Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas

Manuel Mager[♠*] **Arturo Oncevay**[♡*] **Abteen Ebrahimi**[♢*] **John Ortega**[Ω]
Annette Rios[ψ] **Angela Fan**[▽] **Ximena Gutierrez-Vasques**[ψ] **Luis Chiruzzo**[△]
Gustavo A. Giménez-Lugo[♣] **Ricardo Ramos**[η] **Ivan Vladimir Meza Ruiz**[♯]
Rolando Coto-Solano[℧] **Alexis Palmer**[♢] **Elisabeth Mager**[♯] **Vishrav Chaudhary**[▽]
Graham Neubig[⋈] **Ngoc Thang Vu**[♠] **Katharina Kann**[♢]
[⋈]Carnegie Mellon University [℧]Dartmouth College [▽]Facebook AI Research
[Ω]New York University [△]Universidad de la República, Uruguay
[η]Universidad Tecnológica de Tlaxcala [♯]Universidad Nacional Autónoma de México
[♣]Universidade Tecnológica Federal do Paraná [♢]University of Colorado Boulder
[♡]University of Edinburgh [♠]University of Stuttgart [ψ]University of Zurich

Abstract

This paper presents the results of the 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. The shared task featured two independent tracks, and participants submitted machine translation systems for up to 10 indigenous languages. Overall, 8 teams participated with a total of 214 submissions. We provided training sets consisting of data collected from various sources, as well as manually translated sentences for the development and test sets. An official baseline trained on this data was also provided. Team submissions featured a variety of architectures, including both statistical and neural models, and for the majority of languages, many teams were able to considerably improve over the baseline. The best performing systems achieved 12.97 ChrF higher than baseline, when averaged across languages.

1 Introduction

Many of the world's languages, including languages native to the Americas, receive worryingly little attention from NLP researchers. According to Glottolog (Nordhoff and Hammarström, 2012), 86 language families and 95 language isolates can be found in the Americas, and many of them are labeled as endangered. From an NLP perspective, the development of language technologies has the potential to help language communities and activists in the documentation, promotion and revitalization of their languages (Mager et al., 2018b; Galla, 2016). There have been recent initiatives to promote research on languages of the Americas (Fernández et al., 2013; Coler and Homola, 2014; Gutierrez-Vasques, 2015; Mager and Meza, 2018; Ortega et al., 2020; Zhang et al., 2020; Schwartz et al., 2020; Barrault et al., 2020).

The AmericasNLP 2021 Shared Task on Open Machine Translation (OMT) aimed at moving research on indigenous and endangered languages more into the focus of the NLP community. As the official shared task training sets, we provided a collection of publicly available parallel corpora (§3). Additionally, all participants were allowed to use other existing datasets or create their own resources for training in order to improve their systems. Each language pair used in the shared task consisted of an indigenous language and a high-resource language (Spanish). The languages belong to a diverse set of language families: Aymaran, Arawak, Chibchan, Tupi-Guarani, Uto-Aztecan, Oto-Manguean, Quechuan, and Panoan. The ten language pairs included in the shared task are: Quechua–Spanish, Wixarika–Spanish, Shipibo-Konibo–Spanish, Asháninka–Spanish, Raramuri–Spanish, Nahuatl–Spanish, Otomí–Spanish, Aymara–Spanish, Guarani–Spanish, and Bribri–Spanish. For development and testing, we used parallel sentences belonging to a new natural language inference dataset for the 10 indigenous languages featured in our shared task, which is a manual translation of the Spanish version of the multilingual XNLI dataset (Conneau et al., 2018). For a complete description of this dataset we refer the reader to Ebrahimi et al. (2021).

Together with the data, we also provided: a simple baseline based on the small transformer architecture (Vaswani et al., 2017) proposed together with the FLORES dataset (Guzmán et al., 2019); and a description of challenges and particular characteristics for all provided resources[1]. We established two tracks: one where training models on the development set after hyperparameter tuning is

*The first three authors contributed equally.

[1]https://github.com/AmericasNLP/americasnlp2021/blob/main/data/information_datasets.pdf

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 202–217

allowed (Track 1), and one where models cannot be trained directly on the development set (Track 2).

Machine translation for indigenous languages often presents unique challenges. As many indigenous languages do not have a strong written tradition, orthographic rules are not well defined or standardized, and even if they are regulated, often times native speakers do not follow them or create their own adapted versions. Simply normalizing the data is generally not a viable option, as even the definition of what constitutes a morpheme or a orthographic word is frequently ill defined. Furthermore, the huge dialectal variability among those languages, even from one village to the other, adds additional complexity to the task. We describe the particular challenges for each language in Section §3.

Eight teams participated in the AmericasNLP 2021 Shared Task on OMT. Most teams submitted systems in both tracks and for all 10 language pairs, yielding a total of 214 submissions.

2 Task and Evaluation

2.1 Open Machine Translation

Given the limited availability of resources and the important dialectal, orthographic and domain challenges, we designed our task as an unrestrained machine translation shared task: we called it *open* machine translation to emphasize that participants were free to use any resources they could find. Possible resources could, for instance, include existing or newly created parallel data, dictionaries, tools, or pretrained models.

We invited submissions to two different tracks: Systems in Track 1 were allowed to use the development set as part of the training data, since this is a common practice in the machine translation community. Systems in Track 2 were not allowed to be trained directly on the development set, mimicking a more realistic low-resource setting.

2.2 Primary Evaluation

In order to be able to evaluate a large number of systems on all 10 languages, we used automatic metrics for our primary evaluation. Our main metric, which determined the official ranking of systems, was ChrF (Popović, 2015). We made this choice due to certain properties of our languages, such as word boundaries not being standardized for all languages and many languages being polysynthetic, resulting in a small number of words per sentence. We further reported BLEU scores (Papineni et al., 2002) for all systems and languages.

2.3 Supplementary Evaluation

To gain additional insight into the strengths and weaknesses of the top-performing submissions, we further performed a supplementary manual evaluation for two language pairs and a limited number of systems, using a subset of the test set.

We asked our annotators to provide ratings of system outputs using separate 5-point scales for adequacy and fluency. The annotation was performed by the translator who created the test datasets. The expert received the source sentence in Spanish, the reference in the indigenous language, and an anonymized system output. In addition to the baseline, we considered the 3 highest ranked systems according to our main metric, and randomly selected 100 sentences for each language. The following were the descriptions of the ratings as provided to the expert annotator in Spanish (translated into English here for convenience):

Adequacy The output sentence expresses the meaning of the reference.

1. Extremely bad: The original meaning is not contained at all.
2. Bad: Some words or phrases allow to guess the content.
3. Neutral.
4. Sufficiently good: The original meaning is understandable, but some parts are unclear or incorrect.
5. Excellent: The meaning of the output is the same as that of the reference.

Fluency The output sentence is easily readable and looks like a human-produced text.

1. Extremely bad: The output text does not belong to the target language.
2. Bad: The output sentence is hardly readable.
3. Neutral.
4. Sufficiently good: The output seems like a human-produced text in the target language, but contains weird mistakes.
5. Excellent: The output seems like a human-produced text in the target language, and is readable without issues.

Language	ISO	Family	Train	Dev	Test
Asháninka	cni	Arawak	3883	883	1002
Aymara	aym	Aymaran	6531	996	1003
Bribri	bzd	Chibchan	7508	996	1003
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 featured in the AmericasNLP 2021 Shared Task on OMT, their ISO codes, language families and dataset statistics. For the origins of the datasets, please refer to the text.

