Brambila. 1976. *Diccionario rarámuri-castellano (tarahumar)*. Obra Nacional de la buena Prensa.

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, Marseille, France. European Language Resources Association.

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

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*.

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

Maximiliano Duran. 2010. La lengua general de los incas. Accessed: : 2023-05-25.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Enora Rice, Cynthia Montaño, John Ortega, Shruti Rijhwani, Alexis Palmer, Rolando Coto-Solano, Hilaria Cruz, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Third Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Association for Computational Linguistics.

Margery Peña Enrique. 2005. Diccionario fraseológico bribri-espanol˜ espanol-bribri, 2nd edn˜. *San Jose: Editorial de la Universidad de Costa Rica.[Google Scholar]*.

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

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

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Cesar Iter and Zenobio Ortiz Cárdenas. 2019. Runasimita yachasun.

Carla Victoria Jara Murillo. 1993. I ttè. historias bribris. *San José: Editorial Universidad de Costa Rica*.

Carla Victoria Jara Murillo. 2018. *Gramática de la lengua bribri*.

Lyrics translate. 2008. Lyrics translate. Accessed: : 2023-05-25.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. Probabilistic finite-state morphological segmenter for wixarika (huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Manuel Mager, Carrillo Dionico, and Ivan Meza. 2020. The wixarika-spanish parallel corpus the wixarika-spanish parallel corpus.(august 2018).

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.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the*

Twelfth Language Resources and Evaluation Conference, pages 2884–2892, Marseille, France. European Language Resources Association.

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

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

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.

Carla Victoria Jara Murillo and Alí García Segura. 2013. *Se'ttö bribri ie: Hablemos en bribri*. Programa de Regionalización Interuniversitaria CONARE.

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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.

Sofía Flores Solórzano. 2017. Corpus oral pandialectal de la lengua bribri.

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

Jörg Tiedemann. 2020. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*.

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.

A Hyperparameters

Hyper-parameter	Value
Batch size	16†
Update freq	1
Max learning rate	0.01
Schedule	inverse square root
Warmup steps	10 000
Adam betas	0.9, 0.98
Label smoothing	0.1‡
Weight decay	0.0001
Dropout	0.3
Clip norm	1e-6
Language pair temperature	3⋆
Number of updates	1M
Valid freq	40K updates + every epoch
Beam size	5

Table 7: Hyper-parameters used to train our models.
†: 64 for NLLB-600M, 2 for NLLB-3.3B.
‡: 0.2 for one of our models, used in Ensemble 5.
⋆: 1 for NLLB-3.3B models (including for backtranslations)

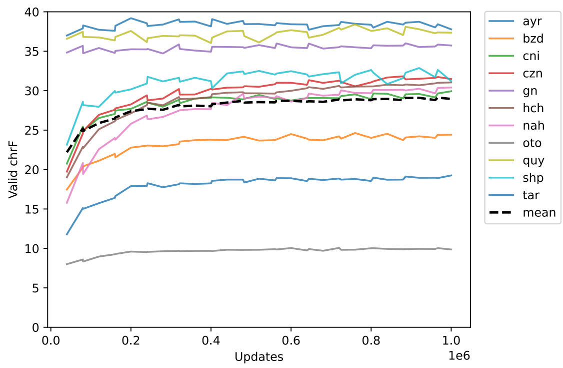

Figure 1: Valid chrF scores during training of our best single model (Submission 3).

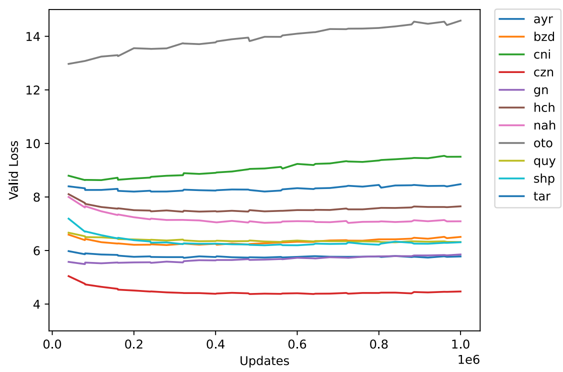

Figure 2: Valid losses during training of our best single model (Submission 3).

Dataset	Source
ashaninka-spanish	Ortega et al. (2020) Cushimariano Romano and Sebastián Q. (2008) Mihas (2011)
aymara-spanish	GlobalVoices (Tiedemann, 2012)
bribri-spanish	Adolfo et al. (1998) Solórzano (2017) Jara Murillo (2018) Murillo and Segura (2013) Jara Murillo (1993) Enrique (2005)
guarani-spanish	Chiruzzo et al. (2020)
hñähñu-spanish	Tsunkua `https://tsunkua.elotl.mx/about/`
wixarika-spanish	Mager et al. (2020)
shipibo_konibo-spanish	Montoya et al. (2019) Galarreta et al. (2017)
raramuri-spanish	Brambila (1976)
quechua-spanish	JW300 (Agić and Vulić, 2019) GlobalVoices (Tiedemann, 2012)
nahuatl-spanish	Axolotl (Gutierrez-Vasques et al., 2016)
chatino-spanish	IUScholar Works `https://scholarworks.iu.edu/dspace/handle/2022/21028`

Table 8: Data provided by the organisers of the 2023 AmericasNLP