3 Languages and Datasets

In this section, we will present the languages and datasets featured in our shared task. Figure 1 additionally provides an overview of the languages, their linguistic families, and the number of parallel sentences with Spanish.

3.1 Development and Test Sets

For system development and testing, we leveraged individual pairs of parallel sentences from AmericasNLI (Ebrahimi et al., 2021). This dataset is a translation of the Spanish version of XNLI (Conneau et al., 2018) into our 10 indigenous languages. It was not publicly available until after the conclusion of the competition, avoiding an accidental inclusion of the test set into the training data by the participants. For more information regarding the creation of the dataset, we refer the reader to (Ebrahimi et al., 2021).

3.2 Training Data

We collected publicly available datasets in all 10 languages and provided them to the shared task participants as a starting point. We will now introduce the languages and the training datasets, explaining similarities and differences between training sets on the one hand and development and test sets on the other.

Spanish–Wixarika Wixarika (also known as Huichol) with ISO code `hch` is spoken in Mexico and belongs to the Yuto-Aztecan linguistic family. The training, development and test sets all belong to the same dialectal variation, Wixarika of Zoquipan, and use the same orthography. However, word boundaries are not always marked according to the same criteria in development/test and train. The training data (Mager et al., 2018a) is a translation of the fairy tales of Hans Christian Andersen and contains word acquisitions and code-switching.

Spanish–Nahuatl Nahuatl is a Yuto-Aztecan language spoken in Mexico and El Salvador, with a wide dialectal variation (around 30 variants). For each main dialect a specific ISO 639-3 code is available.[2] There is a lack of consensus regarding the orthographic standard. This is very noticeable in the training data: the train corpus (Gutierrez-Vasques et al., 2016) has dialectal, domain, orthographic and diachronic variation (Nahuatl side). However, the majority of entries are closer to a Classical Nahuatl orthographic "standard".

The development and test datasets were translated to modern Nahuatl. In particular, the translations belong to Nahuatl Central/Nahuatl de la Huasteca (Hidalgo y San Luis Potosí) dialects. In order to be closer to the training corpus, an orthographic normalization was applied. A simple rule based approach was used, which was based on the most predictable orthographic changes between modern varieties and Classical Nahuatl.

Spanish—Guarani Guarani is mostly spoken in Paraguay, Bolivia, Argentina and Brazil. It belongs to the Tupian language family (ISO `gnw, gun, gug, gui, grn, nhd`). The training corpus for Guarani (Chiruzzo et al., 2020) was collected from web sources (blogs and news articles) that contained a mix of dialects, from pure Guarani to more mixed Jopara which combines Guarani with Spanish neologisms. The development and test corpora, on the other hand, are in standard Paraguayan Guarani.

Spanish—Bribri Bribri is a Chibchan language spoken in southern Costa Rica (ISO code `bzd`). The training set for Bribri was extracted from six sources (Feldman and Coto-Solano, 2020; Margery, 2005; Jara Murillo, 2018a; Constenla et al., 2004; Jara Murillo and García Segura, 2013; Jara Murillo, 2018b; Flores Solórzano, 2017), including a dictionary, a grammar, two language learning textbooks, one storybook and the transcribed sentences from

[2] ISO 639-3 for the Nahutal languages: `nci, nhn, nch, ncx, naz, nln, nhe, ngu, azz, nhq, nhk, nhx, nhp, ncl, nhm, nhy, ncj, nht, nlv, ppl, nhz, npl, nhc, nhv, nhi, nhg, nuz, nhw, nsu, xpo, nhn, nch, ncx, naz, nln, nhe, ngu, azz, nhq, nhk, nhx, nhp, ncl, nhm, nhy, ncj, nht, nlv, ppl, nhz, npl, nhc, nhv, nhi, nhg, nuz, nhw, nsu,` and `xpo`.

one spoken corpus. The sentences belong to three major dialects: Amubri, Coroma and Salitre.

There are numerous sources of variation in the Bribri data (Feldman and Coto-Solano, 2020): 1) There are several different orthographies, which use different diacritics for the same words. 2) The Unicode encoding of visually similar diacritics differs among authors. 3) There is phonetic and lexical variation across dialects. 4) There is considerable idiosyncratic variation between writers, including variation in word boundaries (e.g. *ikíe* vrs *i kie* "it is called"). In order to build a standardized training set, an intermediate orthography was used to make these different forms comparable and learning easier. All of the training sentences are comparable in domain; they come from either traditional stories or language learning examples. Because of the nature of the texts, there is very little code-switching into Spanish. This is different from regular Bribri conversation, which would contain more borrowings from Spanish and more code-switching. The development and test sentences were translated by a speaker of the Amubri dialect and transformed into the intermediate orthography.

Spanish—Rarámuri Rarámuri is a Uto-Aztecan language, spoken in northern Mexico (ISO: `tac, twr, tar, tcu, thh`). Training data for Rarámuri consists of a set of extracted phrases from the Rarámuri dictionary Brambila (1976). However, we could not find any description of the dialectal variation to which these examples belong. The development and test set are translations from Spanish into the highlands Rarámuri variant (`tar`), and may differ from the training set. As with many polysynthetic languages, challenges can arise when the boundaries of a morpheme and a word are not clear and have no consensus. Native speakers, even with a standard orthography and from the same dialectal variation, may define words in a different standards to define word boundaries.

Spanish—Quechua Quechua is a family of languages spoken in Argentina, Bolivia, Colombia, Ecuador, Peru, and Chile with many ISO codes for its language (`quh, cqu, qvn, qvc, qur, quy, quk, qvo, qve`, and `quf`). The development and test sets are translated into the standard version of Southern Quechua, specifically the Quechua Chanka (Ayacucho, code: `quy`) variety. This variety is spoken in different regions of Peru, and it can be understood in different areas of other countries, such as Bolivia or Argentina. This is the variant used on Wikipedia Quechua pages, and by Microsoft in its translations of software into Quechua. Southern Quechua includes different Quechua variants, such as Quechua Cuzco (`quz`) and Quechua Ayacucho (`quy`). Training datasets are provided for both variants. These datasets were created from JW300 (Agić and Vulić, 2019), which consists of Jehovah's Witness texts, sentences extracted from the official dictionary of the Minister of Education (MINEDU), and miscellaneous dictionary entries and samples which have been collected and reviewed by Huarcaya Taquiri (2020).

Spanish–Aymara Aymara is a Aymaran language spoken in Bolivia, Peru, and Chile (ISO codes `aym, ayr, ayc`). The development and test sets are translated into the Central Aymara variant (`ayr`), specifically Aymara La Paz jilata, the largest variant. This is similar to the variant of the available training set, which is obtained from Global Voices (Prokopidis et al., 2016) (and published in OPUS (Tiedemann, 2012)), a news portal translated by volunteers. However, the text may have potentially different writing styles that are not necessarily edited.

Spanish-–Shipibo-Konibo Shipibo-Konibo is a Panoan language spoken in Perú (ISO `shp` and `kaq`). The training sets for Shipibo-Konibo have been obtained from different sources and translators: Sources include translations of a sample from the Tatoeba dataset (Gómez Montoya et al., 2019), translated sentences from books for bilingual education (Galarreta et al., 2017), and dictionary entries and examples (Loriot et al., 1993). Translated text was created by a bilingual teacher, and follows the most recent guidelines of the Minister of Education in Peru, however, the third source is an extraction of parallel sentences from an old dictionary. The development and test sets were created following the official convention as in the translated training sets.

Spanish—Asháninka Asháninka is an Arawakan language (ISO: `cni`) spoken in Peru and Brazil. Training data was created by collecting texts from different domains such as traditional stories, educational texts, and environmental laws for the Amazonian region (Ortega et al., 2020; Romano, Rubén and Richer, Sebastián, 2008; Mihas, 2011). The texts belong to domains

such as: traditional stories, educational texts, environmental laws for the Amazonian region. Not all the texts are translated into Spanish, there is a small fraction of these that are translated into Portuguese because a dialect of pan-Ashaninka is also spoken in the state of Acre in Brazil. The texts come from different pan-Ashaninka dialects and have been normalized using the AshMorph (Ortega et al., 2020). There are many neologisms that are not spread to the speakers of different communities. The translator of the development and test sets only translated the words and concepts that are well known in the communities, whereas other terms are preserved in Spanish. Moreover, the development and test sets were created following the official writing convention proposed by the Peruvian Government and taught in bilingual schools.

Spanish-–Otomí Otomí (also known as Hñähñu, Hñähño, Ñhato, Ñûhmû, depending on the region) is an Oto-Manguean language spoken in Mexico (ISO codes: `ott, otn, otx, ote, otq, otz, otl, ots, otm`). The training set[3] was collected from a set of different sources, which implies that the text contains more than one dialectal variation and orthographic standard, however, most texts belong to the Valle del Mezquital dialect (`ote`). This was specially challenging for the translation task, since the development and test sets are from the Ñûhmû de Ixtenco, Tlaxcala, variant (`otz`), which also has its own orthographic system. This variant is especially endangered as less than 100 elders still speak it.

3.3 External Data Used by Participants

In addition to the provided datasets, participants also used additional publicly available parallel data, monolingual corpora or newly collected data sets. The most common datasets were JW300 (Agić and Vulić, 2019) and the Bible's New Testament (Mayer and Cysouw, 2014; Christodouloupoulos and Steedman, 2015; McCarthy et al., 2020). Besides those, GlobalVoices (Prokopidis et al., 2016) and datasets available at OPUS (Tiedemann, 2012) were added. New datasets were extracted from constitutions, dictionaries, and educational books. For monolingual text, Wikipedia was most commonly used, assuming one was available in a language.

4 Baseline and Submitted Systems

We will now describe our baseline as well as all submitted systems. An overview of all teams and the main ideas going into their submissions is shown in Table 2.

4.1 Baseline

Our baseline system was a transformer-based sequence to sequence model (Vaswani et al., 2017). We employed the hyperparameters proposed by Guzmán et al. (2019) for a low-resource scenario. We implemented the model using Fairseq (Ott et al., 2019). The implementation of the baseline can be found in the official shared task repository.[4]

4.2 University of British Columbia

The team of the University of British Columbia (`UBC-NLP`; Billah-Nagoudi et al., 2021) participated for all ten language pairs and in both tracks. They used an encoder-decoder transformer model based on T5 (Raffel et al., 2020). This model was pretrained on a dataset consisting of 10 indigenous languages and Spanish, that was collected by the team from different sources such as the Bible and Wikipedia, totaling 1.17 GB of text. However, given that some of the languages have more available data than others, this dataset is unbalanced in favor of languages like Nahuatl, Guarani, and Quechua. The team also proposed a two-stage fine-tuning method: first fine-tuning on the entire dataset, and then only on the target languages.

4.3 Helsinki

The University of Helsinki (`Helsinki`; Vázquez et al., 2021) participated for all ten language pairs in both tracks. This team did an extensive exploration of the existing datasets, and collected additional resources both from commonly used sources such as the Bible and Wikipedia, as well as other minor sources such as constitutions. Monolingual data was used to generate paired sentences through back-translation, and these parallel examples were added to the existing dataset. Then, a normalization process was done using existing tools, and the aligned data was further filtered. The quality of the data was also considered, and each dataset was assigned a weight depending on a noisiness estimation. The team used a transformer sequence-to-sequence model trained via two steps. For their main submission they first trained on data which

[3] Otomí online corpus: https://tsunkua.elotl.mx/about/

[4] https://github.com/AmericasNLP/americasnlp2021

Team	Langs.	Sub.	Data	Models	Multilingual	Pretrained
CoAStaL (Bollmann et al., 2021)	10	20	Bible, JW300, OPUS, Wikipedia, New collected data	PB-SMT, Constrained Random Strings	No	No
Helsinki (Vázquez et al., 2021)	10	50	Bible, OPUS, Constitutions, Normalization, Filtering, Back-Translation	Transformer NMT	Yes, all ST languages + Spanish-English	No
NRC-CNRC (Knowles et al., 2021)	4	17	No external data, preoricessing, BPE Dropout.	Transofrmer NMT	Yes, 4-languages	No
REPUcs (Moreno, 2021)	1	2	JW300, New dataset, Europarl	Transformer NMT.	Yes, with Spanish-English	Spanish-English pretraining
Tamalli (Parida et al., 2021)	10	42	-	WB-SMT. Transformer NMT,	10-languages	No
UBC-NLP (Billah-Nagoudi et al., 2021)	8	29	Bible, Wikipedia	Transformer T5	10-Languages	New T5
UTokyo (Zheng et al., 2021)	10	40	Monolingual from other languages. Data	Transformer	Yes	New mBART
Anonymous	8	14	-	-	-	-

Table 2: Participating team (*Team*) with system description paper, number of languages that system outputs were submitted for (*Langs.*), total number of submissions (*Sub.*), external data (*Data*), models (*Models*), if training was multilingual (*Multilingual*), and if pretraining was done (*Pretrained*). More details can be found in the text.

was 90% Spanish–English and 10% indigenous languages, and then changed the data proportion to 50% Spanish–English and 50% indigenous languages.

4.4 CoAStaL

The team of the University of Copenhagen (CoAStaL) submitted systems for both tracks (Bollmann et al., 2021). They focused on additional data collection and tried to improve the results with low-resource techniques. The team discovered that it was even hard to generate correct words in the output and that phrase-based statistical machine translation (PB-SMT) systems work well when compared to the state-of-the-art neural models. Interestingly, the team introduced a baseline that mimicked the target language using a character-trigram distribution and length constraints without any knowledge of the source sentence. This random text generation achieved even better results than some of the other submitted systems. The team also reported failed experiments, where character-based neural machine translation (NMT), pretrained transformers, language model priors, and graph convolution encoders using UD annotations could not get any meaningful results.

4.5 REPUcs

The system of the Pontificia Universidad Católica del Perú (REPUcs; Moreno, 2021) submitted to the the Spanish–Quechua language pair in both tracks. The team collected external data from 3 different sources and analyzed the domain disparity between this training data and the development set. To solve the problem of domain mismatch, they decided to collect additional data that could be a better match for the target domain. The used data from a handbook (Iter and Ortiz-Cárdenas, 2019), a lexicon,[5] and poems on the web (Duran, 2010).[6] Their model is a transformer encoder-decoder architecture with SentencePiece (Kudo and Richardson, 2018) tokenization. Together with the existing parallel corpora, the new paired data was used for finetuning on top of a pretrained Spanish–English translation model. The team submitted two versions of their system: the first was only finetuned on JW300+ data, while the second one additionally leveraged the newly collected dataset.

4.6 UTokyo

The team of the University of Tokyo (UTokyo; Zheng et al., 2021) submitted systems for all languages and both tracks. A multilingual pretrained encoder-decoder model (mBART; Liu et al., 2020) was used, implemented with the Fairseq toolkit (Ott et al., 2019). The model was first pretrained on a huge amount of data (up to 13GB) from var-

[5] https://www.inkatour.com/dico/
[6] https://lyricstranslate.com/

Lang.	Rank	Team	Sub	BLEU	ChrF
aym	1	Helsinki	2	2.80	**31.0**
	2	Helsinki	1	2.91	30.2
	3	Helsinki	3	2.35	26.1
	4	UTokyo	1	1.17	21.4
	5	CoAStaL	1	1.11	19.1
	6	UBC-NLP	2	0.99	19.0
	7	UBC-NLP	4	0.76	18.6
	8	UTokyo	2	1.18	14.9
	9	Anonym	1	0.01	7.3
	Rank	Team	Sub	BLEU	ChrF
bzd	1	Helsinki	2	5.18	**21.3**
	2	Helsinki	1	4.93	20.4
	3	CoAStaL	1	3.60	19.6
	4	Helsinki	3	3.68	17.7
	5	UTokyo	1	1.70	14.3
	6	UBC-NLP	2	0.94	11.3
	7	UTokyo	2	1.28	11.2
	8	UBC-NLP	4	0.89	11.1
	9	Anonym	1	0.14	6.1
Lang.	Rank	Team	Sub	BLEU	ChrF
cni	1	Helsinki	2	6.09	**33.2**
	2	Helsinki	1	5.87	32.4
	3	Helsinki	3	5.00	30.6
	4	CoAStaL	1	3.02	26.5
	5	UTokyo	1	0.20	21.6
	6	UTokyo	2	0.84	18.9
	7	UBC-NLP	2	0.08	18.3
	8	UBC-NLP	4	0.09	17.8
	9	Anonym	1	0.08	11.4
Lang.	Rank	Team	Sub	BLEU	ChrF
gn	1	Helsinki	2	8.92	**37.6**
	2	Helsinki	1	8.18	36.7
	3	Helsinki	3	5.97	31.1
	4	NRC-CNRC	0	4.73	30.4
	5	NRC-CNRC	4	5.27	30.3
	6	NRC-CNRC	2	4.06	28.8
	7	UTokyo	1	3.21	26.5
	8	CoAStaL	1	2.20	24.1
	9	UTokyo	2	3.18	23.3
	10	NRC-CNRC	3	0.64	16.3
	11	Anonym	1	0.03	8.5
Lang	Rank	Team	Sub	BLEU	ChrF
hch	1	Helsinki	2	15.67	**36.0**
	2	Helsinki	1	14.71	34.8
	3	NRC-CNRC	0	14.90	32.7
	4	NRC-CNRC	2	13.65	31.5
	5	Helsinki	3	13.72	31.1
	6	CoAStaL	1	8.80	25.7
	7	UTokyo	1	7.09	23.8
	8	NRC-CNRC	3	4.62	20.0
	9	UBC-NLP	2	5.52	19.5
	10	UBC-NLP	4	5.09	18.6
	11	UTokyo	2	6.30	18.4
	12	Amonym	1	0.06	8.1

Lang	Rank	Team	Sub	BLEU	ChrF
nah	1	Helsinki	2	3.25	**30.1**
	2	Helsinki	1	2.8	29.4
	3	NRC-CNRC	0	2.13	27.7
	4	NRC-CNRC	2	1.78	27.3
	5	Helsinki	3	2.76	27.3
	6	UTokyo	1	0.55	23.9
	7	CoAStaL	1	2.06	21.4
	8	UTokyo	2	0.98	19.8
	9	UBC-NLP	2	0.16	19.6
	10	NRC-CNRC	3	0.14	18.1
	11	Anonym	2	0.09	10.3
	12	Anonym	3	0.09	9.7
	13	Anonym	4	0.08	9.5
	14	Anonym	1	0.04	8.7
Lang	Rank	Team	Sub	BLEU	ChrF
oto	1	Helsinki	2	5.59	**22.8**
	2	Helsinki	1	3.85	19.1
	3	CoAStaL	1	2.72	18.4
	4	Helsinki	3	2.9	18.1
	5	UTokyo	2	2.45	15.2
	6	UTokyo	1	0.12	12.8
	7	Anonym	1	0.15	10.2
	8	UBC-NLP	2	0.04	8.4
	9	UBC-NLP	4	0.04	8.3
Lang	Rank	Team	Sub	BLEU	ChrF
quy	1	Helsinki	2	5.38	**39.4**
	2	Helsinki	1	5.16	38.3
	3	REPUcs	2	3.1	35.8
	4	UTokyo	1	2.35	33.2
	5	UTokyo	2	2.62	32.8
	6	Helsinki	3	3.56	31.8
	7	CoAStaL	1	1.63	26.9
	8	Anonym	2	0.23	10.3
	9	Anonym	4	0.13	9.8
	10	Anonym	1	0.06	9.0
	11	Anonym	3	0.03	6.6
Lang	Rank	Team	Sub	BLEU	ChrF
shp	1	Helsinki	2	10.49	**39.9**
	2	Helsinki	1	9.06	38.0
	3	CoAStaL	1	3.9	29.7
	4	Helsinki	3	6.76	28.6
	5	UTokyo	1	0.33	16.3
	6	UTokyo	2	0.46	15.5
	7	UBC-NLP	2	0.23	12.4
Lang	Rank	Team	Sub	BLEU	ChrF
tar	1	Helsinki	2	3.56	**25.8**
	2	Helsinki	1	3.24	24.8
	3	NRC-CNRC	0	2.69	24.7
	4	NRC-CNRC	2	2.1	23.9
	5	Helsinki	3	1.8	21.6
	6	NRC-CNRC	3	0.83	16.5
	7	CoAStaL	1	1.05	15.9
	8	UTokyo	1	0.1	12.2
	9	UBC-NLP	2	0.05	10.5
	10	UBC-NLP	4	0.1	10.5
	11	UTokyo	2	0.69	8.4

Table 3: Results of Track 1 (development set used for training) for all systems and language pairs. The results are ranked by the official metric of the shared task: ChrF. One team decided to send a anonymous submission (*Anonym*). Best results are shown in bold, and they are significantly better than the second place team (in each language-pair) according to the Wilcoxon signed-ranked test and Pitman's permutation test with $p<0.05$ (Dror et al., 2018).

ious high-resource languages, and then finetuned for each target language using the official provided data.

4.7 NRC-CNRC

The team of the National Research Council Canada (NRC-CNRC; Knowles et al., 2021) submitted systems for the Spanish to Wixárika, Nahuatl, Rarámuri and Guarani language pairs for both tracks. Due to ethical considerations, the team decided not to use external data, and restricted themselves to the data provided for the shared task. All data was preprocessed with standard Moses tools (Koehn et al., 2007). The submitted systems were based on a Transformer model, and used BPE for tokenization. The team experimented with multilingual models pretrained on either 3 or 4 languages, finding that the 4 language model achieved higher performance. Additionally the team trained a Translation Memory (Simard and Fujita, 2012) using half of the examples of the development set. Surprisingly, even given its small amount of training data, this system outperformed the team's Track 2 submission for Rarámuri.

4.8 Tamalli

The team Tamalli[7] (Parida et al., 2021) participated in Track 1 for all 10 language pairs. The team used an IBM Model 2 for SMT, and a transformer model for NMT. The team's NMT models were trained in two settings: one-to-one, with one model being trained per target language, and one-to-many, where decoder weights were shared across languages and a language embedding layer was added to the decoder. They submitted 5 systems per language, which differed in their hyperparameter choices and training setup.

5 Results

5.1 Track 1

The complete results for all systems submitted to Track 1 are shown in Table 3. Submission 2 of the Helsinki team achieved first place for all language pairs. Interestingly, for all language pairs, the Helsinki team also achieved the second best result with their Submission 1. Submission 3 was less successful, achieving third place on three pairs. The NRC-CNRC team achieved third place for Wixárika, Nahuatl, and Rarámuri, and fourth for Guarani.The lower automatic scores of their systems can also be partly due to the team not using additional datasets. The REPUcs system obtained the third best result for Quechua, the only language they participated in. CoAStaL's first system, a PB-SMT model, achieved third place for Bribri, Otomí, and Shipibo-Konibo, and fourth place for Ashaninka. This suggests that SMT is still competitive for low-resource languages. UTokyo and UBC-NLP were less successful than the other approaches. Finally, we attribute the bad performance of the anonymous submission to a possible bug. Since our baseline system was not trained on the development set, no specific baseline was available for this track.

5.2 Track 2

All results for Track 2, including those of our baseline system, are shown in Table 5.

Most submissions outperformed the baseline by a large margin. As for Track 1, the best system was from the Helsinki team (submission 5), winning 9 out of 10 language pairs. REPUcs achieved the best score for Spanish–Quechua, the only language pair they submitted results for. Their pretraining on Spanish–English and the newly collected dataset proved to be successful.

Second places were more diverse for Track 2 than for Track 1. The NRC-CNRC team achieved second place for two languages (Wixarika and Guarani), UTokyo achieved second place for three languages (Aymara, Nahuatl and Otomí), and the Helsinki team came in second for Quechua. Tamalli only participated in Track 2, with 4 systems per language. Their most successful one was submission 1, a word-based SMT system. An interesting submission for this track was the CoAStaL submission 2, which created a random generated output that mimics the target language distribution. This system consistently outperformed the official baseline and even outperformed other approaches for most languages.

5.3 Supplementary Evaluation Results

As explained in §2, we also conducted a small human evaluation of system outputs based on adequacy and fluency on a 5-points scale, which was performed by a professional translator for two language-pairs: Spanish to Shipibo-Konibo and

[7]Participating universities: Idiap Research Institute, City University of New York, BITS-India, Universidad Autónoma Metropolitana-México, Ghent University, and Universidad Politécnica de Tulancingo-México

System	aym	bzd	cni	gn	hch	nah	oto	quy	shp	tar	Avg.
Baseline	49.33	**52.00**	42.80	55.87	41.07	54.07	36.50	59.87	52.00	43.73	48.72
Helsinki-5	**57.60**	48.93	**55.33**	**62.40**	**55.33**	**62.33**	**49.33**	**60.80**	**65.07**	**58.80**	57.59
NRC-CNRC-1	-	-	-	57.20	50.40	58.94	-	-	-	53.47	55.00*

Table 4: Results of the NLI analysis. * indicates that the average score is not directly comparable as the number of languages differs for the given system.

Otomí.[8] This evaluation was performed given the extremely low automatic evaluation scores, and the natural question about the usefulness of the outputs of MT systems at the current state-of-the-art. While we selected two languages as a sample to get a better approximation to this question, further studies are needed to draw stronger conclusions.

Figure 1 shows the adequacy and fluency scores annotated for Spanish–Shipibo-Konibo and Spanish–Otomí language-pairs. considering the baseline and the three highest ranked systems according to ChrF. For both languages, we observe that the adequacy scores are similar between all systems except for `Helsinki`, the best ranked submission given the automatic evaluation metric, which has more variance than the others. However, the average score is low, around 2, which means that only few words or phrases express the meaning of the reference.

Looking at fluency, there is less similarity between the Shipibo-Konibo and Otomí annotations. For Shipibo-Konibo, there is no clear difference between the systems in terms of their average scores. We note that `Tamalli`'s system obtained the larger group with the relatively highest score. For Otomí, the three submitted systems are at least slightly better than the baseline on average, but only in 1 level of the scale. The scores for fluency are similar to adequacy in this case. Besides, according to the annotations, the output translations in Shipibo-Konibo were closer to human-produced texts than in Otomí.

We also show the relationship between ChrF and the adequacy and fluency scores in Figure 2. However, there does not seem to be a correlation between the automatic metric and the manually assigned scores.

[8] In the WMT campaigns, it is common to perform a crowd-sourced evaluation with several annotators. However, we cannot follow that procedure given the low chance to find native speakers of indigenous languages as users in crowd-sourcing platforms.

5.4 Analysis: NLI

One approach for zero-shot transfer learning of a sequence classification task is the translate-train approach, where a translation system is used to translate high-resource labeled training data into the target language. In the case of pretrained multilingual models, these machine translated examples are then used for finetuning. For our analysis, we used various shared task submissions to create different sets of translated training data. We then trained a natural language inference (NLI) model using this translated data, and used the downstream NLI performance as an extrinsic evaluation of translation quality.

Our experimental setup was identical to Ebrahimi et al. (2021). We focused only on submissions from Track 2, and analyzed the `Helsinki-5` and the `NRC-CNRC-1` system. We present results in Table 4. Performance from using the `Helsinki` system far outperforms the baseline on average, and using the `NRC-CNRC` system also improves over the baseline. For the four languages covered by all systems, we can see that the ranking of NLI performance matches that of the automatic ChrF evaluation. Between the Helsinki and Baseline systems, this ranking also holds for every other language except for Bribri, where the Baseline achieves around 3 percentage points higher accuracy. Overall, this evaluation both confirms the ranking created by the ChrF scores and provides strong evidence supporting the use of translation-based approaches for zero-shot tasks.

6 Error Analysis

To extend the analysis in the previous sections, Tables 6 and 7 show output samples using the best ranked system (`Helsinki-5`) for Shipibo-Konibo and Otomí, respectively. In each table, we present the top-3 outputs ranked by ChrF and the top-3 ranked by Adequacy and Fluency.

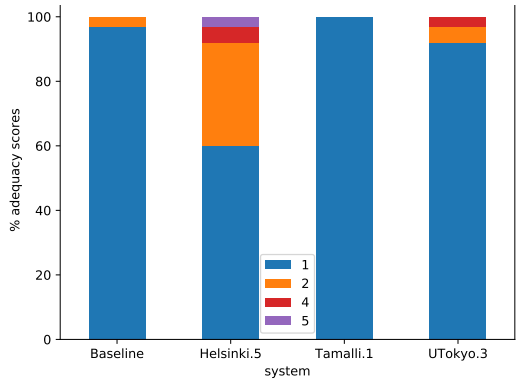
For Shipibo-Konibo, in Table 6, we observe that the first three outputs (with the highest ChrF) are quite close to the reference. Surprisingly, the ad-

Lang.	Rank	Team	Sub	BLEU	ChrF
aym	1	Helsinki	5	2.29	**28.3**
	2	Helsinki	4	1.41	21.6
	3	UTokyo	3	1.03	20.9
	4	Tamalli	1	0.03	20.2
	5	Tamalli	3	0.39	19.4
	6	UBC-NLP	3	0.82	18.2
	7	UBC-NLP	1	1.01	17.8
	8	UTokyo 4		1.34	17.2
	9	CoAStaL	2	0.05	16.8
	10	Tamalli	2	0.07	16.6
	11	Baseline	1	0.01	15.7
	12	Tamalli	5	0.12	15.1
Lang.	Rank	Team	Sub	BLEU	ChrF
bzd	1	Helsinki	5	2.39	**16.5**
	2	Tamalli	3	1.09	13.2
	3	UTokyo	3	1.29	13.1
	4	Helsinki	4	1.98	13.0
	5	Tamalli	1	0.03	11.3
	6	UBC-NLP	1	0.99	11.2
	7	UBC-NLP	3	0.86	11.0
	8	CoAStaL	2	0.06	10.7
	9	Tamalli	5	0.36	10.6
	10	UTokyo	4	1.13	10.4
	11	Baseline	1	0.01	6.8
	12	Tamalli	2	0.25	3.7
Lang.	Rank	Team	Sub	BLEU	ChrF
cni	1	Helsinki	5	3.05	**25.8**
	2	Tamalli	1	0.01	25.3
	3	Helsinki	4	3.01	23.6
	4	UTokyo	3	0.47	21.4
	5	CoAStaL	2	0.03	21.2
	6	Tamalli	3	0.18	18.6
	7	UTokyo	4	0.76	18.4
	8	UBC-NLP	1	0.07	17.8
	9	UBC-NLP	3	0.09	17.6
	10	Tamalli	5	0.07	17.4
	11	Tamalli	2	0.06	13.0
	12	Baseline	1	0.01	10.2
Lang.	Rank	Team	Sub	BLEU	ChrF
gn	1	Helsinki	5	6.13	**33.6**
	2	Helsinki	4	4.10	27.6
	3	NRC-CNRC	1	2.86	26.1
	4	UTokyo	3	3.16	25.4
	5	UTokyo	4	2.97	25.1
	6	Tamalli	5	1.90	20.7
	7	Baseline	1	0.12	19.3
	8	Tamalli	3	1.03	18.7
	9	Tamalli	1	0.05	17.2
	10	CoAStaL	2	0.03	12.8
	11	Tamalli	2	0.13	10.8
Lang.	Rank	Team	Sub	BLEU	ChrF
hch	1	Helsinki	5	9.63	**30.4**
	2	NRC-CNRC	1	7.96	26.4
	3	Helsinki	4	9.13	25.4
	4	UTokyo	3	6.74	22.9
	5	UTokyo	4	6.74	21.6
	6	Tamalli	1	0.01	21.4
	7	Tamalli	3	5.02	20.6
	8	UBC-NLP	1	5.10	19.4
	9	CoAStaL	2	2.07	19.1
	10	UBC-NLP	3	4.95	18.6
	11	Tamalli	5	4.71	16.9
	12	Baseline	1	2.20	12.6
	13	Tamalli	2	3.29	9.4

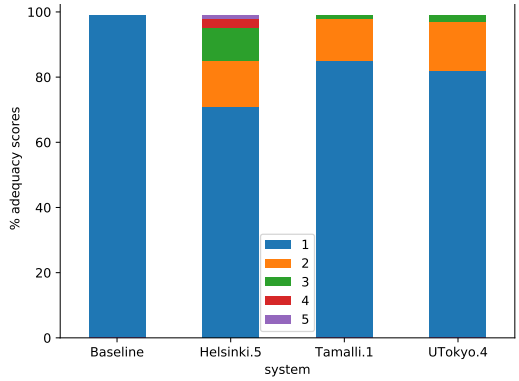
Lang	Rank	Team	Sub	BLEU	ChrF
nah	1	Helsinki	5	2.38	**26.6**
	2	Helsinki	4	2.02	24.3
	3	UTokyo	4	1.2	23.8
	4	NRC-CNRC	1	0.83	23.7
	5	UTokyo	3	0.29	23.6
	6	Tamalli	1	0.03	21.8
	7	UBC-NLP	1	0.12	19.5
	8	UBC-NLP	3	0.15	18.8
	9	CoAStaL	2	0.03	18.4
	10	Tamalli	3	0.11	17.4
	11	Tamalli	5	0.10	16.6
	12	Baseline	1	0.01	15.7
	13	Tamalli	4	0.08	14.5
	14	Tamalli	2	0.03	11.2
Lang	Rank	Team	Sub	BLEU	ChrF
oto	1	Helsinki	5	1.69	**14.7**
	2	Helsinki	4	1.37	14.1
	3	UTokyo	4	1.28	13.3
	4	UTokyo	3	0.05	12.5
	5	Tamalli	1	0.01	11.8
	6	Tamalli	3	0.12	11.0
	7	CoAStaL	2	0.03	10.1
	8	UBC-NLP	1	0.03	8.2
	9	UBC-NLP	3	0.03	8.1
	10	Tamalli	5	0.01	7.4
	11	Baseline	1	0.00	5.4
	12	Tamalli	2	0.00	1.4
Lang	Rank	Team	Sub	BLEU	ChrF
quy	1	REPUcs	1	2.91	34.6
	2	Helsinki	5	3.63	34.3
	3	UTokyo	4	2.47	33.0
	4	UTokyo	3	2.1	32.8
	5	Baseline	1	0.05	30.4
	6	Tamalli	5	0.96	27.3
	7	Tamalli	3	0.64	26.3
	8	Helsinki	4	2.67	25.2
	9	Tamalli	1	0.22	24.4
	10	Tamalli	2	0.69	23.2
	11	CoAStaL	2	0.02	23.2
Lang	Rank	Team	Sub	BLEU	ChrF
shp	1	Helsinki	5	5.43	**32.9**
	2	Helsinki	4	4.53	29.4
	3	Tamalli	1	0.06	20.4
	4	UTokyo	3	0.71	17.5
	5	CoAStaL	2	0.04	17.3
	6	UTokyo	4	0.64	16.4
	7	Tamalli	3	0.31	14.9
	8	Tamalli	5	0.28	12.5
	9	UBC-NLP	1	0.16	12.4
	10	Baseline	1	0.01	12.1
	11	Tamalli	2	0.09	8.9
Lang	Rank	Team	Sub	BLEU	ChrF
tar	1	Helsinki	5	1.07	**18.4**
	2	Tamalli	1	0.04	15.5
	3	Helsinki	4	0.81	15.5
	4	NRC-CNRC	1	0.27	14.3
	5	UTokyo	3	0.06	12.3
	6	UTokyo	4	0.06	11.9
	7	CoAStaL	2	0.06	11.3
	8	UBC-NLP	1	0.08	10.2
	9	UBC-NLP	3	0.06	10.2
	10	Tamalli	4	0.05	8.9
	11	Tamalli	3	0.04	8.4
	12	Tamalli	5	0.02	7.3
	13	Baseline	1	0.00	3.9
	14	Tamalli	2	0.01	2.8

Table 5: Results of Track 2 (development set *not* used for training) for all systems and language pairs. The results are ranked by the official metric of the shared task: ChrF. Best results per language pair are shown in bold, and they are significantly better than the second place team (in each language-pair) according to the Wilcoxon signed-ranked test and Pitman's permutation test with $p<0.05$ (Dror et al., 2018).

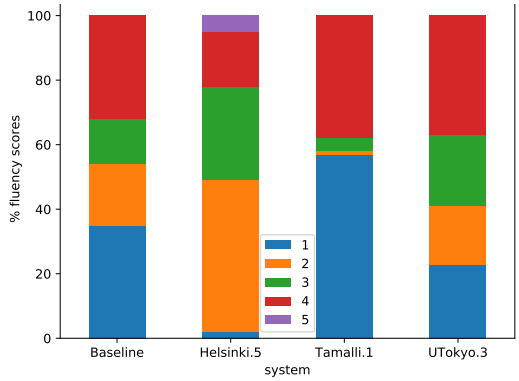
(a) Shipibo-Konibo: Adequacy



(b) Otomí: Adequacy



(c) Shipibo-Konibo: Fluency



(d) Otomí: Fluency

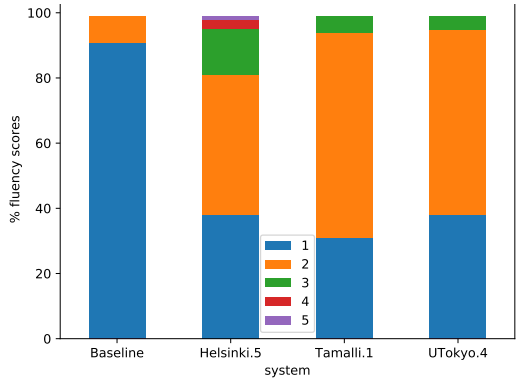


Figure 1: Adequacy and fluency distribution scores for Shipibo-Konibo and Otomí.

equacy annotation of the first sample is relatively low. We can also observe that many subwords are presented in both the reference and the system's output, but not entire words, which shows why BLEU may not be a useful metric to evaluate performance. However, the subwords are still located in different order, and concatenated with different morphemes, which impacts the fluency. Concerning the most adequate and fluent samples, we still observe a high presence of correct subwords in the output, and we can infer that the different order or concatenation of different morphemes did not affect the original meaning of the sentence.

For Otomí, in Table 7, the scenario was less positive, as the ChrF scores are lower than for Shipibo-Konibo, on average. This was echoed in the top-3 outputs, which are very short and contain words or phrases that are preserved in Spanish for the reference translation. Concerning the most adequate and fluent outputs, we observed a very low overlapping of subwords (less than in Shipibo-Konibo), which could only indicate that the outputs preserve part of the meaning of the source but they are expressed differently than the reference. Moreover, we noticed some inconsistencies in the punctuation, which impacts in the ChrF overall score.

In summary, there are some elements to explore further in the rest of the outputs: How many loanwords or how much code-switched text from Spanish is presented in the reference translation? Is there consistency in the punctuation, e.g., period at the end of a segment, between all the source and reference sentences?

7 Conclusion

This paper presents the results of the AmericasNLP 2021 Shared Task on OMT. We received 214 submissions of machine translation systems by 8 teams. All systems suffered from the minimal amount of data and the challenging orthographic, dialectal and domain mismatches of the training and test set. However, most teams achieved huge improvements over the official baseline. We found that text cleaning and normalization, as well as domain adaptation played large roles in the best performing systems. The best NMT systems were multilingual approaches with a limited size (over massive multilingual). Additionally, SMT models also performed well, outperforming larger pretrained submissions.

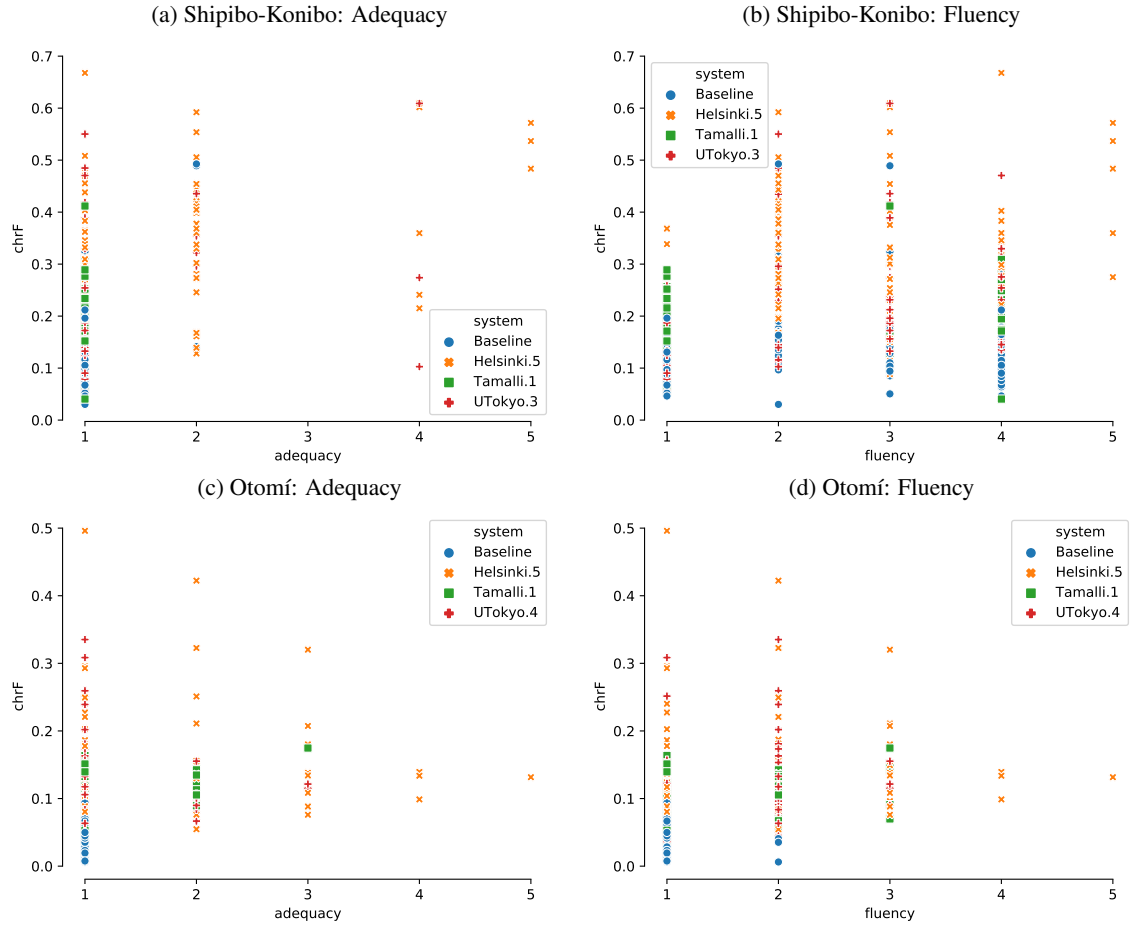


Figure 2: Relationship between ChrF scores and annotations for adequacy (left) and fluency (right).

Scores	Sentences
C: 66.7	SRC: Un niño murió de los cinco.
A: 1	REF: Westiora bakera mawata iki pichika batiayax.
F: 4	OUT: Westiora bakera pichika mawata iki.
C: 60.9	SRC: Sé que no puedes oírme.
A: 4	REF: Eanra onanke min ea ninkati atipanyama.
F: 3	OUT: Minra ea ninkati atipanyamake.
C: 60.1	SRC: Necesito un minuto para recoger mis pensamientos.
A: 4	REF: Eara westiora minuto kenai nokon shinanbo biti kopi.
F: 3	OUT: Westiora serera ea kenai nokon shinanbo biti.
C: 57.1	SRC: Hoy no he ido, así que no lo he visto.
A: 5	REF: Ramara ea kama iki, jakopira en oinama iki.
F: 5	OUT: Ramara ea kayamake, jaskarakopira en oinyamake
C: 53.6	SRC: El U2 tomó mucha película.
A: 5	REF: Nato U2ninra kikin icha película bike.
F: 5	OUT: U2ninra icha pelicula bike.
C: 48.3	SRC: No teníamos televisión.
A: 5	REF: Noara televisiónma ika iki.
F: 5	OUT: Televisiónmara noa iwanke.

Table 6: Translation outputs of the best system (`Helsinki`) for Shipibo-Konibo. Top-3 samples have the highest ChrF (C) scores, whereas the bottom-3 have the best adequacy (A) and fluency (F) values.

Scores	Sentences
C: 49.6	SRC: Locust Hill oh claro, sí, genial
A: 1	REF: Locust Hill handa hâ
F: 4	OUT: Locust Hill ohbuho jä'i
C: 42.2	SRC: Kennedy habló con los pilotos.
A: 4	REF: Kennedy bi ñama nen ya pilotos.
F: 3	OUT: Kennedy bi ñäui ya pihnyo.
C: 32.2	SRC: ¿Te gustan los libros de Harry Potter o no?
A: 4	REF: ¿ di ho-y ya ynttothoma on Harry Potter a hin?
F: 3	OUT: ¿ Gi pefihu na rä libro ra Harry Potter o hina?
C: 13.1	SRC: Un niño murió de los cinco.
A: 5	REF: nã mehtzi bidû on ya qda
F: 5	OUT: N'a ra bätsi bi du ko ya kut'a.
C: 13.9	SRC: Él recibe ayuda con sus comidas y ropa.
A: 4	REF: na di hiâni mâhte nen ynu ynñuni xi áhxo
F: 4	OUT: Nu'a hä häni ko ya hñuni ne ya dutu.
C: 13.3	SRC: Ni siquiera entendió la ceremonia nupcial, ni siquiera sabía que se había casado, en serio–
A: 4	REF: Hin bi ôccode na nînthadi, hin mipâca guê bin miqha nthâdi,maqhuani ngu -a.
F: 4	OUT: Inbi bädi te ra nge'a bi nthati, bi ot'e ra guenda...

Table 7: Translation outputs of the best system (`Helsinki`) for Otomí. Top-3 samples have the highest ChrF (C) scores, whereas the bottom-3 have the best adequacy (A) and fluency (F) values.

Acknowledgments

We would like to thank translators of the test and development set, that made this shared task possible: Francisco Morales (Bribri), Feliciano Torres Ríos and Esau Zumaeta Rojas (Asháninka), Perla Alvarez Britez (Guarani), Silvino González de la Crúz (Wixarika), Giovany Martínez Sebastián, Pedro Kapoltitan, and José Antonio (Nahuatl), José Mateo Lino Cajero Velázquez (Otomí), Liz Chávez (Shipibo-Konibo), and María del Cármen Sotelo Holguín (Rarámuri). We also thank our sponsors for their financial support: Facebook AI Research, Microsoft Research, Google Research, the Institute of Computational Linguistics at the University of Zurich, the NAACL Emerging Regions Funding, Comunidad Elotl, and Snorkel AI. Additionally we want to thank all participants for their submissions and effort to advance NLP research for the indigenous languages of the Americas. Manuel Mager received financial support by DAAD Doctoral Research Grant for this work.

References

Željko Agić and Ivan Vulić. 2019. JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

El Moatez Billah-Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusoglu. 2021. IndT5: A Text-to-Text Transformer for 10 Indigenous Languages. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*, Online. Association for Computational Linguistics.

Marcel Bollmann, Rahul Aralikatte, Héctor Murrieta-Bello, Daniel Hershcovich, Miryam de Lhoneux, and Anders Søgaard. 2021. Moses and the character-based random babbling baseline: CoAStaL at AmericasNLP 2021 shared task. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*, Online. Association for Computational Linguistics.

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

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Matthew Coler and Petr Homola. 2014. *Rule-based machine translation for Aymara*, pages 67–80. Cambridge University Press.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Adolfo Constenla, Feliciano Elizondo, and Francisco Pereira. 2004. *Curso Básico de Bribri*. Editorial de la Universidad de Costa Rica.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Maximiliano Duran. 2010. Lengua general de los incas. http://quechua-ayacucho.org/es/index_es.php. Accessed: 2021-03-15.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, Gustavo A. Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. AmericasNLI: Evaluating Zero-shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-resource Languages.

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

Dayana Iguarán Fernández, Ornela Quintero Gamboa, Jose Molina Atencia, and Oscar Elías Herrera Bedoya. 2013. Design and implementation of an "Web API" for the automatic translation Colombia's language pairs: Spanish-Wayuunaiki case. In *Communications and Computing (COLCOM), 2013 IEEE Colombian Conference on*, pages 1–9. IEEE.

Sofía Flores Solórzano. 2017. Corpus oral pandialectal de la lengua bribri. http://bribri.net.

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.

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

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.

Ximena Gutierrez-Vasques. 2015. Bilingual lexicon extraction for a distant language pair using a small parallel corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 154–160.

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).

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.

Diego Huarcaya Taquiri. 2020. Traducción automática neuronal para lengua nativa peruana. Bachelor's thesis, Universidad Peruana Unión.

Cesar Iter and Zenobio Ortiz-Cárdenas. 2019. *Runasimita yachasun. Método de quechua*, 1 edition. Instituto Francés de Estudios Andinos, Lima.

Carla Victoria Jara Murillo. 2018a. *Gramática de la Lengua Bribri*. EDigital.

Carla Victoria Jara Murillo. 2018b. *I Ttè Historias Bribris*, second edition. Editorial de la Universidad de Costa Rica.

Carla Victoria Jara Murillo and Alí García Segura. 2013. *Se' ttö' bribri ie Hablemos en bribri*. EDigital.

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2021. NRC-CNRC Machine Translation Systems for the 2021 AmericasNLP Shared Task. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*, Online. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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.

James Loriot, Erwin Lauriault, Dwight Day, and Peru. Ministerio de Educación. 1993. *Diccionario Shipibo-Castellano*.

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

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018b. 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 and Ivan Meza. 2018. Hacia la traducción automática de las lenguas indígenas de méxico. In *Proceedings of the 2018 Digital Humanities Conference*. The Association of Digital Humanities Organizations.

Enrique Margery. 2005. *Diccionario Fraseológico Bribri-Español Español-Bribri*, second edition. Editorial de la Universidad de Costa Rica.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. *Oceania*, 135(273):40.

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 12th 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.* WI:Clarks Graphics.

Oscar Moreno. 2021. The REPU CS' spanish–quechua submission to the AmericasNLP 2021 shared task on open machine translation. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*, Online. Association for Computational Linguistics.

Sebastian Nordhoff and Harald Hammarström. 2012. Glottolog/Langdoc:Increasing the visibility of grey literature for low-density languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3289–3294, Istanbul, Turkey. European Language Resources Association (ELRA).

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

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shantipriya Parida, Subhadarshi Panda, Amulya Dash, Esau Villatoro-Tello, A. Seza Doğruöz, Rosa M. Ortega-Mendoza, Amadeo Hernández, Yashvardhan Sharma, and Petr Motlicek. 2021. Open Machine Translation for Low Resource South American Languages (AmericasNLP 2021 Shared Task Contribution). In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*, Online. Association for Computational Linguistics.

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.

Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. Parallel Global Voices: a Collection of Multilingual Corpora with Citizen Media Stories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 900–905, Portorož, Slovenia. European Language Resources Association (ELRA).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21:1–67.

Romano, Rubén and Richer, Sebastián. 2008. Ñaantsipeta asháninkaki birakochaki. www.lengamer.org/publicaciones/diccionarios/.

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*.

Michel Simard and Atsushi Fujita. 2012. A poor man's translation memory using machine translation evaluation metrics. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. 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 AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*, Online. Association for Computational Linguistics.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020. ChrEn: Cherokee-English machine translation for endangered language revitalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 577–595, Online. Association for Computational Linguistics.

Zheng, Francis, Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Low-Resource Machine Translation Using Cross-Lingual Language Model Pretraining. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*, Online. Association for Computational Linguistics.